FACULTY OF ENGINEERING
DEPARTMENT OF CIVIL ENGINEERING
TRAFFIC ENGINEERING AND INFRASTRUCTURE PLANNING SECTION
KASTEELPARK ARENBERG 40, B-3001 HEVERLEE, BELGIUM

KATHOLIEKE
UNIVERSITEIT
LEUVEN

Course H111

Traffic Demand Modelling

May 1998

L.H. Immers
J.E. Stada

Last update:	14 October 1998

Foreword

This text offers an introduction to the field of traffic demand modelling. The subject is part of course H111 that is given to students of the Department of Civil Engineering of the Catholic University of Leuven.

In writing this text, we were particularly inspired by the work of Manheim, Sheffi and d'Ortúzar and Willumsen.

We express our thanks to those students who provided helpful comments.

The authors would appreciate any other comments on the text.

Heverlee, May 1998

L.H. Immers

J.E. Stada

Translated from the Dutch by L. Hurley (2002)

email:
ben.immers@bwk.kuleuven.ac.be
jim.stada@bwk.kuleuven.ac.be

tel:
0032-16-321669
0032-16-321670

address:
Katholieke Universiteit Leuven
Departement Burgerlijke Bouwkunde
Park van Arenberg 40
B-3001 Heverlee
Belgium

Contents

1. An analysis of the transport system

Transport arises as a consequence of the spatial division between economic and social activities. The aim of this chapter is to describe the interaction between the total of social and economic activities and the transport system. The conceptual model that ensues serves as a framework in which the other subjects to be discussed in this course can be placed in their correct context. The description is based on Manheim [1979][1§]. The chapter closes with some remarks about the role of modelling in traffic planning.

1.1 Social change

Society undergoes constant change. Due to the strong interactions between the transport sector and society (traffic largely derives from social life and its associated pattern of activities) changes in the social sphere can have important consequences for the functioning of the transport system.

Basically, there are three separate types of changes that affect the functioning of the transport system.

- *Changes in the demand for transport.* Population increase, increased incomes and changes in land-use affect the demand for transport.
- *Changes in transport technology*. Technological innovation is not confined to the different means of transport only. Innovation also occurs in the form of new transport concepts such as special lanes for target groups, information systems for the travelling public, road pricing, etc.
- *Changes in value judgements*. Decisions in the transport sphere can have far-reaching implications. Today there is a greater appreciation of the social consequences of particular measures and of the environmental impact inherent in traffic.

We will begin our analysis with a systematic description of the transport system and of its interaction with its socio-economic environment. Next, we will identify the options that are available to influence the direction of development and to determine the effects of these steering measures.

The basic assumptions of the analysis are as follows:

- The transport system in an area is seen as a coherent multi-modal system. The term *multi-modal* means that a number of forms of transport or "modalities" are included in the equation. The term *transport system* refers to the totality of transportation modes, networks, terminals, transportation services, etc.
- Because of the large number of interactions, the transport system can only be looked at in conjunction with the social, economic and political pattern of a given area. From this point, we will call this framework of social, economic and political activities the *activity system.*

[§] Notes refer to the list of references at the end of the text.

1.2 Interaction between the transport system and the activity system

Figure 1-1 shows the interaction between the transport system *T* and the activity system *A*. The interaction between these systems results in a flow pattern *F* in the multi-modal network. By the *flow pattern F* we mean the totality of transportation flows in terms of origin, destination, transportation modes, routes, departure times, volumes, etc. and the associated level of service of the traffic system, for example in terms of travel times. In fact, the *F* variable describes the situation of the transport system at a particular moment in time.

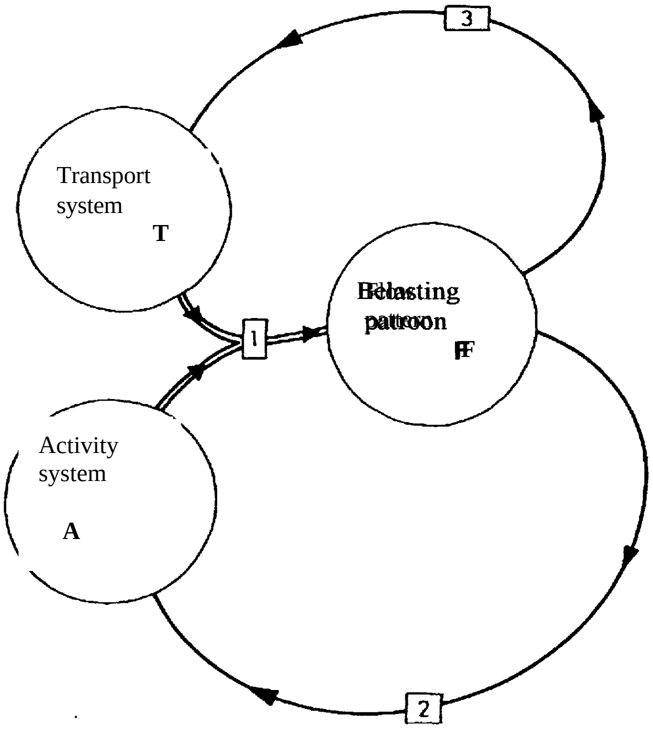


Figure 1-1 Interaction between transport system and activity system.[1]

The following relationships are indicated in Figure 1-1:

Relationship 1: A given transport system and a given activity system will result in a given flow pattern on the multi-modal transport network.

Relationship 2: The use of the multi-modal network, as well as the associated level of service in the traffic system and the consequent use of scarce resources will lead to alterations in the activity system.

Relationship 3: The use of the multi-modal network, and the traffic pattern associated with it, will lead institutions and operators to alter the quality of their services and their infrastructure.

1.2.1 Transportation Options

There are many ways of intervention both in the transport system and in the activity system. It is important to realise that different groups can make different decisions. These include, for example, individual customers of transport services, travel operators that offer travel services, and institutions that formulate and execute transport policy. The opportunities for intervention, the options, can be formulated both for the transport system and for the activity system.

As for the transport system, options for the following aspects are available:

- *Technology.* Examples here are the introduction of new transportation concepts such as containers, the development of new propulsion techniques and innovations in the area of road construction.
- *Networks.* Networks are defined by a set of nodes and a set of links that connect the nodes. One can choose between all kinds of network configurations such as radial, concentric or grid patterns. In addition, the properties of networks can be changed. These properties, such as capacity and speed limits are often attributed to the network links.
- *Vehicles.* The number of vehicles to be used and their properties can be influenced. Such decisions can be taken by managers of transportation companies, but also by policy makers through legislation.
- *System operating policies.* Regulation can be done through price intervention, subsidies, or other fiscal measures. Regulation can also be done through the legislature. Also included are operational decisions in respect of scheduling, types of service to be offered, and routing.
- *Organisational policies.* The structural organisation of the transport system has its repercussions on the transport system. Here one can think of the distribution of responsibilities between the various institutions or of the internal organisation inside transportation enterprises.

In the activity system, the options can also be divided into a number of categories:

- *Travel options.* This concerns the options available to every potential user of the transportation system, i.e. the choice whether or not to travel, the choice of transport mode, the time of travel and the route. The combined result of all these choices constitutes the demand for transport.
- *Spatial dispersion of population and economic activities.* Economic and social factors determine the location of residence and the place of work as well as the scale of the economic activity. They also influence the need for transport. This applies, on the one hand, to factors that can be actively directed through the implementation of government policy. And on the other hand it applies to processes which are free from outside intervention. In both cases, however, we speak of options in the activity system.

1.2.2 Consequences or impacts of transportation

The consequences or impacts of the measures to be taken are manifold. Furthermore, the consequences may turn out to be positive for one group and negative for another. This aspect must be born in mind during the evaluation process. A possible subdivision of the consequences may be as follows:

- *User impacts*. These apply to the individual traveller and those in the goods transporting business. The load on the transport system is experienced by the user in terms of travel times, costs, comfort, etc.
- *Transport operator impacts*. These concern the consequences for the operators of public transport companies, etc. Here, costs and returns are important factors.
- *Environmental impacts*. Impacts on the environment, such as air pollution, noise pollution and fragmentation of the landscape have gained in influence on the design of the transportation system
- *Functional impacts*. These are the implications for the functional quality of the activity system. One can think of changing incomes in retail outlets and in land prices.
- *Governmental impacts*. The impacts of the transport system can lead to administrative consequences in the area of legislature, in the allocation of responsibilities, etc.

1.2.3 The prediction problem.

We must now find a way to predict the consequences of a variety of available interventions or options. It is important that the implications are optimally estimated or calculated in order to guarantee maximum certainty in the solution to a problem that has been identified, or to realise a formulated target. This problem of prognosis has been schematised in Figure 1-2.

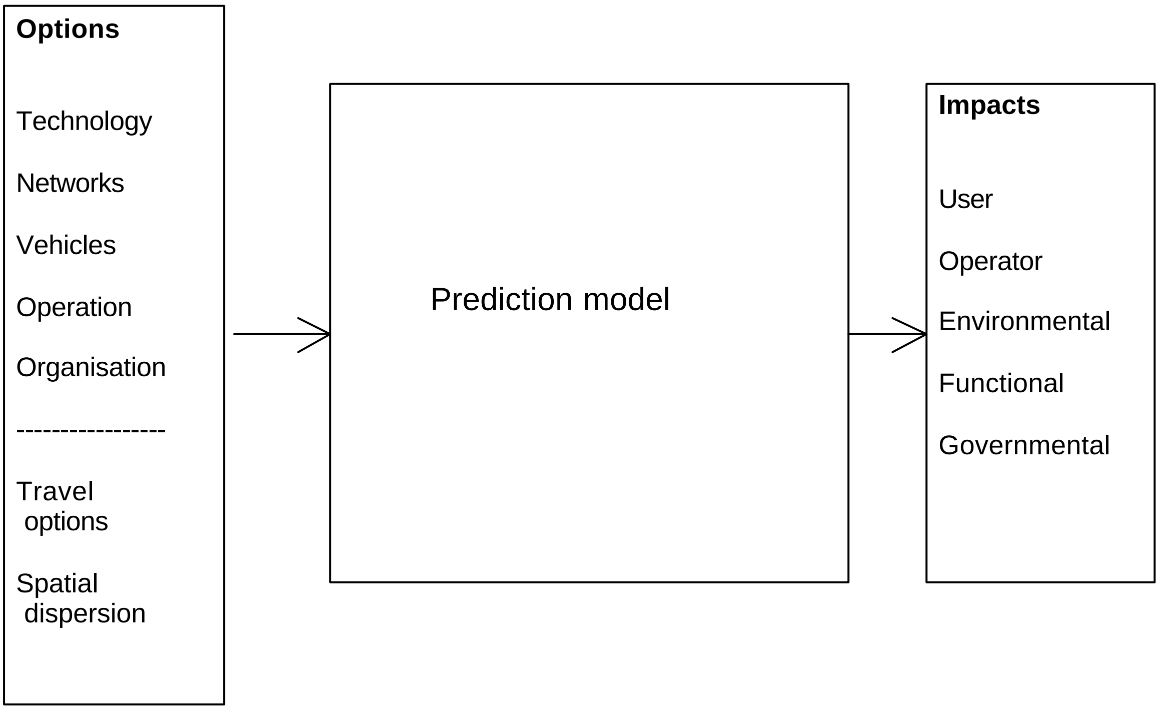


Figure 1-2 The prediction problem.[1]

To calculate the effects of interventions on the transport system, we need to predict the flow pattern F for a given transport system T and a given activity system A.

1.3 *Demand and supply on the transport market*

The starting point in the prognosis is that a market can be identified for the transport sector that functions independently of other markets. The following variables and functions are important in describing the functioning of this transportation market:

Variables:
T the transport system
A the activity system
F the flow pattern
S the level of service (travel times, fares, etc.)
V the volume of the traffic flows

Functions:
J supply function: $S = \mathrm{J}(V,T)$
D demand function: $V = \mathrm{D}(S,A)$

A *supply function* (see Figure 1-3a) gives the level of service as a function of the volume of the traffic flows in any given transport system. In economic terms, a supply function indicates the behaviour of the supplier of producer. In our case, the supply function indicates the service that a given transport system can offer at different rates of traffic flow. This is why the supply function is sometimes called a service function. The level of service incorporates a number of elements. Amongst these, the travel time and the costs of travel are important elements. Changes in the transportation system change the shape of the supply function.

A *demand function* (see Figure 1-3b)describes the size of the transport flows as a function of the level of service in a given activity system. The function is read starting from the vertical axis. As the level of service rises, so the demand rises. Changes in the activity system lead to changes in the demand function.

The flow pattern *F* of the multi-modal network is defined as the combination of the transportation flows *V* with their associated level of service *S*. The load pattern *F° = (V°, S°)* for a given transportation system and a given activity system can be calculated on the basis of the equilibrium that emerges between demand and supply, as indicated in Figure 1-3c.

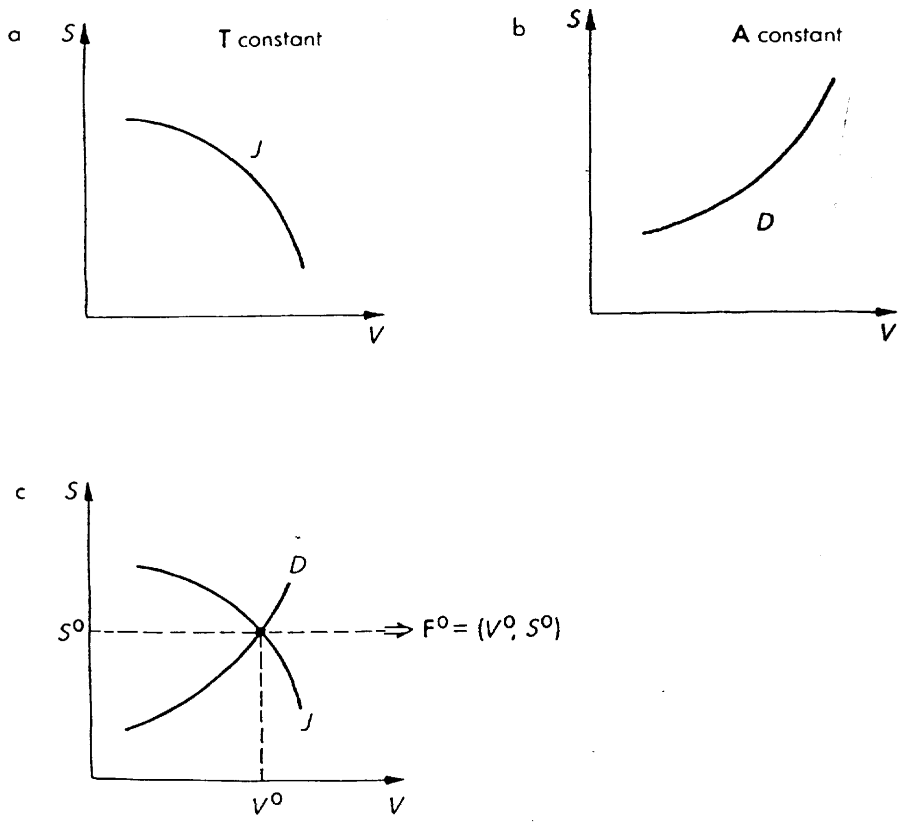


Figure 1-3 Equilibrium on the transport market.[1]

In addition to the general case as shown in the previous figure, Figure 1-4 shows two other functional relations.

Do note that the vertical axis now shows the travel time, not the level of service. The travel time is taken here as a measure of the service level. This leads to a reversal of the vertical axis.

The first example, Figure 1-4a, concerns a constant supply function. This could apply to a road of infinite capacity, for example. Travel time does not alter as a result of a change in traffic volume.

The second case, Figure 1-4b, shows a constant demand function. The demand is not responsive to changes in the level of service. This could apply to passengers, for example of public transport, who have no alternative transport options. These are sometimes called forced travellers or "captives".

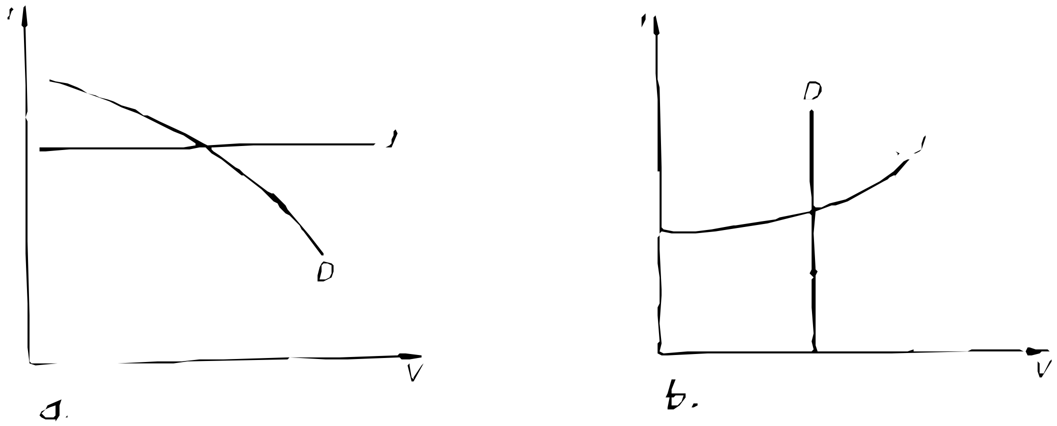


Figure 1-4 Constant supply and demand function.[1]

1.4 Consequences of the introduction of new transportation facilities

When the flow pattern indicates congestion at certain points, government institutions can, for example, decide to provide new infrastructure. Public transport companies can also decide to adapt their services. This feedback from the flow pattern to the transportation system was indicated by relationship 3 in Figure 1-1.

So what happens when the transportation system is altered? Let's assume that the new transportation system is an improvement on the old one. Figure 1-5 shows the supply function for both the old system (J°) and the new system (J^1).

Again, as in Figure 1-4, the vertical axis shows the travel time! An increase in the time spent travelling means a lower level of service.

If we were to introduce a new improved transportation system immediately, this would lead to a reduction in travel time. A new equilibrium will occur in which the reduced travel time will entice more consumers to avail of the service. The equilibrium is achieved with the values V^1 and t^1.

We must, however, keep in mind, that the time-lapse between a possible identification of insufficient capacity and the planning and bringing on steam of additional capacity will take many years. In other words, by the time the new service is eventually introduced, the economic and demographic developments, such as population increase, increased car-ownership and so on, have led to an altered demand function from D° to D^2.

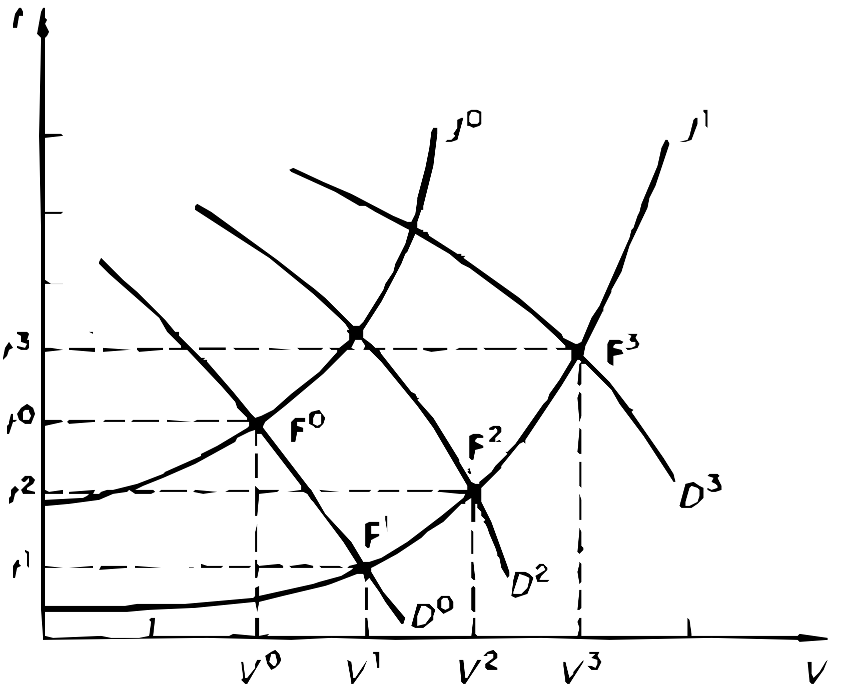


Figure 1-5 Equilibrium in the short and long run.[1]

The new equilibrium F^2 will be achieved several weeks or months after the introduction of the new service. This time-lapse is a function of the adjustment required by people to familiarise themselves with the changed departure times, routes or transport modes. We call this equilibrium that was indicated by relationship 1 in Figure 1-1 the short-run equilibrium. It emerges through changes in travel behaviour.

The improved quality of the transport system, moreover, induces another development. Firms will appear in areas that have become more accessible, people on higher incomes will move out of town to settle in easily accessible rural areas, etc. In short, reacting to and sometimes even anticipating changes in the quality of the infrastructure, relocation of activities and possible new developments can occur. Due to the improved transport system, people will choose to live at a greater distance from their workplace and new customers are likely to come forward to use the improved transport system. This will, in its turn, give rise to an increase in the demand function. The demand function will shift to a higher level. This process is called "activity shift" and is shown in relationship 2 of Figure 1-1. This development is a slow one and so the demand function D^3 of the new situation in the activity system will only be realised in the long term. The equilibrium F^3, which is a consequence of alterations in the activity system, is called the long-run equilibrium. This equilibrium will usually not be achieved because new services will in the meantime have been introduced which will incline the system towards a new equilibrium. The long-run equilibrium, therefore, may serve to indicate trends in development.

Note, in Figure 1-5, that the travel time t^3 in the final situation is higher than the original travel time $t°$ with which we began our analysis. The level of service has, therefore, gone down, in spite of the provision of new infrastructure! Thus there is a possibility that the implications of the introduction of a new service on the demand pattern are such as to eventually reduce the quality of the traffic system in the new situation.

If, as explained above, the quality of the transport system has decreased in spite of the introduction of new traffic facilities, it would not be correct to conclude that the provision of this new facilities was pointless. After all, the volume of traffic in the new situation is also much larger. Circumstances determine whether we can speak of a favourable or unfavourable development. Remember that the increased demand is the result of a shift in the activity system. This can, on the one hand, imply that the same number of users travel longer distances. Or it can mean that the distance travelled per user shows a nominal increase only, but that the number of users has increased. A combination of the two is, naturally, also possible. If the increased demand was mainly a function of an increase in distance travelled by users, one could speak of an unfavourable effect. Where the provision of the new transport service leads to mobility for more users, one could - with some reservation - see this as a positive result.

We refer the reader to chapter 5.3.1 for additional observations regarding the consequences of changes in the transport system.

1.5 ***Model types***

Based on the analysis above concerning the mechanisms of equilibrium that apply to the transport market, a list model types can be drawn up that are required to calculate the consequences of policy measures. These concern the following models:

- *Demand models* that determine the scale of demand as a function of the service level.
- *Supply models* that, depending on the measures to be taken, determine the level of service as a function of load on the network.
- *Short-run equilibrium models* that, on the basis of demand and supply, determine the scale of the traffic flows in a network.
- *Long-run equilibrium models* that describe the interactions between changes in the infrastructure and the spatial distribution of activities.
- *Impact models* that indicate which implicit consequences are involved in the provision of improved service levels, such as the necessary investments, environmental impacts, social effects, safety, etc.

We need models that belong to the five types described above, to solve our problem of prognosis, which was schematically presented in Figure 1-2.

Figure 1-6 shows the aforementioned model types in their mutual context.

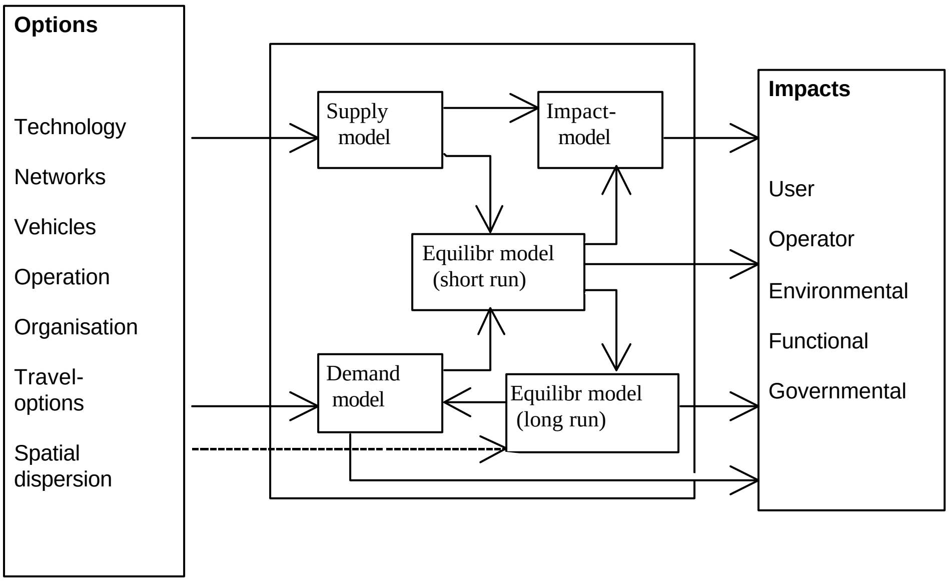


Figure 1-6 Model types needed for prognosis

1.6 Practical implementations of the conceptual model

The conceptual model presented in this chapter can be used to place the multitude of complex transport phenomena that we encounter on a daily basis, in their appropriate context. The model provides an insight in the fundamental interactions between the transport system and the socio-economic system in which it is embedded. The model is still too abstract, however, to serve as a basis for making transport calculations.

In chapter 2, we will discuss the traditional traffic demand model. This model is used in many places around the world to make real transport calculations. The traditional traffic demand model consists of a number of sub-models, which we can see as practical implementations of the conceptual model discussed in this chapter.

Below is an overview of the 5 types of model described above and their parallel in the traditional traffic demand model.

- Demand models appear in the traditional traffic demand model in a number of sub-models, namely the *production/attraction model,* the *distribution model,* and the *mode choice model.*
- Supply models are reflected in the so-called *time-loss functions,* which indicate the relation between travel time or travel costs and the flow rate on road sections.
- Short-run equilibrium models are used in the so-called *assignment models,* which determine the routes in a network, by taking into account that traffic flows themselves influence the travel times on links in the network.
- Long-run equilibrium models, which reflect the influence of the flow pattern on the activity system, are often implemented in the form of a scenario-approach. Here one postulates a certain plausible future socio-economic development that is used to calculate the consequences for the transport system. Then there are models that try to directly predict the effect of the flow pattern on the spatial and socio-economic

development. However, in practice these models are, as yet, little used. We will not discuss such long-run equilibrium models in this course.

- Impact models have been developed in large numbers. There are models that can calculate the impact of traffic on air pollution, noise pollution and safety. Because these models are very specialised, they will not be discussed in this general introductory course.

1.7 Models in the planning process

Figure 1-7 indicates where in the planning process, the prediction model can be applied. (Models can also be used in other phases of the planning process. Because they are not discussed in this chapter, these models are not shown in the figure.)

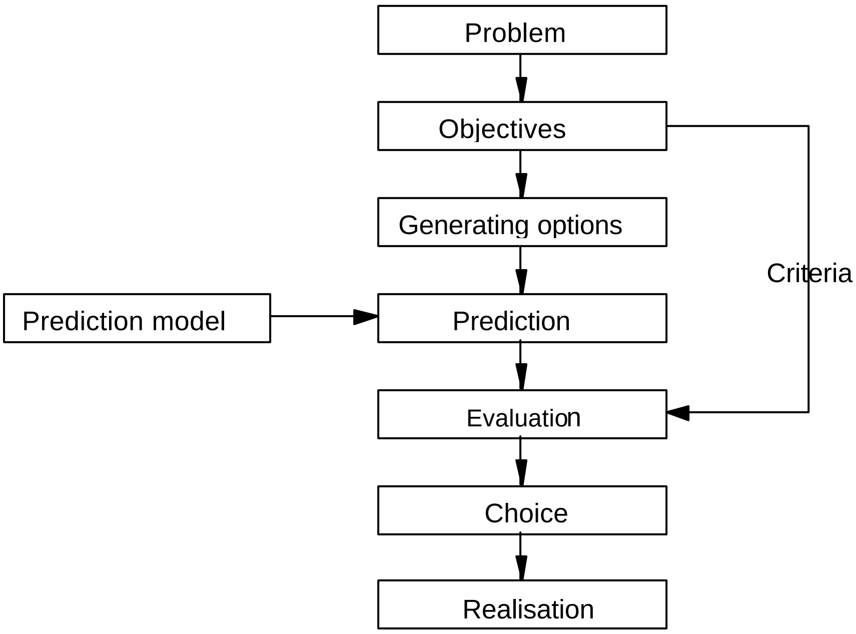


Figure 1-7 Planning process.

There are some important reasons why models are used:

- A model, especially if it is in the form of a computer programme, enables one to include complex interactions that could easily be overlooked without the use of models or that could be incorrectly interpreted. A model does not always have to come in the form of a computer application, but it lends itself very well to the purpose.
- When one compares them to the implementation costs, model calculations are a very cost-effective means by which to give answers to the consequences of various alternative options. However, one must keep the limitations of models in mind. Using a model means that concessions are made in regard to the reproduction of reality. Models are a simplified reproduction of a part of reality. If this was not so, we might as well take reality itself as our model.

We conclude this chapter with some important criteria needed to assess the prognosis models:

- *Relevance*. The model must be able to calculate the impact of every single measure of intervention that one wants to investigate.
- *Accuracy*. The results of the model and the observations must agree to a reasonable extent. Note that it is unrealistic to expect great accuracy in traffic models, when compared to models in the exact sciences. Traffic engineering is no exact science, but a field of knowledge that lies between the exact and the social sciences.
- *Theoretical foundation*. The formulation of the model should, ideally, be based on a solid theoretical foundation. Those models that depend on a simple extrapolation of observed behaviour have only a limited field of application, both in place and time.
- *Simplicity.* The simplicity of a model should be seen as its mark of quality. Generally, simple models are also more robust, i.e. they are more resistant to input errors than complex models.
- *Validation*. Matching the model results with the observations is called the calibration or the empirical fitting of a model. Validating a model, however, means testing the model's capability to make predictions. Here one must not use data that have been used in the calibration. Every model should be properly validated.
- *Practical applicability.* One must be able to apply the model in the framework drawn up in regard to available time, funds, and personnel. This relates particularly to the gathering of input data required for the model.

1.8 Summary

Transport evolves primarily through the spatial separation of economic and social activities. The combined social and economic activities, including the political and other consultative- and decision-making structures, is called the activity system. The activity system represents the demand side in the realisation of transport. The totality of transportation services is called the transport system and this represents the supply side. Demand and supply result in an equilibrium, the flow pattern. The flow pattern is expressed in a particular volume of transport with attendant attributes such as travel times, levels of congestion, etc.

The equilibrium is dynamic, not static, in that a certain volume of transport can, in its turn, trigger changes in the activity- and the transport systems, resulting in a new equilibrium. This process is one of perpetual change.

2 Structure of the traditional traffic demand model.

In this chapter we give an introduction to the traditional traffic demand model. This model can be seen as an elaboration of the conceptual model described in chapter 1. The model is called the traditional traffic demand model because years of research and application have led to a commonly accepted model structure. This structure evolved in the sixties and remained more or less the same in spite of enormous progress in modelling techniques.

The development of traffic demand models began in the fifties in the United States, where elaborate models were built amongst others for the cities of Detroit and Chicago. In the sixties, traffic models began to be used in England. From England it spread to the rest of Europe. The widespread use of traffic modelling in Flanders began relatively late (in the early nineties). Today, every province has its own multi-modal traffic model.

There is an extensive library of literature on the topic. One of the best references is Ortúzar and Willumsen (1995)[2].

We begin the treatment of the structure of the traditional traffic demand model with a short exposition of the function of this model and a discussion of some of the terms used in common model theory.

2.1 Function of the traffic demand model

Both the activity- and the transport system are subject to constant change. These changes either occur autonomously, or they are planned. By autonomous developments we mean societal changes outside of our sphere of influence. Examples include technological developments, changes in incomes structures, changing attitudes towards work and leisure time, etc. Developments can also result from deliberately planned intervention. Examples here include the construction of new infrastructure, measures to stimulate alternative means of transport, for example through pricing mechanisms, the pursuit of particular planning goals, etc.

The function of the traffic demand model that will be presented in this chapter is to calculate the equilibrium that results from a given situation in the activity- and the transport system (short-run). The calculated traffic flows can be used to design traffic facilities. Very important also are the external effects of traffic flow. Here, one must mainly think in terms of negative effects such as environmental degradation, loss of time and money, and congestion-induced irritation.

In chapter 1 we showed that the traffic flows influence the activity system and the transport system. The effects are largely long-term, and include such processes as the relocation of work and living areas and the adaptation of current infrastructure or the construction of new infrastructure. Current traffic models are not quite up to handling such feedback. Possible quantitative models that could adequately describe this complicated process are still in the development phase.

Below are a number of questions on traffic flows that can be explained using travel demand models:

- How does the transport pattern in an area change following the construction of a new motorway?
- What are the consequences of locating employment to the outskirts of a city?
- Which are the optimum locations for work- and urban areas to be assigned in a region?

These questions give an idea of the scale level to which traffic models tend to be applied. The traffic models that are dealt with in this chapter work at the level of an urban area or a region such as a province or county. They definitely do not deal with the description of the traffic pattern on a single road or junction.

2.2 Model concept

A model is a simplified representation of a part of reality.

Models can be classified according to their way of representation. The representation can be concrete or physical, as in scale models. Abstract models belong to a totally different class. There are many kinds of abstract models. The kind of model that interests us in this chapter is the mathematical model. The traffic model is a mathematical model. We will return to the common structure of a mathematical model later.

We can also classify models according to their end-use. Here, we distinguish between descriptive, predictive and normative models. Descriptive models are confined to the schematic representation of a phenomenon. They do not aim to explain this phenomenon. A predictive model, or a prognosis model, has a greater reach. Starting from the current state of a phenomenon, and having knowledge of probable future influences, it is used to predict the future situation. One can follow a simple trajectory, for example, by extrapolating trends. Or one can try to reach a deeper understanding of the relevant phenomenon by developing a theory. In the last case one speaks of a causal model. The classic traffic demand model that we will discuss later on is an example of a causal prognosis model. Then there are normative models. Here one decides on a particular norm or objective, namely a goal function, or an objective function that needs to be optimised. Next, one attempts to determine which conditions need to be met in order to achieve the optimal situation. Normative models also come under the name of prescriptive or optimisation models.

There are still more possible classifications. We can, for example, classify models according to the role that time plays in the description of phenomena. When the flows in a traffic model are time-dependant, we speak of a dynamic model. If we assume that the flows are constant over a specific period of time, we have a static model. Lastly, we mention the classification that looks at the role of chance played in the model. There are many models that use chance or stochastic variables. These are so-called stochastic models. Stochastic variables are not used in deterministic models.

In conclusion, we can state that the traditional traffic demand model that we are to discuss is an abstract, mathematical model. It is a static, largely causal predictive model, that features both deterministic and stochastic sub-models. The trend over the last few years has been to develop dynamic traffic models.

2.2.1 Mathematical models

Mathematical models, including those in the area of traffic engineering, comprise systems of mathematical equations where the behaviour of the variable *Y* is deduced from a number of variables X_i.

$$Y = f(X_i, a_j)$$

Where:

Y Dependent (or the to be explained) variable

X_i Independent (or explanatory) variables

a_j Parameters

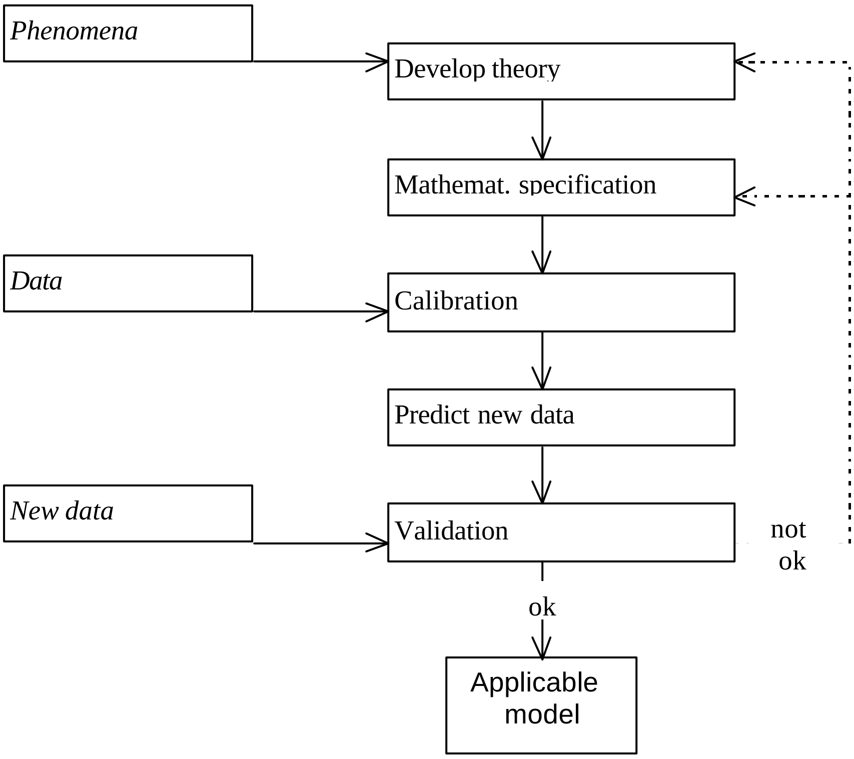


Figure 2-1 Construction of a mathematical model

Figure 2-1 shows that a specific procedure is followed in the construction of a mathematical model. The process of observation and thought about a phenomenon in the area of a traffic engineering, leads to the formulation of a theory. This theory is reflected in a mathematical specification of the model.

The model-specifications include:

- A determination of the functional form of the equations
- A specification of the independent variables

A possible functional form for a theoretical model could be a linear additive form such as:

$$Y = a_1 X_1 + a_2 X_2 + \ldots.$$

Or a multiplicative form such as:

$$Y = aX_1 X_2 \ldots$$

2.2.2 Calibration and validation

In addition to the independent variables X , the specified mathematical model shows a number of parameters a . Calibrating the model means that we determine the values for the parameters in order to ensure maximum agreement between the values calculated through the model and the original observations. The observations for a traffic model from questionnaires and traffic counts apply to a specific point in time that we use as a point of reference. One usually says that the model is calibrated for a specific baseline year. Other terms used for calibration are: the ‘estimation’ of a model and the ‘fitting’ of a model.

Even when a model had been calibrated it still does not mean that the model can be used to make predictions. When there are a sufficiently large number of parameters, it is, in principle, possible to correctly fit practically every model. Validating a model means that the model predictions are compared with new observations. *These observations may not already have been used in the calibration!* It is possible, for example, to validate a model by using it to ‘predict’ a situation from the past and to compare this with the actual situation in the present. (“back casting”).

A validated model, lastly, can be used to make a prognosis for a specific target year. This does require some caution. Traffic models are based on an analysis of observed travel behaviour. They are, therefore, only valid for circumstances that do not deviate too much from the circumstances that form the basis of the analysis. The accuracy of the predictions, moreover, will decrease as the target year lies further in the future. This is so because the models do not take account of possible gradual changes in traffic behaviour.

2.3 Structure traffic demand model

The scale at which traffic services are used, such as the volume of traffic on a road and the number of train passengers on a given route, arises from a number of choices that are made by individual transportation consumers.

The choices faced by the individual include:

- The choice whether or not to travel
- The choice of the time of departure
- The choice of destination
- The choice of transport mode.
- The choice of route.

The list above suggests that the choice process consists of a number of separate choices and that the choices occur in a definite order. This will usually not be the case. Some choices will be made simultaneously, i.e. all at once and directly inter-linked, and not in a particular order.

In order to make the problem mathematically tractable, we generally assume that the choices can be modelled separately. As for the sequence of the models, the mutual relationship between choice of destination, choice of transport mode and choice of departure-time causes a problem. Because of the interwoven nature of these choices, it is best to group them in one model segment.

Departure time modelling is implicitly incorporated in the first sub-model (the production- and attraction model). This sub-model calculates the number of journeys undertaken for a specific period, for example the morning peak period.

The sub-models need input data. The traffic pattern is the result of a large number of individual choices. But we cannot model every individual choice. Combination or aggregation is unavoidable if we want to achieve a workable model. The study area is, therefore, divided into a number of zones. The aim is to distinguish zones that are socially and economically homogenous. Inside a zone, trips are grouped with respect to trip-purposes and person-types. In addition the area is provided with schematised networks for homogenous transport modes.

The following sub-models are used (see Figure 2-2)

- *Production/attraction model.* A production model describes the number of journeys that are generated in a zone as a function of a number of personal characteristics and characteristic features of the environment. The total number of journeys produced per zone is calculated, without reference, as yet, to the destinations of these trips. The trips are specified as to point of time (e.g. peak or off-peak). A zone can both generate trips and attract them. An attraction model describes the total number of trips that a zone attracts independent of the origin, as a function of characteristics such as employment rate and retail area. The productions and attractions that have been calculated are also called trip-ends. Taken over a sufficiently long period of time, the total number of departures that are calculated over the combined zones must equal the total of arrivals. To this end, the results of the production- and attraction models are adjusted, if necessary. This is called *balancing* of productions and attractions.

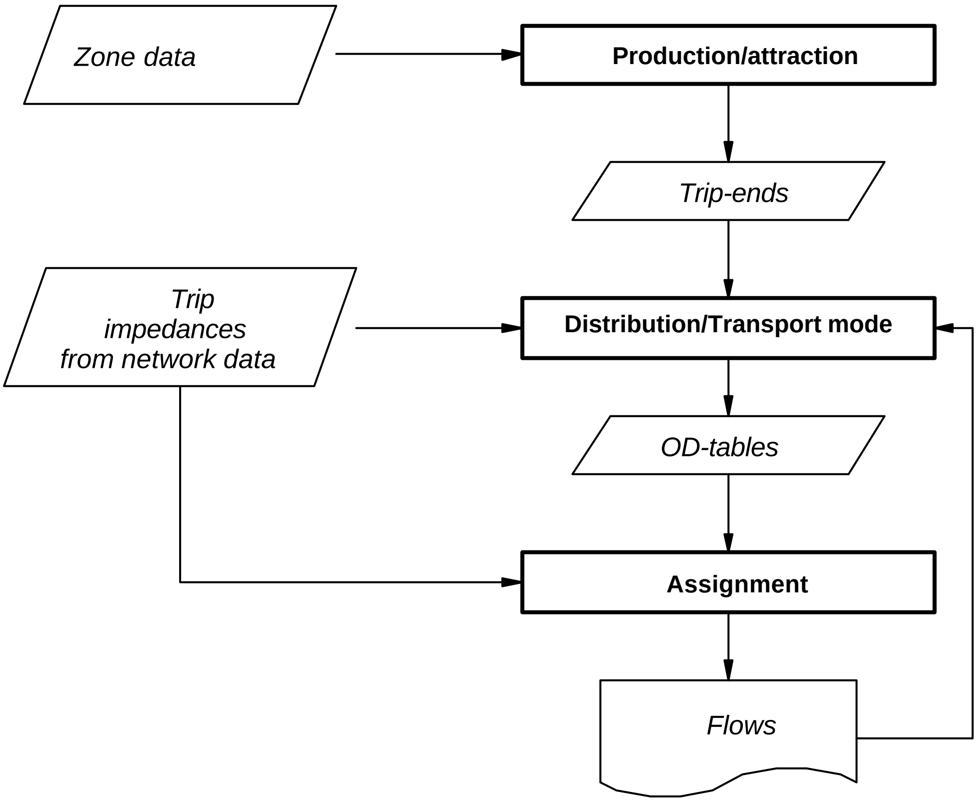


Figure 2-2 Structure of the traditional traffic demand model

- Distribution/Mode choice model

  Distribution model. In the distribution model the trips originating in a certain zone *i*, that have been calculated in the production model, are distributed over possible destinations *j*. The trips that have been calculated in the attraction model with zone *j* as their destination, are distributed over the possible points of origin *i*. The connection between points of origin and destination is calculated as a function of the ease or resistance with which the distance between *i* and *j* can be bridged. Depending on whether one distinguishes several trip purposes or person-types, the calculations in this sub model can deliver one or several origin-destination tables. In an origin-destination table (OD-table) the rows of the table represent the origins and the columns represent the destinations. The entries in the table represent the trips between a certain origin and a destination.

  Mode choice model. A mode choice model calculates which type of transport travellers use as a function of personal characteristics and the relevant transport modes. The calculated distribution amongst the various transport modes is called the “modal split”. The calculation results in a further subdivision of OD-tables according to traffic mode.

- *Traffic assignment model.* Even for a single mode of travel, there often are several possible routes between an origin and a destination. In a traffic assignment model (also called a route choice model) the trips between the origins and destinations specified in the OD-tables are assigned to the possible routes in the network according to characteristic of these routes (distance for example). The assignment is done for each traffic mode separately, and results in the traffic flows on the links of the various networks. The calculated traffic flows imply certain definite journey duration times. In order to ensure overall consistency of the variables in the model, the travel duration times are compared with those used in the distribution/mode choice model, and if necessary an iteration step is added.

2.4 Overview

The discussion of the structure of the traffic demand model in this chapter highlights the significance of the "choice" concept. The observed traffic pattern is the outcome of the choice behaviour of a large number of individuals. A general theory of individual choice behaviour is needed. One theory that can explain many transport phenomena is the so-called discrete choice theory. We will, therefore, precede our detailed discussion of the traffic demand model in later chapters by a discussion of this theory.

Before a traffic model can be applied, we need data concerning the study area and concerning the networks that will be used for the trips. We will examine this subject in a separate chapter.

The last four chapters will then look at the production/attraction model, the distribution model, the mode choice model and the traffic assignment model.

3 Discrete choice theory

The discrete choice theory that will be discussed in this chapter is a general theory applicable in situations where people are expected to choose between mutually exclusive alternatives. The discrete choice theory, which originates in the sciences of psychology and economy, appeared to be most suitable to the modelling of choice situations in traffic engineering. The standard reference in this area is the book by Ben Akiva and Lerman (1985)[3] .

The basic principle of this theory is that when an individual is confronted with a situation in which he has the choice between a number of mutually exclusive alternatives, he will attribute a certain valuation or utility to each alternative. This utility is a function of the characteristics of the alternatives and of the characteristics of the person making the choice. The choice will fall on the alternative that shows the greatest utility. The problem is that utilities are not immediately observable and measurable. What can be observed are the characteristics of the alternatives (as, for example, travel time and costs for a certain travel mode), that we consider to influence the utility that an individual assigns to an alternative.

We can never be certain, however, that we have accounted for all the characteristics that influence the utility. And even if we included all the characteristics, the individual will, for reasons that elude our observation, not always judge the alternatives in the same way. This is why the utilities are modelled as chance variables. The choice models that are formulated using the discrete choice theory, will, therefore, indicate the probability with which various alternatives will be chosen.

The name given to the set of alternatives from which choices can be made is called the choice set. Some choice sets are inherently continuous, which means that they contain an infinite number of elements. Assume, for example, that one must decide on the amount of basic materials that are used in the manufacture of a particular product. This is a case of a continuous choice set. However, in this chapter we are interested in discontinuous choice sets, i.e. sets which contain a number of finite discrete points only. This explains the term 'discrete choice theory'.

3.1 Logitmodel

Imagine that an individual finds him self in a particular choice situation. He has the choice of K alternatives. We want to know which alternative he will choose. Now, the discrete choice theory postulates that he will assign a certain quantified value to each alternative. This value is the called the "utility" of the alternative. Having assigned a utility to each of the alternatives, he will chose the alternative with the highest utility.

The utility of the alternative i for an individual is a function of the characteristics of the alternative and the individual characteristics of that person. It can happen that individuals who are apparently in exactly identical choice situations arrive at different choices. This can be caused by a number of factors:

- *Unobserved characteristics*. It is possible that we failed to take account of particular characteristics that are very important for the individual. This failure can be due to ignorance, or through a lack of data.
- *Measurement errors*. Some characteristics can be subject to measurement errors. If it is a matter of a choice between two routes, for example, it is possible that the travel times that have been estimated by the researcher differ from the actual experienced travel times.
- *Incorrect specification of the utility function*. The utility is a function of a number of characteristics. We usually adopt a simple linear function of the characteristics. There are occasions, however, when the various characteristics need to be combined in a different way.

The effect of the factors mentioned above is that the utility that we calculated for a particular person is an average which leaves room for variation. To account for this situation, the utility of the alternative *a* is written as a stochastic variable U_a, which consists of a systematic (non-stochastic) component V_a that represents the observed characteristics of alternative *a*, and a stochastic component e_a (a so-called error term):

$$U_a = V_a + e_a$$

The error term e_a has a probability distribution with an expected value (mean) equal to zero. This means that the utility U_a has a probability distribution with a mean value equal to the known, observed value V_a.

The probability Pr(*a*) of a randomly picked person choosing the alternative *a* is equal to the probability that the utility of alternative *a* is greater than the utilities of all the other alternatives:

$$\Pr(a) = \Pr(U_a > U_k) \quad voor\ alle\ k \neq a$$

Example:

Imagine a choice situation with two alternatives:

$$U_1 = 3 + e_1$$
$$U_2 = 2 + e_2$$

The probability that the first alternative is chosen is as follows:

$$\Pr(1) = \Pr(U_1 > U_2) = \Pr(3 + e_1 > 2 + e_2) = \Pr(e_1 - e_2 > -1)$$

Assume a uniform probability distribution for e_1 between -2 and +2 and that $e_2 = 0$.

We then find that:

Pr(1) = 0.75 en Pr(2) = 0.25.

The example clearly shows that the probabilities of the various alternatives will depend on the probability distribution of the error terms. The question arises what probability distribution we should adopt for the error terms.

One could adopt a so-called Multivariate Normal Distribution for the error terms. A Multivariate Normal Distribution is a distribution where the variances may differ and where there may be mutual dependency between the probability distributions. This is the most general case. A model using the Multivariate Normal Distribution for the error terms is called a *probit model*. The disadvantage of a probit model is that it is not possible to write the results in a closed analytical form and, if one wants to apply it, simulation techniques (Monte Carlo techniques) need to be used. The calculations, moreover, become highly involved if the number of choice alternatives increases.

The best-known discrete choice model, the *logitmodel*, arises when one makes the following three simplifying assumptions:

- *First assumption*: Approach the Normal distribution by a *Gumbel distribution*. This probability distribution is very similar to a Normal distribution. However, the mathematical form is simpler than that of a Normal distribution, and amenable to analytical manipulation. This makes it easier to manipulate it in an analysis. The cumulative density function for the Gumbel distribution is:

  $$F(x) = \exp(-\exp(-m(x-h)))$$

  In this formula ? is the modus (the highest point) in the distribution, and μ is a dispersion parameter.

  The mean of the Gumbel-distribution is:

  $m = h + g / m$ $\quad$ g is Euler' s constant ≈ 0.577

  The variance is:

  $s^2 = p^2 / 6m^2$

  The mean of the error term m = 0, so h = -g / m
  Figure 3-1 shows the Normal distribution and the Gumbel-division, both normalised with means equal to 0 and variances equal to 1.

- *Second assumption*. Assume that the probability distributions for all error terms are *identical*, in other words, that all have the same variance.
- *Third assumption*. Assume that the probability distribution for all the error terms are statistically *independent*.

With these assumptions, it can be shown that the probability of choosing alternative *a* out of a total of *K* alternatives is:

$$\Pr(a) = \frac{e^{mV_a}}{\sum_{k=1}^{K} e^{mV_k}}$$

We call this formula the *multinomial logit model* or, in short, the logit model. The name "logit" derives from the so-called logistic function, which has an S-shaped graph. We will come across it further on in this chapter. The formula shows that the probability that alternative a is chosen, depends on the observed utilities of the alternatives, and also on the dispersion parameter μ. Because all utility components are multiplied by the same constant factor, the dispersion parameter μ can, in practice, not be estimated separately. We can give μ an arbitrary value. If we let $\mu = 1$, the logit model becomes:

$$\Pr(a) = \frac{e^{V_a}}{\sum_{k=1}^{K} e^{V_k}}$$

Note that the utilities have been scaled by a factor of $1/\mu$ compared to of the utility components V, with which we started the derivation of the logit model.

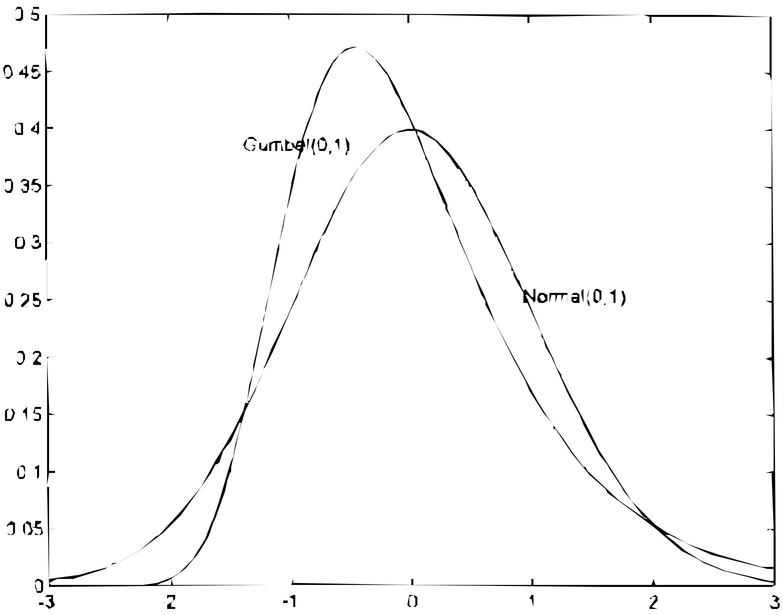


Figure 3-1 Normal amd Gumbel distributions (normalized)

Suppose that one wants to apply the logit model to the problem of mode choice. In that case, the symbols have the following meaning:

Pr(a)	=	the probability that a will be chosen.
V_k	=	the observable utility of travel mode k
K	=	the number of alternative travel modes

If $K = 2$, this is called a binary logit model and if $K > 2$ it is called a multi-nomial logit model.

The observable utilities V_k are, as we said, a function of the characteristics of the alternatives and of the personal characteristics. For this function, one usually takes a linear function as illustrated in the following example:

Example

Imagine a situation in which one can choose between three transport modes: car, bus and bicycle. Assume that the observable utilities V_k for a particular group of people (who have the same personal characteristics) can be given by the following functions:

$$\begin{array}{lllll} V_{car} & = 1.0 & & - 0.15 * K_{car} & - 0.10 * T_{car} \\ V_{bus} & = & & - 0.15 * K_{bus} & - 0.10 * T_{bus} \\ V_{bicycle} & = & -0.5 & & - 0.10 * T_{bicycle} \end{array}$$

In this example, T and K are, respectively, travel time and travel costs, and they have the following values:

	Car	Bus	Bicycle
T (min)	5	15	20
*K (BEF*100)*	0.20	0.17	-

Now we can calculate the probabilities that a particular travel mode will be chosen by individuals in this group as follows:

$$V_{car} = 0.47 \qquad V_{bus} = -1.53 \qquad V_{bicycle} = -2.50$$

$$\begin{array}{lll} \text{Pr(car)} & = e^{0.47} / (e^{0.47} + e^{-1.53} + e^{-2.50}) & = 84.3\ \% \\ \text{Pr(bus)} & & = 11.4\ \% \\ \text{Pr(bicycle)} & & = \ \ 4.3\ \% \end{array}$$

3.2 Specification of a logit model

The example above the functions for the observable utilities V_k were given. Determining these functions (this is called the specification of the model) is far from trivial. We can distinguish a number of phases in the specification process:

3.2.1 Functional form

It is usual to apply a linear function of the characteristics. These functions give reasonable results in practice. The advantage of a linear function is that the estimation of the parameters from the observations is easier. Specification of a linear function is, however, not mandatory; non-linear functions are also applicable.

3.2.2 Variables

Two kinds of variables occur in the functions of the observable utility:

- *Generic variables* are variables with the same coefficient (unequal to zero) in all functions for the various alternatives. In the example above, the travel time is a generic variable. The basic assumption here is that one minute of travel time has a similar impact on the utility, whether it concerns a minute of time travelled by car, bus, or bicycle.
- *Alternative-specific variables* are used in one or more, but not in all utility functions. They have the same coefficients in the utility-functions in which they are used. In the example above, travel costs is an alternative-specific variable. An alternative-specific variable is often used in the utility function of only one alternative. If we had, in the example above, reason to suppose that the travel time of the three alternatives exerted different influences on the choice, we would have specified three different alternative-

specific variables for the travel times of the car, bus and bicycle, and different coefficients would have been estimated. Constants are a special form of alternative-specific variables (also called alternative-specific constants). In the example above we two alternative-specific constants are used.

The example only used characteristics of the alternatives as variables. We worked on the premise that the functions were estimated for a group of individuals with the same personal characteristics. In general, however, the personal characteristics of the individual who makes a decision could be used in the utility functions. For example, socio-economic variables such as income, age, etc. Because the socio-economic variables for one individual will not differ across the various alternatives, they need to be specified as alternative-specific variables.

3.2.3 Calibration

When the functional form and the variables in the functions for the observable utilities have been determined, we can, with the aid of observations, estimate the parameter values (coefficients). There are standard statistical procedures for this purpose. The estimation of the parameters of traffic models is not dealt with in this introductory course.

3.3 Graphic illustration of the binary logit model

Suppose that there is a choice between two alternatives: alternative 1 and alternative 2. This implies for a binary logit model that:

$$\Pr(1) = \frac{e^{V_1}}{e^{V_1} + e^{V_2}}$$

Divide the numerator and the denominator by exp(V_1) to get:

$$\Pr(1) = \frac{1}{1 + e^{-(V_1 - V_2)}}$$

We now plot Pr(1) as a function of (V_1-V_2). See Figure 3-2.

The curve that emerges is called a logistic curve. As expected, the probabilities for both alternatives are equal to 0.5 when the observable utilities are equal (i.e. when V_1-V_2 = 0. If V_1 is large compared to V_2 (i.e. if V_1-V_2 >> 0). Pr(1) approaches 1 asymptotically.

When the reverse is the case and V_1 is much smaller than V_2, Pr(1) approaches 0.

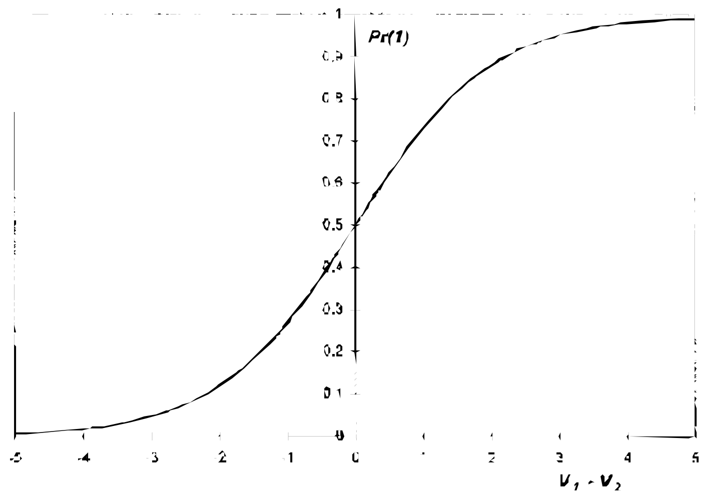


Figure 3-2 The logistic curve

3.4 Aggregate and disaggregate models

The logit model discussed above is an example of a disaggregate model. This means that the model describes the behaviour of individuals, or of groups of individuals, who share the same personal characteristics. The majority of models that we discuss in the other chapters of this course are aggregate models. These models use averages for large groups of individuals, for example over a zone.

Although a disaggregate model enables us to calculate choice probabilities on an individual basis, we are really interested in the prediction of travel behaviour for an entire area. Because the logit model is non-linear, aggregating individual probabilities to probabilities for an entire area is a non-trivial exercise. It is not correct to simply accept the average values of the explanatory variables across the entire area. We can illustrate this by an example. See Figure 3-3.

The correct average value for a group of 2 people A and B is (Pr(A) + Pr(B))/2. If we take the average values of the explanatory variables we get the incorrect value of Pr((V_A + V_B)/2), which is indicated by point C in the figure. To deal with the problem, a number of procedures have been proposed. One possible solution is to reduce the aggregation error by the introduction of a classification into person types.

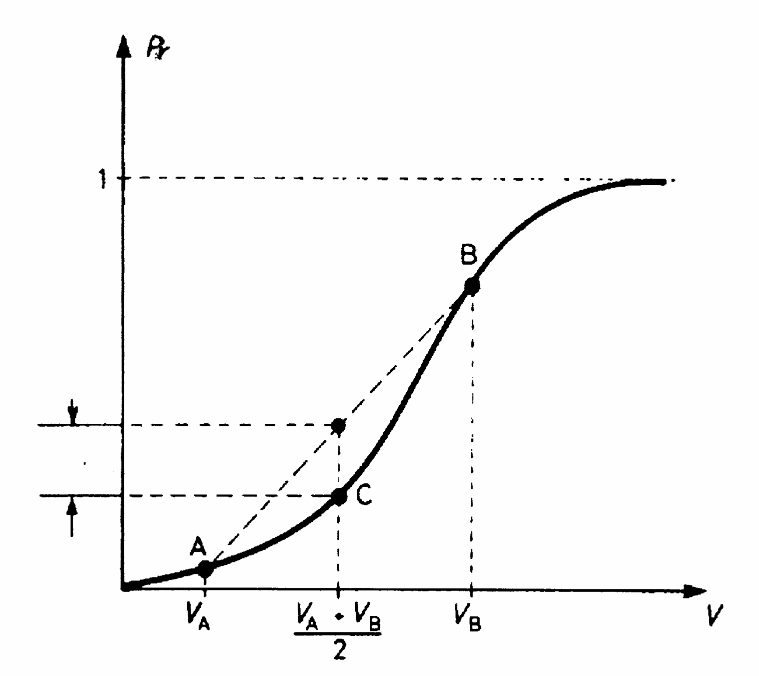


Figure 3-3 Error made when aggregating by means.[2]

When applying the logit model it is always advisable to use disaggregate data. If these are not available then there is no other option than to work with average values for the observable utilities over a whole geographical area. The disadvantages of such a procedure have been explained above.

3.5 Restrictions of the logit model

The assumptions we made in the derivation of the logit model, namely identical error terms with the same variance and mutually independent error terms, have led to an easily manageable model. However, if these assumptions have not been met, the application of the logit model will lead to incorrect results.

We will illustrate this by two examples. The problem in both examples is one of route choice. Since the logit model can, in principle, be used for all kinds of choice situations, it should be noted that the problems could also occur in other contexts.

Example 1 (non-identical error terms)

When we looked at the binary logit model, we noted that the probabilities for both alternatives are a function of the difference of the observable utilities of the alternatives. Suppose that one writes the observable utility in a route choice model as a linear function of the travel times of the alternative routes. The logit model would then show that the probabilities regarding the choice of the various routes that could be taken, are functions of the difference in travel time between the alternatives.

Observe the route choice problem in Figure 3-4. In the first case (Figure 3-4a), the travel times along both routes are respectively 5 and 10 minutes. In the second case, travel times are 125 and 120 minutes. The difference in travel time via the alternative routes is, in both cases, 5 minutes. Here, the logit model would predict that, in both cases, the traffic would spread itself over the alternative routes in the same proportions. This leads to an illogical result, for one would not expect that a difference in travel time of 5 minutes on a total journey of 5 to 10 minutes would give the same result as a similar difference of 5 minutes, but now over a journey of 2 hours.

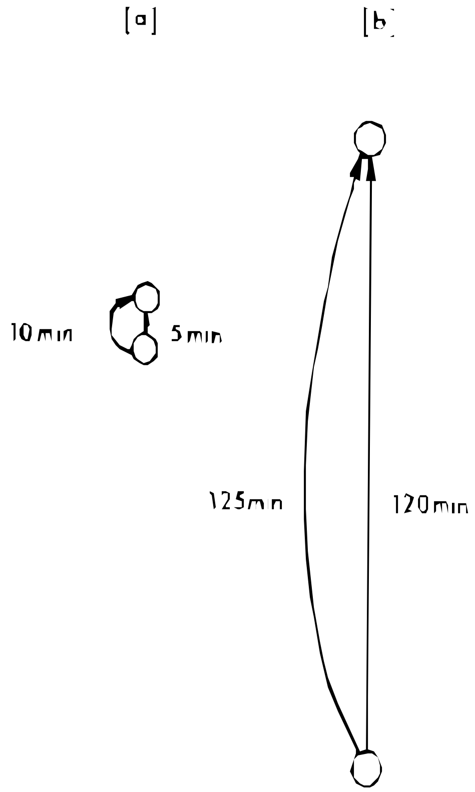


Figure 3-4 Logit route choice(1).[5]

The cause of the incorrect result lies in the fact that in the derivation of the logit model it is assumed that the variances of all the error terms are identical. This is not so in this example. The variation in the perception of travel times for long journeys will be larger than for short journeys.

Example 2a (error terms not statistically independent)

In the network of Figure 3-5 there are three routes between O and D. The travel time along each route is 1 hour. The lower two routes overlap with one another. The degree of overlap is indicated by r. Since the travel times along the three routes are equal, the logit model predicts that the traffic will divide itself in equal proportions over the network. Each route gets 1/3 of the traffic. If r is small, in other words, if the routes show little overlap, as indicated in Figure 3-5b, then the result appears to be reasonable, because the travellers will judge the three routes as more or less equal alternatives. But when there is a significant overlap, as in Figure 3-5c, this is no longer the case. Half of the traffic will probably choose the highest route, and the rest will divide itself into two parts, each quarter taking one of the lower routes.

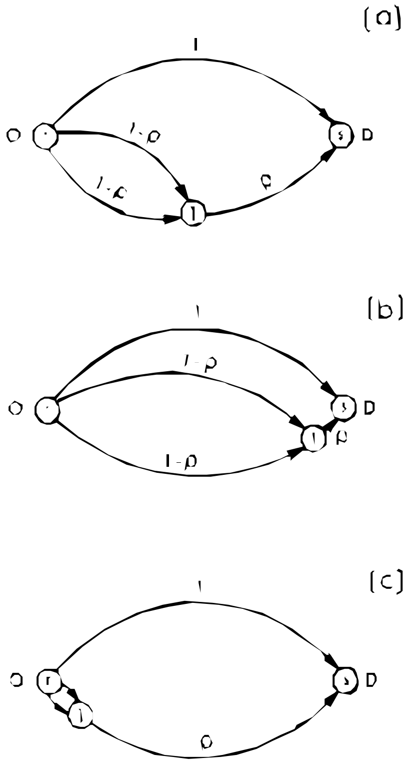


Figure 3-5 Logit route choice (2).[5]

The cause of the incorrect results here lies in the fact that the error terms for the alternatives are not statistically independent, although this is implied in the derivation of the logit model. Because both lower routes overlap significantly, the error terms are in fact strongly correlated.

Example 2b (error terms not statistically independent)

Another example where the logit model gives incorrect results is the so-called blue bus-red bus problem.

Assume that 50% of travellers in a particular city choose to take the car, and the other 50% the bus. This means that the utilities of both transport options are on an equal footing. Now suppose that the bus company decides to paint half of the busses blue and the other half red. The traveller now has three options, namely the car, the blue bus and the red bus. The utilities of the three alternatives are equal. The logit model will predict that each of the three alternatives has one chance in 3 of being chosen. Thus, the share held by the car would lower from 50% to 33%. In reality the percentage of the car remains at 50% ,of course, and the other travellers will divide themselves evenly over the blue and red busses.

Just as in the route choice problem of example 2a, the erroneous result is caused by the fact that the error terms are not statistically independent. The two bus-alternatives are, in fact, identical, and their error terms are, therefore, completely correlated.

3.6 Hierarchical logit models

3.6.1 Simultaneous or sequential choice

The previous paragraphs have shown that the logit model is a good and simple instrument by which to analyse choice situations, but that we must be aware of its limitations.

The logit model gives incorrect results in the following cases:

- When the alternatives are not independent (the error terms of the alternatives are correlated in that case);
- When the variances in the perception of the alternatives differ strongly from each other (the error terms of the alternatives are then not identical).

If we are to use the logit model in its simplest form we must, therefore, ensure that the alternatives are observed as clearly different and independent possibilities. There must also be sufficient reason to suppose that the variations in the perception of a certain alternative will not diverge too much from the variations in the perception of the other alternatives.

If the conditions stated above are not met we can use a probit model instead of a logit model. The probit model uses a Multi-Variate Normal Distribution and the limitations of the logit model do not apply. The disadvantages of the probit model, however, are that it can not be written in a mathematically closed form and that, for a large number of alternatives, the calculations are very laborious. In addition, its application requires data regarding the co-variance between the error terms of the alternatives, and these data are often not available.

Another approach that enables us to continue to use the advantages of the logit model, is the application of the so-called *hierarchical logit model.* The hierarchical logit model is suitable in those cases in which we suspect that a number of alternatives are correlated. In that case, the choice process is divided over a number of phases or levels. It is necessary that alternatives on the same level remain sufficiently distinguishable, in order that a logit model can be applied on this level. Because the choice between alternatives *on the same level* happens simultaneously, one speaks of a *simultaneous* choice. It is now assumed that the choices on the *different levels* happen one after the other. In other words we are dealing with a *sequential* choice structure. We will illustrate this basic idea, one that comes down to the application of conditional probabilities, using a number of examples.

One could, in the example 2a, which was discussed in the previous paragraph, choose, first for the highest or lowest route, and only afterwards, if one had decided on the lowest route, on sub-route 2 or 3.

In the case of the blue bus/red bus problem, the choice is primarily one between car and bus. One decides on the blue or red bus only when one has first chosen in favour of the bus per sé.

An hierarchical logit model can also be used in the following case. Suppose there is a choice of a journey by car, bus or train. This could, to start with, be seen as a choice between private transport (car) and public transport (bus and train). The choice whether to take the bus or the train only occurs in the second instance, when the decision on public transport has already been made.

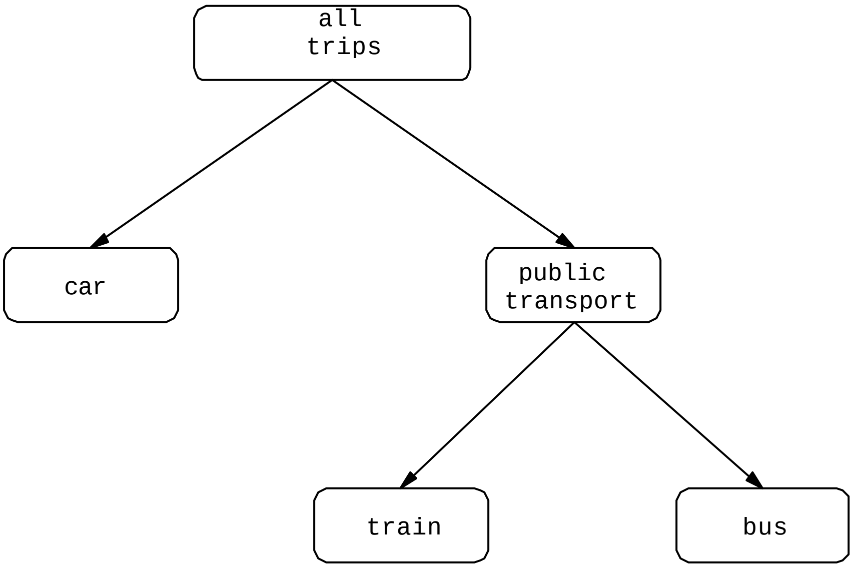


Figure 3-6 Sequential choice structure

An hierarchical logit model for the last example can be represented in a schematic way as shown in Figure 3-6. The alternatives bus and train can actually be seen as a choice in favour of public transport. They are, therefore, brought together in a separate level. A logit model is used at the highest level in order to determine the shares taken by the car and by the public transport. In order to do this one must give the alternative "public transport" a utility value. The utility for the public transport is a function of the utilities that have been distinguished for the bus and train separately. The utility of the public transport sector thus calculated is called a *combined utility.*

3.6.2 Calculation method hierarchical logit model

We will illustrate the calculation method of the hierarchical model, using a problem of destination choice and travel mode choice.

The set of destinations is indicated by *D*, the set of available travel modes to a particular destination *d* by M_d. Suppose we have reason to believe that someone who is about to travel decides first for destination *d* Î *D* and then for travel mode *m* Î M_d to reach this destination. In this case the choice structure looks like Figure 3-7. The total of available alternatives is subdivided in a number of subsets. These subsets consist of a number of alternatives that share certain characteristics so that one can expect the utilities to show some correlation. These subsets are also called "nests". This is why an hierarchical logit model is sometimes called a *nested logit model.*

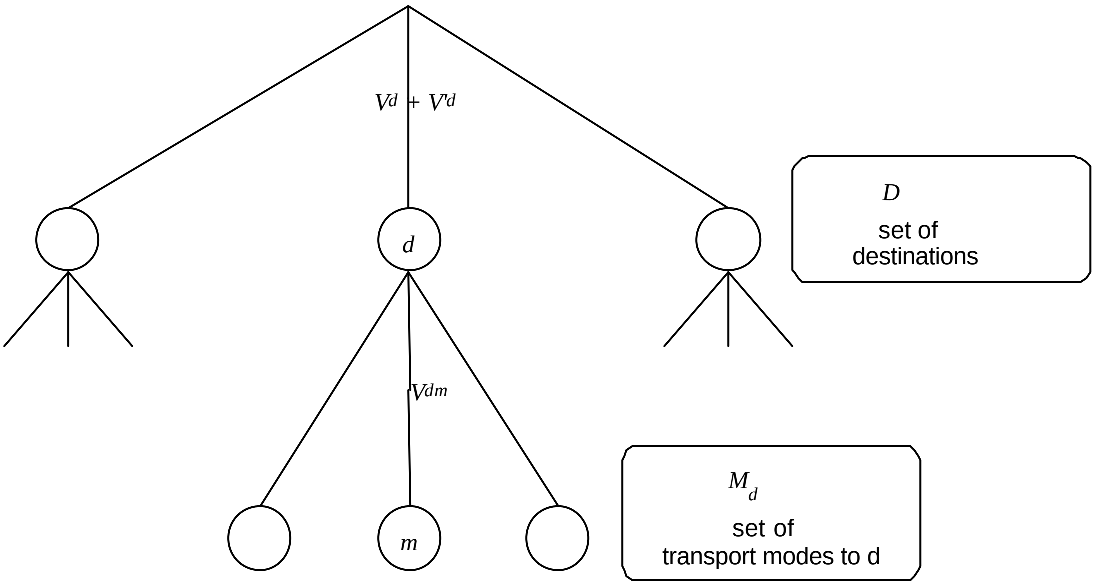


Figure 3-7 Calculation method hierarchical logit model.

We envisage the total observable utility of a destination-mode of transport combination to be composed from two components:

V_{dm}: a component in the observable utility that varies with traffic mode *m*, also inside a certain choice for the destination *d*.

V_d: a component of the observable utility that is independent of the chosen mode of transport. The only determining factors are the characteristics of the destination *d*.

From now on we will assume that utilities can be added up. The total utility of an alternative is equal to the arithmetical sum of the utility components of that alternative.

The conditional probability of choosing traffic mode *m*, given that destination *d* has already been chosen, is given by:

$$\Pr(m/d) = \frac{e^{m_2 V_{dm}}}{\sum_{m' \in M_d} e^{m_2 V_{dm'}}}$$

If we want to apply a logit model to the destination choice at the highest level of the hierarchy, we must assign an observable utility to each of the destinations. This observable utility consists of the component V_d, and also of a *combined utility*, a sort of replacement value that characterises M_d, the set of available travel modes to a certain destination *d*. To this end, the most obvious solution would be to assign the maximum of the observable utilities V_{dm} to the destinations concerned. One must remember, however, that utilities are stochastic variables. If we assume that the error terms in a nest *d* show a Gumbel distribution with a dispersion parameter of m_2^d , it can be shown that the expected value $V¢_d$

of the maximum of the observable utilities of all modes of transport to a certain destination *d* are given by:

$$V'_d = \frac{1}{m_2^d} \ln \sum_{m' \in M_d} e^{m_2^d V_{dm'}}$$

The probability that destination *d* will be chosen can now be written as:

$$\Pr(d) = \frac{e^{m_1(V_d + V'_d)}}{\sum_{d' \in D} e^{m_1(V_{d'} + V'_{d'})}}$$

The probability that the combination of destination and travel mode (*d*,*m*) will be chosen is equal to the probability of destination *d* multiplied by the conditional probability of travel mode *m*, given destination *d*:

$$\Pr(d,m) = \frac{e^{m_1(V_d + V'_d)}}{\sum_{d' \in D} e^{m_1(V_{d'} + V'_{d'})}} * \frac{e^{m_2^d V_{dm}}}{\sum_{m' \in M_d} e^{m_2^d V_{dm'}}}$$

The dispersion parameters m_1 *and* m_2^d have been maintained in these formulas because they are not necessarily equal. This plays a role when combining two sequential logit models, and is expressed by the exponents in the first factor of the formula above, where m_1 is multiplied by $V¢_d$.

It is not possible to estimate the distribution parameters m_2^d and the utilities V_{dm} separately. Thus we set $m_2^d = 1$, which means scaling V_{dm} by a factor of $1/m_2^d$. We also set $q_d = m_1 / m_2^d$, which leads to the following:

$$\Pr(d,m) = \frac{e^{q_d(V_d + LS_d)}}{\sum_{d' \in D} e^{q_d(V_{d'} + LS_{d'})}} * \frac{e^{V_{dm}}}{\sum_{m' \in M_d} e^{V_{dm'}}}$$

Here:

$$LS_d = \ln \sum_{m' \in M_d} e^{V_{dm'}}$$

The above expression is called a *logsum*.

Finally, it is not possible to separately estimate both factors in the expression $q_d\ V_d$. We assume a scaling of V_d with a factor of $1/q\ _d$. This eventually leads to the following formula for the hierarchical logit model:

$$\Pr(d,m) = \frac{e^{V_d + q_d\ LS_d}}{\sum_{d'\in D} e^{V_{d'} + q_d\ LS_{d'}}} * \frac{e^{V_{dm}}}{\sum_{m'\in M_d} e^{V_{dm'}}}$$

In this formula the q_d are empirically determined parameters. If we now calibrate the hierarchical logit model with the available observations we must determine both the values of the coefficients in the functions for the observable utilities V_d and V_{dm} and the values of q_d.

To gain consistent results it can be shown that the following must be the case:

$m_1 \leq m_2^d$ or, equivalently: $0 < q_d \leq 1$.

Since *m*, the distribution parameter in the Gumbel distribution, is inversely proportional to the variance, the variance in the error terms at the first level of choice is larger, or at most equal to that on the second choice level.

If one finds $q\ _d = 1$ for all nests *d*, then the hierarchical logit model is algebraic equivalent to the normal (non-nested) multi-nomial logit model. Because in that case we get:

$$\Pr(d,m) \quad = \quad \frac{e^{V_d} \cdot \sum_{m'\in M_d} e^{V_{dm'}}}{\sum_{d'\in D} (e^{V_{d'}} \cdot \sum_{m'\in M_{d'}} e^{V_{d'm'}})} * \frac{e^{V_{dm}}}{\sum_{m'\in M_d} e^{V_{dm'}}} \quad = \quad \frac{e^{(V_d + V_{dm})}}{\sum_{d'\in D, m'\in M_{d'}} e^{(V_{d'} + V_{d'm'})}}$$

This means that the utilities allocated by the individual inside a subset are not correlated and that, therefore, a sequential choice model should not be used. We are then dealing with a simple simultaneous choice model.

A value of $0 < q\ _d < 1$ means that the utilities in a nest *d* are correlated, and the more so as $q\ _d$ approaches zero.

If, lastly, we find that $q_d \leq 0$ of $q_d > 1$ for a nest *d*, it indicates that the postulated model with its subdivision into nests is incorrect. We must then try another choice structure.

If the choice structure consists of a large number of nests, it may be difficult to estimate a separate parameter q_d for each nest. For simplicity sake it is therefore assumed that the distribution parameters m_2^d for all nests *d* at the same level are equal. In that case the index *d* may be left out of the parameter q_d in the formulas above.

We will look at the problem of simultaneous and sequential choice structure again in the chapter on traffic mode choice.

3.7 Summary

The logit model, which is one of the models of discrete choice theory, is very suitable to analyse choices in traffic engineering. The models are based on the assignment of values to characterise the attractiveness of each of the alternatives. The valuations are called utilities. Since we do not know all of the characteristics that define the utility of an alternative, the utility has a stochastic component. The choice distribution over the alternatives is given in terms of probabilities that can then be aggregated over an area or population group. In its simplest form, the logit model has a number of limitations, which can be partly overcome with the aid of a sequential choice structure.

Since the logit model will often be referred to in later chapters, we dealt with it early in this course.

4 Zones and Networks.

A traffic demand model applies to a particular geographical study area. In principle, trips in this area can begin and end at any address, and travellers can choose from all roads, streets and other transport options. Because of the sheer volume of data, however, it is not practical to gather and analyse data based on individual information. We construct a simplified model of reality by introducing the following elements:

- *Zones*; the area to be studied is divided into a number of zones. We study the trips from and to these zones. We assume that all trips begin and end at an imaginary point inside this zone, which is called the centroid of that zone.
- *Networks*; the transport system consists of a number of networks, that represent the available transport modalities. The network is an abstraction of reality. The detailed level of representation depends on the problem to be solved.

Since the design of an outline of the study area into zones and networks strongly depends on the problem to be solved, it is not possible to hold to stringent rules. The intention behind this chapter is to give some general guidelines that may be helpful in the outline design.

4.1 Area zoning

We distinguish the *study area* and a surrounding *area of influence*. Both areas are divided into zones, called respectively the internal and the external zones. In the study area we investigate the traffic flows from and to each zone. As for the area of influence we only examine traffic flows that start or end inside the study area. When traffic moves between two external zones, we only look at the traffic that crosses the study area.

Important parameters are the number of zones to be used and their size. Each zone has a fictitious point, usually situated in the point of gravity of the area, from which all trips from and to the zone are supposed to depart and arrive. This point, called the *centroid* is linked to the network by *connectors*. Trips between two zones, the *interzonal* traffic, occur on the network. Traffic that does not leave the zone, the *intrazonal* traffic, has its departure- and arrival point in the same centroid and is not analysed.

This means that zones must not be too large. If they are too large, a sizeable part of the traffic does not appear on the network and will, therefore, fall outside the analysis. Nor can zones be too small. Small zones require numerous input data. This increases the costs of the study, hampers interpretation of the results and increases the chance of mistakes.

For urban and provincial studies practice has shown that zones with a population of 1000 to 2000 people work reasonably well. Nevertheless, it is possible to deviate from this value, depending on available funds, the scale or the goal of the study.

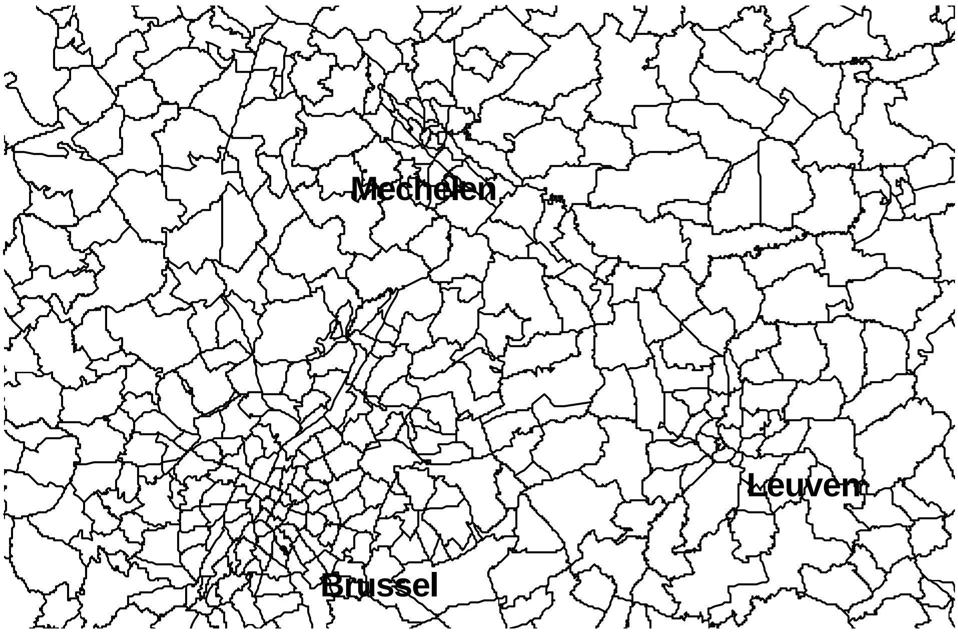


Figure 4-1 Zoning in the model for Flemish Brabant.

The following values can be taken as guidelines for the number of zones. Large urban or regional studies typically have 300 to 500 zones. The provincial traffic demand model for Flemish Brabant (see Figure 4-1) comprises about 1000 zones. This may be considered to be a large number. Fifty to 100 zones are enough for small-scale studies.

In terms of traffic production, zones need to be about of equal size. One needs to aim, moreover, for homogeneity in the determining factors of traffic production and attraction. Since land-use largely determines traffic production and attraction, it is advisable to divide the zones in such a way as to achieve optimal homogenous land-use inside a zone.

The zone borders should coincide with the borders of administrative units. Examples are districts used by National Institutes for Statistics, voting districts, municipalities, counties or provinces. Using such administrative units eases access to socio-economic data. Zone borders should, preferably, also coincide with natural barriers such as rivers, canals, railways, etc. Since there are only a limited number of places where these natural barriers can be crossed, the comparison of model results with field counts is simplified. The shape of the zones should be as compact as possible, because this limits the number of mistakes in the calculation of distances.

When dividing the area of influence into zones it is usual to increase the size of the zones as a function of the distance to the study area. Since most trips will travel over a relatively short distance, the number of relevant trips between the area of influence and the study area will rapidly decrease as the distance increases.

Only one zone-division usually suffices for all stages in the traditional traffic model. When available time and funds allow, one may deviate from this procedure. It has been shown, for example, that an adequate modelling of public transport benefits from a more detailed study area. The provincial model of Antwerp, for example, shows nearly every public transport stop, especially in the town itself. In such cases, an hierarchical division into zones and sub-zones would be applicable.

4.2 Networks

The transport system is represented by a network comprising of *nodes* and *links* that connect the nodes (see Figure 4-2). The network model is a simplified reproduction of the real network. The network is used to calculate the travel times between points of origin and points of destination. The calculated results of the traffic model can, additionally, be reproduced on a network graph.

When dealing with modalities such as car, bicycle and walking, the model networks are immediate derivatives of the physical network. The links of the network represent the roads. The nodes of the network are the intersections. Nodes in the model network are also used to mark changes in road types and the sites, for example, of bridges and other specific infra-structural facilities.

Attributes that characterise the network are assigned to the links. Examples of link-characteristics are length, speed, travel time, capacity, etc. No characteristics are assigned to the nodes of the model network. Specific characteristics of intersections, such as long waiting times for some exits or the prohibition to use certain turns, can be modelled by adding extra (dummy) links.

In contrast to a road network, a public transport network represents the physical infrastructure, as well as *lines* that represent the services carried out on that network. In a model network these lines are defined as a series of consecutive links to which variables such as frequency, capacity and travel times are assigned. Stops, feeder networks to and from stations and bus stops, and points of transfer must also be defined. (see Figure 4-3)

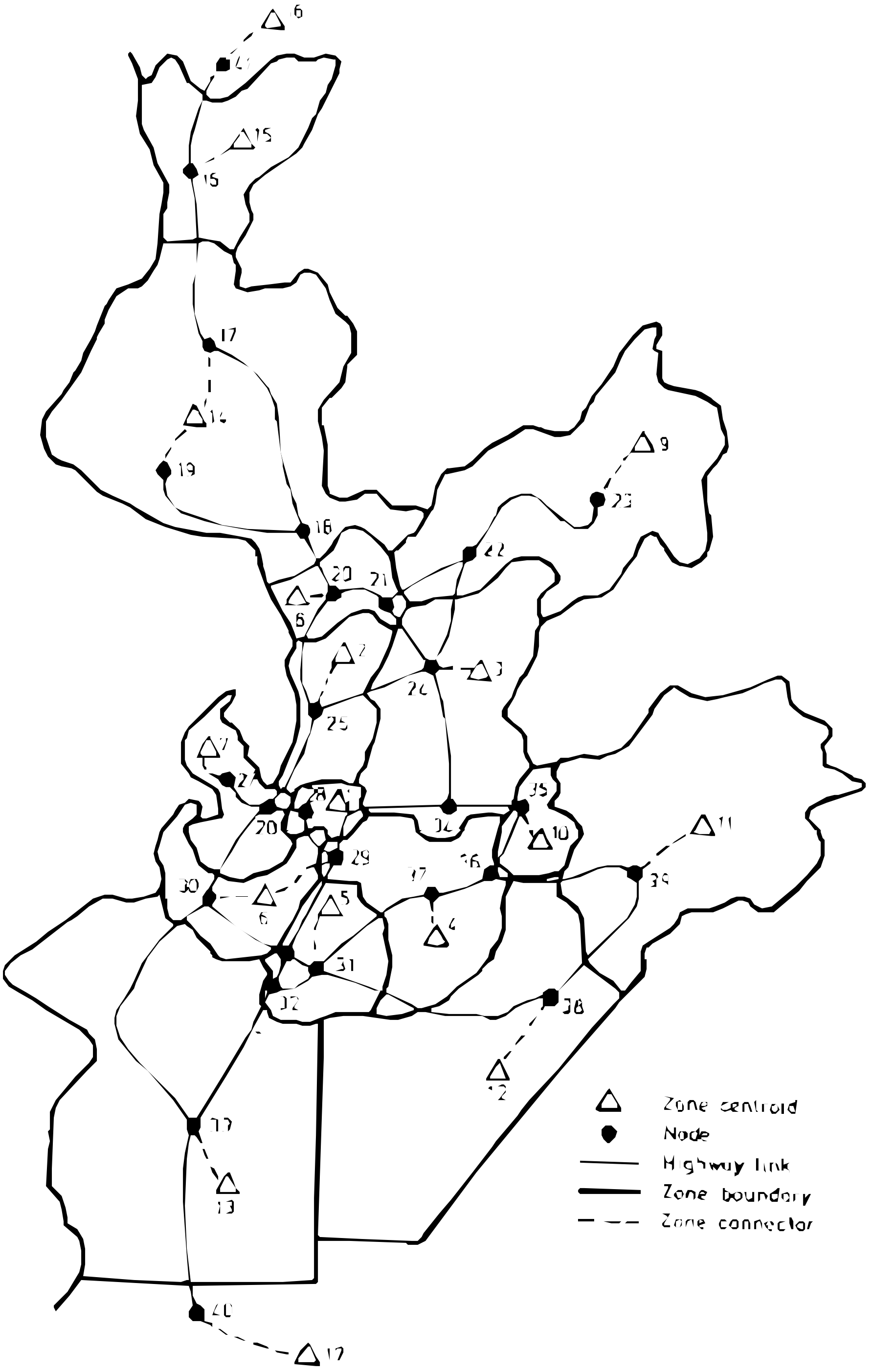


Figure 4-2 Outline of zones and road network.[2]

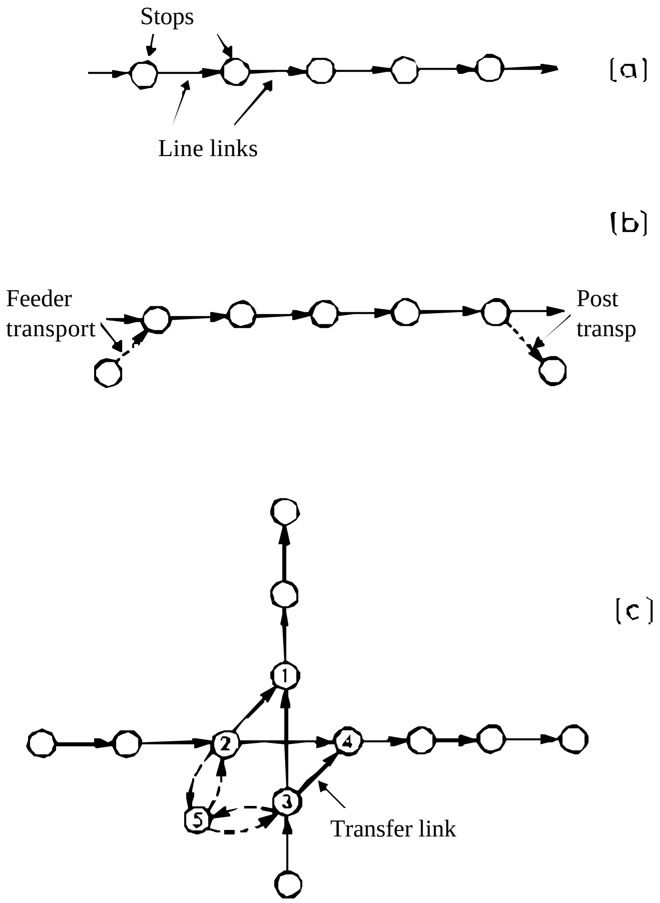


Figure 4-3 Public transport network.[5]

These days it is usual to draw up separate networks for the different transport modalities. This means that one implicitly assumes a trip to be completed via one transport mode. There is a tendency, however, to switch to the application of so called *multi-modal network models*. Here, the networks for the different transport modes are linked via *transfer links.* This means, for example, that a trip consisting of a car journey to the station followed by a train journey, can be adequately modelled.

As we noted in the discussion of the area division, each zone has a centroid. Centroids are points where the traffic of the zone in question enters or leaves the network. Each centroid is linked to the rest of the network by one or more connectors. The connectors are a schematic representation of the local street pattern inside a zone. Only the traffic that originates from the zone in question can use it. Ongoing traffic between two other zones is not considered to use the connectors of an intervening zone. It is important to link the connectors to the network in such a way as to imitate the real situation as close as possible.

A model network is a *directed network*. This means that every link has a direction assigned to it. A road with two-way traffic, therefore, is represented by two links. Even when some computer models show only one link on the screen, the road is presented by two network links internally.

It is usually unnecessary to reproduce the smallest detail in the model of the physical network. Large networks require many input data and are, therefore, expensive. The probability of errors, moreover, increases. Computing optimal routes takes a lot of computer time for large networks. Computing time increases with the number of nodes raised to a power between 2 and 3.

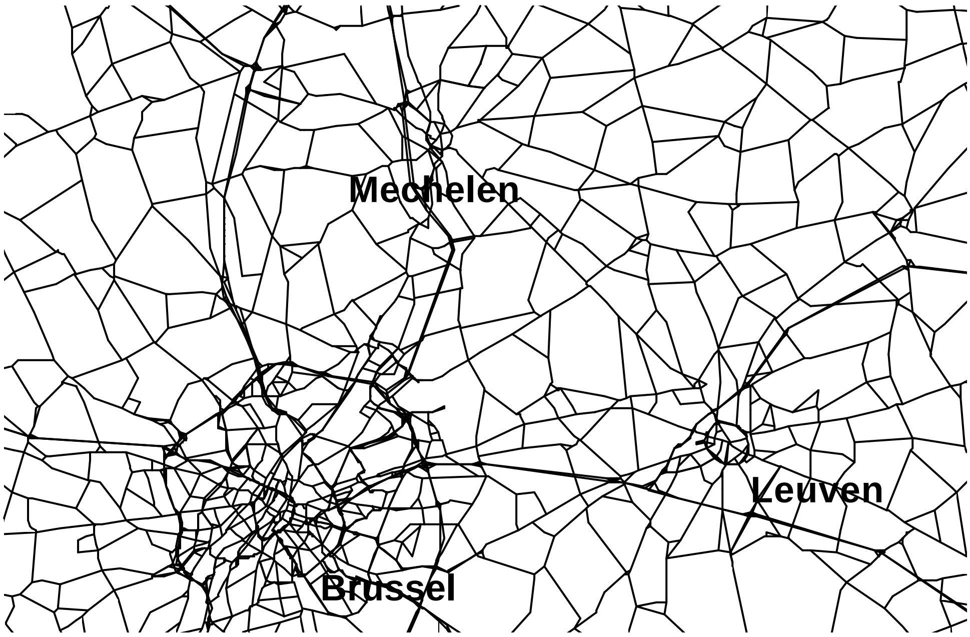


Figure 4-4 Road network in the model for Flemish Brabant

Roads in a network are usually classified according to their function. One distinguishes, for example, motorways, main roads, secondary roads, local roads and urban streets. It is advisable to also represent those links in the model that lie one level below the level of interest. If one wants to study the motor way network, for example, then one should also incorporate the main roads in the network model.

The following values can serve as a guideline for the size of the network. A typical network for an urban or region holds in the order of 1000 to 5000 nodes. The car network for the province of Flemish Brabant (Figure 4-4) has about 10000 nodes, and can thus be seen as a large network. For approximate studies networks of several hundred nodes are sufficient.

5 Production and attraction

The goal of the production/attraction phase in the classical traffic demand model is to predict the total number of produced respectively attracted trips for each zone in the study area. The prediction is based on the socio-economic data of a zone. In this phase, two related models are used.

- *Production model*. This model calculates the total number of trips produced per zone, irrespective of the zone of destination.
- *Attraction model*. This model calculates the total number of trips attracted to a zone, irrespective of the zone of origin.

The best results in the design of good production and attraction models are achieved when a breakdown is made according to trip purpose and other characteristics. We will begin this chapter with a short explanation of some commonly used terms. Next we will look at the classification of trips. This will be followed by a discussion of the factors that influence the production and attraction of trips. We will then look at the most widely used methods to calculate production and attraction, namely regression-analysis and category analysis. The logit model also lends itself very well to the determination of production and attraction, but it is not used much yet for this purpose in practice. We will give an example to illustrate the use of the logit model for the calculation of production. The chapter will end with some general remarks about the stability of the calculated production- and attraction parameters, and on how to balance production and attraction.

5.1 Terminology

A *trip* occurs when someone moves from a place where he has undertaken a specific activity to another place where he will undertake a new activity. The starting point of the trip is called the *origin* and the finishing point the *destination*. A trip can be made using one mode of transport or a sequence of transport modes. In this context, walking is also taken as a mode of transport. We call the movement of a person between two successive points, using one mode of transport, a *journey*. Thus, a trip can consist of several journeys. A trip that originates in O and has destination D, for example, can consist of a bicycle journey from H to point 1, a train journey from point 1 to point 2, and a walking journey from point 2 to B.

Trips that begin or end at home are called *home-based* trips. All other trips are *non-home-based* trips.

A *trip chain* consists of a number of trips of which the first begins in one’s own home, while the last returns there. One could, for example, make a trip from home to work in the morning and a return trip from work back to the home in the evening. Together, they are called a trip chain. In the first trip, the residence was the point of origin and the workplace the point of destination. In the second trip, the reverse happened. Trips are undertaken for specific purposes, for example, to get to work or to go shopping. This is called the *purpose* of the trip.

The terms origins and destinations (or *departures* and *arrivals*) do not always have the same meaning as the terms *production* and *attraction*. In general, the zone of origin produces trips and the zone of arrival attracts them. There is one exception to this rule: in home-based trips, the zone in which the home is located is always considered to have produced the trips. This means, for example, that in the case of home-based work trips, even though traffic in the evening peak period travels back to the home zone, the home zone is still considered to have produced the trip.

5.2 Trip classification

5.2.1 Classification according to trip purpose

It has been found that production and attraction models yield better results if the trips are classified according to trip purpose and if separate models are then made for the different purpose categories.

For most purposes, the home is the origin or destination of a trip. Primary in these home-based trips are the trips to work, but other purposes may also play an important role, such as education, shopping and social or recreational activities. Trips that do not start at the home and that do not end there (non-home-based trips) generally represent a small part of the total number of trips and are, therefore, usually not further classified according to purpose. Thus we get the following breakdown:

	Purpose	*Abbreviatiom*
Home-based	work	HBW (home-based work)
	education shopping social / recreational other	HBO (home-based other)
Non home-based		NHB (non home-based)

5.2.2 Classification according to the departure time of the trip

Trips are divided into those that occur during peak periods (morning or evening) and those that occur outside of the peak period. The contributions of the different trip purposes strongly depend on the time of day during which the trips are undertaken.

The trips that are undertaken with the purpose of work or education, usually occur during peak periods and are called *mandatory trips*. Trips that are undertaken for shopping purposes, social/recreational purposes and other purposes, are less obligatory and are, therefore, called *optional trips*. This means that one can decide not to make the trip at all. Even if one does decide to make the trip, there is more freedom in the choice of departure time.

5.2.3 Classification according to personal characteristics.

Because trip behaviour is strongly influenced by socio-economic factors, a classification into these factors can often be useful. The following characteristics are sometimes used for this classification:

- Level of income
- Car ownership
- Size and structure of the household

The most common classification is the one into car ownership (a classification, for example, into 0,1 or more cars per household).

5.2.4 Classification according to transport mode

The following categories can be used in a classification of transport modes used for trips:

- Walking
- Bicycle
- Car (when necessary divided into: driver and passenger)
- Public transport

It used to be quite common to develop separate production- and attraction models for the various modes of transport. This is rarely done nowadays. Today, the assignment of trips to different modes of transport is carried out in a later stage of the calculation. We return to this point in the chapter on transport mode choice.

5.3 Factors that influence production and attraction

5.3.1 Factors that influence production

The following factors influence the production of a zone:

- *Households characteristics*
  -Income
  -Household structure (number going to work, number going to school, age …)
  -Car ownership
- *Zone characteristics*
  -Land use
  -Land price
  -Residential density, rate of urbanisation
- *Accessibility*
  -Extent of transport options from the zone.
  -Quality of transport options from the zone

Zone characteristics and household characteristics have been frequently used in studies. The characteristic ‘accessibility’, on the other hand, is hardly ever used, neither in production nor in attraction models. Ignoring the influence variable of accessibility in the models means that the production and attraction of a zone are insensitive to changes in the transport system. This is a significant shortcoming in current models.

In addition to what has already been said about this subject in chapter 1.4, we find that a change in the transport system usually results in four effects on the pattern of trip-making:

- *Generative effect*: an in- or decrease in the total number of person-kilometres;
- *Distributive effect*: a different distribution of the origins over the destinations;
- *Temporal effect*: a shift in the timing of the trips;
- *Substitution effect*: a shift to another mode of transport.

These effects can occur in combination. This significantly complicates an analysis of the phenomenon.

By not including the factor of accessibility, the generative effect of a change in infrastructure is not reflected in the production- and attraction models. The distributive effect and the substitution effect càn be determined using the distribution models which will be discussed in chapters 6 and 7. The temporal effect could be accounted for by using a so-called departure-time model. Although such models have been developed, they are little used in practice.

It is likely that a change in infrastructure will have little generative effect on the mandatory trips to work and school, at least in the short term. A generative effect will, most likely, occur for optional trips.

The temporal effect of a change in the transport system is possibly more important than has previously been assumed. Congestion problems will cause many travellers to leave earlier or later. This phenomenon is called “peak spreading”. If improvements in the infrastructure lead to a reduction of congestion problems, many people will, fairly quickly, revert to the original peak period, causing renewed congestion during the peak period, thus defeating the intended effect of the improvements.

Since accessibility definitely is an important factor in determining the production and attraction of zones, there has been an effort to develop models that incorporate the generative and temporal effects caused by a change in accessibility. A number of models have been suggested, but, to date, there has been little agreement regarding the accuracy of these models. This is due to the difficulty in quantifying the concept of accessibility. For this reason, this type of model is rarely used in practice and we will not discuss it further in this course.

5.3.2 Factors that influence attraction

The following factors influence the attraction rate of zones:

- *Number of employees*
- *Land-use*
  -Industrial (type of industry, occupied area)
  -Educational facilities
  -Shops (floor area, sales)
  -Service sector (hospitals, banks, government institutions, conference centres …)
  -Recreational (sport centres, tourist- or amenity sites, theatres …)

-Storage and transfer (harbours, airports …)

- *Accessibility*

  -Extent of transport options to the zone

  -Quality of transport options to the zone

The zones that attract trips are usually the areas of employment. This is primary so because these areas offer employment, but also because these areas attract supply and other services.

As with trip production, accessibility most likely plays an important role.

5.3.3 Factors that influence production/attraction in goods transport

In the previous section we have listed the factors that influence the production and attraction in personal transport. However, goods transport also plays an important role in traffic- and transport policy. Goods traffic counts for a large part of all road-traffic. About 15% of all vehicles on the primary road network are goods vehicles. This percentage varies between 2% and 20% on urban networks. The following factors influence the generation and attraction of goods transport.

- Number of employees in a company
- Company turnover
- Built-up area
- Size of an industrial complex
- Type of company
- Accessibility of a company

5.4 Regression analysis

Regression analysis is the most frequently used method to calculate productions and attractions.

In a linear regression model we try to predict a variable *Y* as a linear function of one or more influence variables X_i

$$Y = a + b_1 X_1 + b_2 X_2 + b_3 X_3 + \cdots$$

The variable *Y* to be predicted is called the dependent variable. The influence variables X_i are the independent variables. The coefficient *a* is the constant factor of the regression equation, the coefficients b_i are the regression coefficients. When there is only one variable *X*, one speaks of simple regression. When there are several variables X_1, X_2 … , we use the term multiple regression.

When regression analysis is used for the development of production and attraction models, the X_i represent the socio-economic influence factors mentioned above, for example income, car-ownership, etc. The dependent variable *Y* represents the number of trips produced or attracted, usually subdivided according to purpose. Most production and attraction models predict the number of trips for a peak period.

The constant factor and the regression coefficients are estimated (calibrated) using socio-economic data gathered over a base year. The calibration uses the method of least squares for which computer programmes are widely available. It is assumed that these coefficients are time-invariant, so that the trips for a forecast year can be determined using the estimated regression equation and expected socio-economic developments.

5.4.1 Production

The development of regression models for production can be based on (aggregate) data per zone or on (disaggregate) data per household.

5.4.1.1 Production based on zonal data

With these models one tries to predict all trips produced by a zone, using socio-economic data that are characteristic for the entire zone. This method should only be used when one has to rely on aggregate zonal data. When more detailed data are available at the household- or at the personal level, it is clear that aggregation to zonal means or totals looses a lot of information. The method will only yield reliable results if zones are fairly homogenous in their socio-economic structure. If this is to be achieved, the zones would, generally, have to be small. This, in turn, increases the costs and complexity of the models in regard to data gathering and calibration.

It should be noted that regression models at the zonal level can use zonal totals, such as trips-totals per zone and cars-totals per zone. But zonal means, such as number of trips per household per zone and number of cars per household per zone, can also be used. The difference between the two methods is small, although the use of zone means is preferred, since this filters the influence of the zone size out of the equation (in so far as this does not appear in the other variables).

5.4.1.2 Production based on household data

We saw in the previous paragraph that a model based on data per household is preferable to a model based on zonal data. This renders the model independent of the chosen zonal break-down (both in terms of the size of the zones and their socio-economic composition). One also avoids the loss of information that comes with the aggregation of data to the zonal level.

In general, the household is taken as the unit, not the individual. This is so due to the assumption that the characteristics of a household (car ownership, for example, or income and composition) determine production, rather than the personal characteristics of an individual.

5.4.2 Attraction

As we have seen, demographic factors play a vital role when determining production. In the case of attraction, conversely, the determining factors are employment in the zone and land-use.

The most important variable for home-based work traffic is, naturally, employment in the attraction zone. As a rough approximation, every place of employment yields one home-based work trip. Information regarding employment is usually available. When this is not the case, one can estimate employment levels on the basis of land-use. A situation similar to that of home-based work trips exists for educational trips.

Estimating the attractions for the other trip purposes is often based on land-use in the attraction zone. The surface area occupied by shops, companies and institutions could express this land-use factor, or another relevant criterion that indicates the importance of the attractor.

When we discussed the regression methods to estimate production, we saw that calibration can be done at the zonal level or on the basis of households. However, when regression analysis is used to calculate the attractions, one usually uses aggregate data at the zonal level.

Lastly, it should be noted that regression equations for normal zones are not used to calculate attraction for special attraction zones of great importance (airports for example). These special zones require separate models, research or counts.

5.4.3 Problems in the application of regression analysis

In this paragraph we present some problems that can occur when applying regression analysis for the development of production and attraction models. It is not our intention to give an exhaustive list; regression analysis is a widely used statistical technique with an extensive literature, to which we refer for more detailed information.

5.4.3.1 Multi-colinearity

This phenomenon occurs when one or more of the independent variables also show a mutual correlation. Obviously one should try to choose the independent variables in such a way as to keep them maximally independent of each other, though this may be difficult in practice.

5.4.3.2 How many and which independent variables?

In general one will try to keep the models as simple as possible, in other words not to use more independent variables than is strictly necessary. Questions that play a role in the decision whether or not to incorporate an independent variable in a multiple regression are the following:

- Are there strong theoretical reasons to introduce a specific variable?
- Is the inclusion of certain variables useful in the calculation of particular policy decisions?
- Does the variable to be introduced contribute sufficiently to the explanation of the production or attraction to be estimated? A technique to test this is a so-called step-wise regression in which the variables are introduced in steps according to their contribution to the explanation of the independent variable.
- Is the future development of the socio-economic variable to be introduced itself easy to predict? If not, it is of little use in the prediction of future productions and attractions.

5.4.3.3 Non-linearities

The regression model assumes a linear relation between trips and socio-economic variables. In reality, however, this relation may not be linear. It is sometimes, though not always, possible to circumvent this difficulty with a number of different techniques. These techniques comprise a transformation of variables (for example taking the logarithms of the variable) or the use of so-called dummy variables. We will not elaborate this point.

5.4.3.4 Constant factor in the regression equation

In order to guarantee that partial results can be added, it is, in the case of regression on the basis of zonal data, necessary for the regression line to pass through the origin. In other words, the constant factor in the regression equation should be equal to zero. If this does not happen the specification of the regression model may be erroneous.

> Example:
>
> Assume that one finds the following regression equation for the number of departures in a zone:
>
> *Departures = 1080 + Population*
>
> For a zone of 2000 residents, therefore, the number of departures is 3080. If the zone is split into two smaller zones, each with 1000 inhabitants, there are 2080 departures per smaller zone and 4160 departures for the entire zone. This contradicts the result of 3080 departures found in the first instance.

5.4.3.5 Extrapolation

Strictly speaking, regression equations apply only to the range of data that were used for the calibration. When one applies regression equations for forecasting purposes, results that fall outside this range can be achieved. Some caution is, therefore, justified when using data that have been calculated in this way.

5.4.3.6 Ecological correlation

When using aggregated data a problem can occur that is known under the name of “ecological correlation”. If the data comes from several sub-populations, we may find a certain (positive, for example) correlation inside the sub-population, while, if we started from the means of the sub-population, we would get a totally different (negative, for example) correlation.

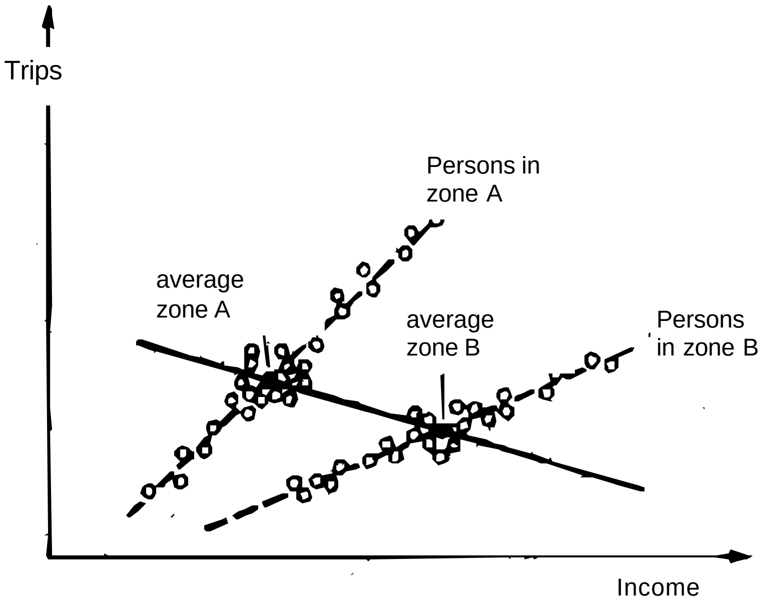


Figure 5-1 Ecological correlation.[2]

An example of this is shown in Figure 5-1. The disaggregated data inside the zone show a positive influence of income on production. If we had used means per zone we would have reached a completely different conclusion.

5.5 Category analysis

In a category analysis the population of the study area is divided into a number of homogenous groups or categories, based on specific socio-economic characteristics. The trip behaviour is determined for each of the categories, with the understanding that this will remain stable over time. If one knows the future composition of a zone in terms of categories of inhabitants, one can calculated future trip behaviour. When compared to regression analysis, this method has some advantages, but also disadvantages.

5.5.1 Production

The principle of category analysis for the calculation of production can best be illustrated by an example.

Example:

Many research projects have shown that the trip production of a household depends primarily on car-ownership, family size and composition, and the income of the household. In the example we assume a classification into three different categories of car-ownership and 4 categories of household size. Having made an inventory of the data, a category analysis could give the results shown in Table 5-1, in which the number of trips per household is given per household category.

Table 5-1 Results of a category analysis.[2]

	Car-ownership of household		
Persons in household	0	1	2+
1	0.12	0.94	
2 of 3	0.60	1.38	2.16
4	1.14	1.74	2.60
5	1.02	1.69	2.60

If a category analysis-table has been made using data for the base year, the next step is to estimate the number of households in each category in a forecast year. The total future production is then found by multiplying those numbers by the trip rates in the category analysis table.

5.5.2 Attraction

Category analysis is rarely used to calculate attraction. It would, in principle, be possible, using, for example, a classification in sectors of employment and employment densities. However, the problems associated with the gathering of sufficient disaggregated data are huge.

5.5.3 Problems in the application of category analysis

Although category analysis has some advantages such as the conceptual simplicity of the method and the fact that non-linearities are easily accommodated for, when compared to regression analysis, it also has its disadvantages.

5.5.3.1 Many calibration data required

The most important disadvantage is the need for large numbers of data. The example above is misleading in terms of the number of categories to be distinguished. In practice one will quickly distinguish three categories of car-ownership, six for incomes and six for household size and composition. This gives 3x6x6 = 108 categories.

Assuming that a minimum of 50 observations per category are required to ensure fairly reliable statistical means, a minimum of about 5000 observations would be needed. In reality, this number will be much higher because the observations will not be equally divided over the categories. In short, a small increase in the number of categories leads to a huge increase in the data required.

5.5.3.2 Which categories now and in the future?

Choosing suitable homogenous categories is a difficult task, although so-called clustering methods are available for this purpose. Another important problem is calculating the future break-down of households in a zone over the various categories. There are no really satisfactory solutions to this problem.

5.6 Use of logit model for calculation of production.

The total number of trips made in an area is the consequence of a large number of individual choices. If we confine the choice set for individuals to the two options whether to make a trip or not, we can use a binary logit model. We shall illustrate this with the following example that was taken from the TransCad manual.[4]

Example

Household characteristics, zone characteristics and the accessibility of the zone determine the production of a zone. These last two characteristics are constant for one specific zone. The production for this specific zone is, therefore, a function of household characteristics only. As noted before, the household is usually taken as a unit, but in this example, we will take the individual as a unit.

We now want to calculate the probability that someone chooses to make a home-based work trip as a function of his or her personal characteristics.

The following personal characteristics are used:

A	= Age	(in years 16-90)
ED	= Education	(scale from 1-17)
G	= Gender	(0 = woman; 1 = man)
MM	= Married man	(0 = no; 1 = yes)
WM	= Married woman	(0 = no; 1 = yes)
CW	= Woman with child under 6.	(0 = no; 1 = yes)

The last four variables are so-called dummy-variables. These are variables, which can only assume the values 0 or 1. Each dummy-variable divides the population into two sub-groups. By using a number of dummy-variables, we can distinguish a number of sub-groups.

Each individual has two choice alternatives. This enables us to use the binary logit model. Choosing alternative 1 means that a trip is made, alternative 2 represents the fact that no trip is undertaken.

The probability that a home-based work trip will be made can be formulated using the following binary logit model (see chapter 3.3)

$$\Pr(1) = \frac{1}{1+e^{-(V_1-V_2)}}$$

Using a data-file, where each line notes the values of the personal characteristics for a person, and whether or not that person made a home-based work trip, we can estimate the functions for the observable utilities V_1 and V_2. Assume that we get the following results:

$$V_1 = -0.47 - 0.05 * A + 0.21 * E + 0.27 * G + 1.59 * MM + 0.31 * MW - 1.74 * CW$$
$$V_2 = 0$$

Since the socio-economic variables for one individual will not vary across the two alternatives they need to be specified as alternative-specific variables (see chapter 3.2). It does not matter in which utility function we include the alternative-specific variable. If we had, for example, included *G* in the function V_2 instead of V_1, we would have achieved the same coefficient 0.27, although prefixed by a minus sign. The end-result of the calculation remains the same because the binary logit-model uses differences in utilities.

We must ensure that all coefficients have signs that agree with our intuitive expectations. For example, the minus sign for the variable *CW* indicates that a woman with a young child is less likely to make a home-based work trip. This is so because the observable utility for the execution of a trip

> becomes smaller when *CW* has a value of 1. The values for the coefficients also enable us to calculate the effect of a change in one of the variables. Assume that the probability of an home-based work trip for someone with an education level of $E = 10$ equals 50%. Check that that this probability rises to 70% if the level of education rises to $E = 14$ and if the other variables do not change.

The logit model discussed above computes the probability that a specific individual will make a trip. However, we want to know the total number of trips made in an area. We will therefore still have to aggregate the individual probabilities, to arrive at a forecast for the entire area. See the remarks concerning this problem in chapter 3.4.

5.7 Stability of production and attraction parameters.

Production- and attraction models are developed to determine the generation of trips in a forecast year. The calibrated parameters, however, are usually based on data from a base year. How can we be certain that these parameters are invariant in time so that they will also apply to future situations?

Research shows that trip behaviour remains fairly stable when the time horizon is not too far away. This requires however, that external social influences do not change too drastically. A considerable increase in fuel prices is an example of this kind of external impact.

Other social tendencies, such as gradual lifestyle changes in society at large, or an ageing population will influence trip behaviour in ways that cannot be reflected in the classical production and attraction models.

So-called long-term *longitudinal impacts* that are of great strategic importance, require a different type of model which are undergoing widespread development at the moment. This type of model will be discussed in another course.

5.8 Balancing production and attraction

If we take a sufficiently long period of time, the total number of departures calculated over all zones, should equal the total number of arrivals in all zones. Separate models are used, however, to calculate productions and attractions. The total number of productions and attractions calculated with these models will, in most cases, differ slightly. Matching the results, in order that the totals are equal, is called balancing productions and attractions.

It is usually assumed that the calculated productions are more reliable than the calculated attractions. This is due to the fact that housing is easier to predict than employment. The total calculated production is, therefore, accepted as the correct value for the total number of trips *T*. If D_j represents the attraction for zone *j* and J the total number of attraction zones, a proportionality factor *f* is determined:

$$f = T / \sum_{j=1}^{J} D_j$$

All calculated attractions are now multiplied by the factor *f*.

Rather than balancing the productions and attractions of the entire OD-table at once, it can be done in stages. To do this, the area is divided into regions and the productions and attractions are balanced per region.

6 Distribution

Using the production- and attraction models discussed in chapter 5 we determined the departures and arrivals for the various zones. However, we still do not know the destination of the trips that depart from a particular zone, nor do we know where the trips that are attracted by a particular zone originated.

The aim of a distribution model can now be described as follows:

Distribute:

- the trips that originate in a particular zone over all destinations
- the trips with a destination in a particular zone over all origins

The complete pattern of trips in the area of study can be represented in a so-called origin-destination table or OD-table. The aim of a distribution model, therefore, is to determine the OD-table for a particular forecast year.

We begin this chapter with a discussion of the OD-table and the notation used. Next we give a very compact description of the basic problem of the distribution calculation and the principle of the methods we can use for its solution.

We will then discuss two methods used for the distribution calculation:

- In the first method we use the trip-data from an existing OD-table for a base-year as a guideline in order to distribute the future productions and attractions (determined in the production- and attraction model) over the new OD-table for the forecast year. Because the old trip data will, in fact, be raised with a growth factor, this model is called the *growth factor model*.
- The second method does not use the existing trips between zones, but the level of travel-resistance or impedance between zones as the measure by which to distribute trips over the cells of an OD-table. Since the resulting formulas somewhat resemble Newton’s law of gravitation, the name *gravity model* is often used. Alternative names for the same model are *interaction model* and *entropy model*. The term interaction model needs no further explanation. The term entropy model came into use when the plausibility of the gravity model was demonstrated with help of the concept of entropy, known from physics and information theory. Later on in the chapter we will show that the gravity model can also be seen as a variation on the growth factor model.

Since the influence of resistance against travel is used in the gravity model, the concept of travel-impedance and the associated concept of a *deterrence function* will precede the discussion of the gravity model itself. Data from an existing OD-table (also called the *base year matrix*) is used to determine the influence of travel impedance. The objective is to determine the shape of the so-called deterrence function for the base year. It is assumed that this deterrence function retains its relevance for the forecast year. Determining the shape of the deterrence function is also called calibrating the gravity model. We raise a few points regarding this kind of calibration in this chapter. A more detailed discussion will be given in a follow-up course.

The chapter ends with a short discussion of a number of application aspects particular to distribution models.

Lastly, we will confine ourselves in this chapter to so-called unimodal (single-transport mode) distribution models. The distribution of trips over several transport modes (the modal split) will be discussed in chapter 7.

6.1 Notation

The pattern of trips in an area is usually given in a number of so-called origin-destination tables or OD-tables. Separate OD-tables tend to be used according to purpose (work, education for example), personal characteristics (whether a car is available or not, for example) and the modes of transport used (car and public transport, for example).

Table 6-1 General form of an OD-table

Departures	Arrivals				$\sum_j T_{ij}$
	1	2	j	n	
1	T_{11}	T_{12}		T_{1n}	O_1
2	T_{21}	T_{22}		T_{2n}	O_2
i			T_{ij}		O_i
m	T_{m1}	T_{m2}		T_{mn}	O_m
$\sum_i T_{ij}$	D_1	D_2	D_j	D_n	$\sum_{ij} T_{ij} = T$

An OD-table is a two-dimensional matrix consisting of m rows and n columns. The rows represent the zones of origin and the columns the destination zones. In most cases a particular zone is both a zone of origin and a destination zone. This means that we usually have a square matrix, i.e.: $m = n$. The cells of row i contain the trips that depart from zone i with the zone of the corresponding column j as destination. The cells on the diagonal from top left to bottom right show the intra-zonal trips, the trips that start and end in the same zone. The other cells represent the inter-zonal trips, i.e. origin and destination are in different zones. The number of trips from i to j (per time unit) is indicated by T_{ij}.

The sum of T_{ij} over all columns in row i represents the total number of trips leaving zone i, which is indicated by O_i. The sum of T_{ij} over all rows in column j represents the total number of trips arriving in zone j and this is indicated by D_j. The sum of all trips in the entire table is indicated by T. The total sum over the entire range of an index is described by placing the appropriate index below the summation sign. Thus we can have the following, for example:

$$\sum_j T_{ij} = O_i \qquad \text{and} \qquad \sum_i T_{ij} = D_j$$

In the distribution problem the T_{ij} are the unknowns. The aim of a distribution model is precisely to determine these T_{ij}. The departures O_i and arrivals D_j act as (boundary) constraints. We get these departures and arrivals, for example, from an application of a production- or attraction model. Depending on the situation, we speak of a doubly or singly constrained distribution model.

• *doubly constrained model*:	both departures and arrivals are known
• *singly constrained model:*	either the departures, or the arrivals are known

6.2 Basic problem of the distribution calculation

For a doubly constrained distribution problem, i.e. the determination of an OD-table where the sums of the rows (the origins) and the sums of the columns (the destinations) are known, the following applies:

$$\sum_j T_{ij} = O_i \qquad \text{voor } i = 1 \ldots m$$

$$\sum_i T_{ij} = D_j \qquad \text{voor } j = 1 \ldots n$$

This adds up to: $m + n - 1$ independent equations.

(If we have set up m equations for the rows and n-1 equations for the columns, then the condition for the last (unused) column provides no additional information. This is why there are $m + n$ -1 independent equations.)

The values to be determined are T_{ij} for $i = 1 \ldots m$ and $j = 1 \ldots n$. Thus there are $m*n$ unknowns.

In a 10x10 table, for example, we only have 19 independent equations with which to determine the 100 cells of the matrix. Thus, there are far more unknowns than equations, the system is undetermined. There is an infinite number of solutions that satisfy the given boundary constraints.

Which of these possible solutions is the correct one, or how will the trips from a particular origin distribute themselves over the destinations?

It seems obvious to assume that the smaller the distance between an origin and a destination, the greater the volume of traffic attracted by the origin-destination pair. In traffic engineering, we do not tend to speak of distances between an origin and a destination but of the more general concept of travel impedance.

We have seen above, that the boundary constraints alone are not sufficient to calculate the OD-table. We need additional conditions. These additional conditions consist of information concerning the impedance between all origins and destinations. How do we get information about the travel impedance between the origins and destinations? In principle, there are two methods:

- The first method uses the distribution of the trips in an existing matrix. This is a known OD-table for a specific base year that serves as the starting point for the calculation of future distributions. After all, the existing distribution of traffic over the OD relations is the result of the distribution of travel impedances between the OD pairs. A model based on this premise is called a growth factor model.

- In the second method we explicitly determine the travel impedance between each OD pair. Data from the base year matrix is then used to determine the influence of impedance on the distribution of trips. The information thus gathered is then used to calculate the future distribution. This type of method is called a *synthetic method*. The best-known synthetic method is the gravity model.

6.3 Growth factor models

Assume that we have access to a base year matrix, possibly from a previous study or otherwise calibrated on the basis of recent data. The aim is to determine an OD-table for a forecast year, say 10 years from now.

Suppose that we also have at our disposal a growth factor for the traffic to be expected over the coming 10 years. That growth factor can be based, for example, on expected economic growth. The expected growth could apply to the entire study area, or we can have information regarding expected growth in production and attraction for the various zones in the study area. Dependent on the available information we can distinguish a number of growth factor methods.

6.3.1 Uniform growth factor

If the information applies to a growth factor for the entire study area, we multiply each cell from the basic matrix by the growth factor. This is the most primitive situation, only suited to a planning term of at most one to two years. The method is also used when time or financial means are inadequate to carry out further research. In most cases we will have access to growth factors that have been differentiated for each zone and that come, for example, from a previous production/attraction calculation.

6.3.2 Singly constrained growth factor model

Assume that we have information about the expected increase in trips for each origin zone. In this case, it is possible to apply the zone-specific growth factor to each row in the matrix. See Table 6-2 for an example. An increase from 255 to 400 is predicted for zone 3. We now multiply all trips in row 3 by a growth factor of $g_3 = 400/255$. The same applies to the other rows.

	1	2	3	4	$\sum_j$	predicted O_i
1	5	50	100	200	355	400
2	50	5	100	300	455	460
3	50	100	5	100	255	400
4	100	200	250	20	570	702
$\sum_i$	205	355	455	620	1635	1962

	1	2	3	4	$\sum_j$	predicted O_i
1	5.6	56.3	112.7	225.4	400	400
2	50.5	5.1	101.1	303.3	460	460
3	78.4	156.9	7.8	156.9	400	400
4	123.2	246.3	307.9	24.6	702	702
$\sum$	257.7	464.6	529.5	701.2	1962	1962

Table 6-2 Example of production-constrained growth factor

6.3.3 Doubly constrained growth factor model

This is a most interesting situation. The reader is advised to take good notice of the method used to solve this problem, because the same principle will return in the discussion of the gravity model.

We now have more information both about the future number of trips that will be produced in the zones of origin, and about the number of future trips attracted by the zones of destination. This leads to a growth factor of g_i per row of the OD-table and a growth factor of *Gj* per column. Which growth factor should we take for cell *ij*? An average growth factor $(g_i + G_j)/2$ is not a good idea. If we apply this kind of average growth factor, neither the constraint on the productions (row totals), nor the constraint on the attractions (column totals) will be met.

In 1965, *Furness* suggested the following iterative method to obtain an OD-table for a forecast year. First match the matrix with the expected future productions by multiplying each row by a (row specific) growth factor *gi*. One will then find that the column totals do not agree with the expected attractions. This is why we now multiply each column in the table that we obtained in the previous step by a new (column specific) factor G_j to achieve this goal. It then becomes obvious that we need to apply new correction factors for the rows, etc. We now repeat this process of balancing until the correction factors for rows and columns converge to a value of 1.0. See Table 6-3 for an example.

	1	2	3	4	$\sum_j$	predicted O_i
1	5	50	100	200	355	400
2	50	5	100	300	455	460
3	50	100	5	100	255	400
4	100	200	250	20	570	702
$\sum_i$	205	355	455	620	1635	
predicted D_j	260	400	500	802		1962

	1	2	3	4	$\sum_j$	predicted O_i
1	5.2	43.6	97.2	254.0	400.0	400
2	44.7	3.8	83.7	327.9	460.1	460
3	76.7	128.7	7.2	187.4	400.0	400
4	133.4	223.9	311.9	32.6	701.8	702
$\sum_i$	260.0	400.0	500.0	801.9	1961.9	
predicted D_j	260	400	500	802		1962

Table 6-3 Example of doubly-constrained growth factor

The algorithm of Furness can be defined as follows:

```
Repeat
    Balance the productions;
    Balance the attractions;
Until convergence
```

It can be shown that this “Furness-process” will, in most cases, converge to a stable solution.

If we summarise all successive multiplication factors of the Furness-process in the factor $a_i = g_{i1} * g_{i2} * \ldots$for the rows and $b_j = G_{j1} * G_{j2} * \ldots$ for the columns, we can write the result of the iteration as follows:

$$T_{ij} = a_i\ b_j\ t_{ij}$$

In this formula a_i and b_j are called *balancing factors* and t_{ij} is the a-priori OD-table or the base-year matrix.

6.3.4 Disadvantages growth factor models

Growth factor models pose the following problems:

- New spatial developments for the study area cannot be accommodated. Although we can calculate the expected production and attraction, we have no suitable start values in the base year matrix.
- The method depends heavily on the reliability of the data in the individual cells of the base year matrix. Erroneous or unreliable data in one or more of the cells in the base year matrix can be reinforced through consecutive correction factors. And, when observations for specific OD-cells are not available, their future value cannot be determined.
- There are instances when problems can occur with the iteration process. We noted above that the Furness-process usually converges to a stable solution. However, when so-called zero-cells occur in the base year matrix, convergence is not always guaranteed. See the example in Table 6-4. The lower table shows the situation after 10 iterations. Convergence does not occur. The problem occurs in the second row of the matrix. Because of the desired column total, the only non-zero-cell is this row can never exceed 400. It is therefore, impossible for the second row total to achieve 460.

	1	2	3	4	$\sum_j$	voorspelde O_i
1	5	50	100	200	355	400
2	0	50	0	0	50	460
3	50	100	5	100	255	400
4	100	200	250	20	570	702
$\sum_i$	155	400	355	320	1230	
voorspelde D_j	260	400	500	802		1962

	1	2	3	4	$\sum_j$	voorspelde O_i
1	3.4	0.7	61.0	355.3	420.4	400
2	0	388.2	0	0	388.2	460
3	65.5	2.8	5.9	345.7	419.9	400
4	191.1	8.3	433.1	101.0	733.5	702
$\sum_i$	260.0	400.0	500.0	802.0	1962.0	
voorspelde D_j	260	400	500	802		1962

Table 6-4 Example of a non-converging Furness process.[2]

- A significant disadvantage of growth factor methods is their inability to incorporate changes in the transport system, in this instance the transport network. Their application possibilities are limited, therefore, in research concerning the impact of new infrastructure, such as the introduction of alternative transport modes or the introduction of tariffs (toll collection, for example).

6.4 Travel impedance and the deterrence function

We have seen that growth factor models have limited applicability. There is a need for distribution models that describe trip behaviour in a more fundamental way. The best known of these so-called synthetic models is the gravity model. Before we undertake a factual description of the gravity model, it is necessary to deal with the concepts of travel impedance and the related deterrence function.

6.4.1 Travel impedance

The effort involved, or the resistance against undertaking a trip is called travel impedance. It would seem obvious to express this impedance simply in terms of the travel time or distance involved. Things are not that simple as can be explained by an example.

> Take the trip between Tilburg in The Netherlands and Leuven in Belgium. Assume that someone wants to travel by car from Tilburg to Leuven. There are two fairly good routes:
>
> - Tilburg-Breda-Antwerp-Brussels-Leuven. This route entirely follows motorways. The length is about 135 km. If there is no serious congestion near Antwerp or Brussels, an average speed of 90 km/hour is attainable. In that case, the trip takes 1.5 hours.
> - Tilburg-Turnhout-Geel-Aarschot-Leuven. This route follows the N9. It is about 100 km long. Some parts of the roads are very good and can be travelled at 70 km/hour. Other parts of the route, however, travel through the centres of towns and villages, which causes considerable delays. The average speed achieved over the entire trajectory of this route may be about 50 km/hour. Travel time, therefore, is 2 hours.
>
> Which of the two alternatives for the car-drive Tilburg-Leuven offers the least impedance? In other words, which will be the most attractive to the driver?
>
> Let's compare travel-costs first. For the sake of simplicity, we only look at the cost of fuel. From this point of view, the journey via the N9 is preferable. The distance is shorter and, due to the lower speeds, fuel-use per kilometre is relatively low.
>
> Next, we compare travel times. In this case, the trip via Antwerp and Brussels is obviously preferable. Travel time is half an hour shorter. And the traveller also attaches a certain value to this gain in travel time.
>
> Other factors such as safety or scenery contribute to the assessment of travel impedance of a route. Assessing such factors is very subjective. When it comes to safety, the journey via Antwerp-Brussels probably wins, but many will prefer the trip through the Mid-Kempen for its natural beauty.
>
> The example shows that assigning a travel impedance to a trip is a difficult problem.

Usually the assessment of travel impedance is confined to (monetary) cost and time elements. Other factors are left out, or, if possible, reduced to cost or time elements.

Monetary costs involved in a car trip are, for example, parking, petrol and toll costs. Costs incurred in travel by public transport can include ticket prices and parking the bicycle. In addition to these so-called variable costs, fixed costs such as amortisation costs for a car or the costs for a season ticket for public transport can also be included.

Travel times for a trip per car are, naturally, the driving time, but also the time spent looking for a parking place and the times spent walking to and from the car park. Of importance in public transport are the travel times in the vehicle, the travel time getting to and from the station, waiting times and transfer times.

The total impedance of a trip from *i* to *j* via route *r* for a specific transport mode can be written as a linear combination of the experienced subjective time duration $T_{ij}{}^{r}$ and monetary costs $K_{ij}{}^{r}$. The minimum of this expression calculated over all possible routes is the travel impedance c_{ij} between *i* and *j*.

$$c_{ij} = \min_{r}(T_{ij}^{r} + K_{ij}^{r}/g)$$

Here the *value of time g* is expressed in money-units per time unit (euro/hour, for example). The value of time *g* indicates that the traveller is prepared to pay *g* money-units for a saved time-unit of travel time. In the formula, the monetary costs $K_{ij}{}^{r}$ have been converted to time units via the *y*. The customary phrase is to say that the travel impedance is expressed in *generalised time.* The unit used is, for example, minutes. The term “cost-minutes” is sometimes used to make a clear distinction between “normal” time units and generalised time units. (The travel impedance can also be expressed in *generalised (monetary) costs.* For the sake of interpretation the expression in time units is preferable; the effort in making a journey is usually associated with a time duration.)

The value of *g* depends on the person (the income of the traveller plays an important role) and on the purpose of the trip (the appreciation of the value of time in commercial traffic is noticeably higher than in other traffic). An accepted average value for *g* is about 5 euro/hour. This value can increase three- or fourfold for commercial traffic.

In the travel impedance formula above $T_{ij}{}^{r}$ and $K_{ij}{}^{r}$ are subjective travel times and monetary costs. In public transport, for example, a minute of waiting time is experienced as a greater nuisance by the traveller than a minute of effective riding time in the vehicle. Something similar happens with monetary costs. Out-of-pocket costs such as parking fees affect the traveller more than hidden costs such as petrol costs. To express this difference in perception, the duration times t_s and costs k_s of the various components *s* which together make up the journey from *i* to *j* via route *r* are multiplied by the weighting factors a_s and b_s:

$$T_{ij}^{r} = \sum_{s} a_s \cdot t_s \qquad \text{en} \qquad K_{ij}^{r} = \sum_{s} b_s \cdot k_s$$

There often is insufficient information to assess the value of the weighting factors, which is why they are often set equal to 1. In public transport, however, some research has been carried out into the weighting of travel time components. The travel time for a trip by public transport is composed of the following components: effective riding time in the vehicle, travelling time to and from the terminal, waiting time and transfer time. The following factors, for example, have been used in the traffic models for Flemish Brabant:

component s	a_s
riding time in vehicle	1.00
time to and from terminal	1.65
waiting time	1.50

Table 6-5 Weighting of travel time public transport (Fl. Brabant)

When in public transport the journey also requires transfers, then, besides the time spent waiting and walking, the traveller also experiences annoyance regarding the break in the journey itself. Extra terms in the impedance formula can also be introduced for this aspect.

From now on we will indicate the travel impedance between two places *i* and *j* by the generalised travel time c_{ij} and assume that all applicable components have been properly accounted for. If we want to distinguish between a number of transport modes we will add the index *m* (for modality). We then use the notation c_{ij}^{m}, which indicates the impedance between *i* and *j* for transport mode *m*.

6.4.2 Deterrence function

Assume that the total number of departures (the trip production) from a zone of origin is known. We want to find out how these trips will distribute themselves over the possible destinations. It is intuitively clear (and it appears from empirical research) that the number of trips to a destination decreases as the distance (or rather the travel impedance) to that destination increases. This travel impedance effect on the distribution of trips is expressed by the *deterrence function* $F(c_{ij})$. Separate deterrence functions are applied depending on the purpose of the trip, on personal characteristics and, as will become clear in chapter 7, on the mode of transport.

Over the years, many mathematical forms have been proposed for the deterrence function. Originally it was assumed that the number of trips would decrease in proportion to the square of the distance, inspired, no doubt by Newton's law of gravitation. Observations, however, did not confirm that hypothesis. Later on travel impedance was used instead of distance, while other exponents in the negative power function were tried. Further on in this chapter we will show that, based on theoretical considerations, the shape of the deterrence function can be approximately described by a negative exponential function, at least over a limited range of the travel impedance. Some functions that have been used are:

$F(c_{ij}) = c_{ij}^{-a}$ negative power function

$F(c_{ij}) = e^{-b\,c_{ij}}$ negative exponential function

$F(c_{ij}) = c_{ij}^{-a} \cdot e^{-b\,c_{ij}}$ combined power- and exponential function

Parameters a and b in the functions above are determined through calibration using observations from the study area. The general shape of the functions for some values of the parameters is given in Figure 6-1.

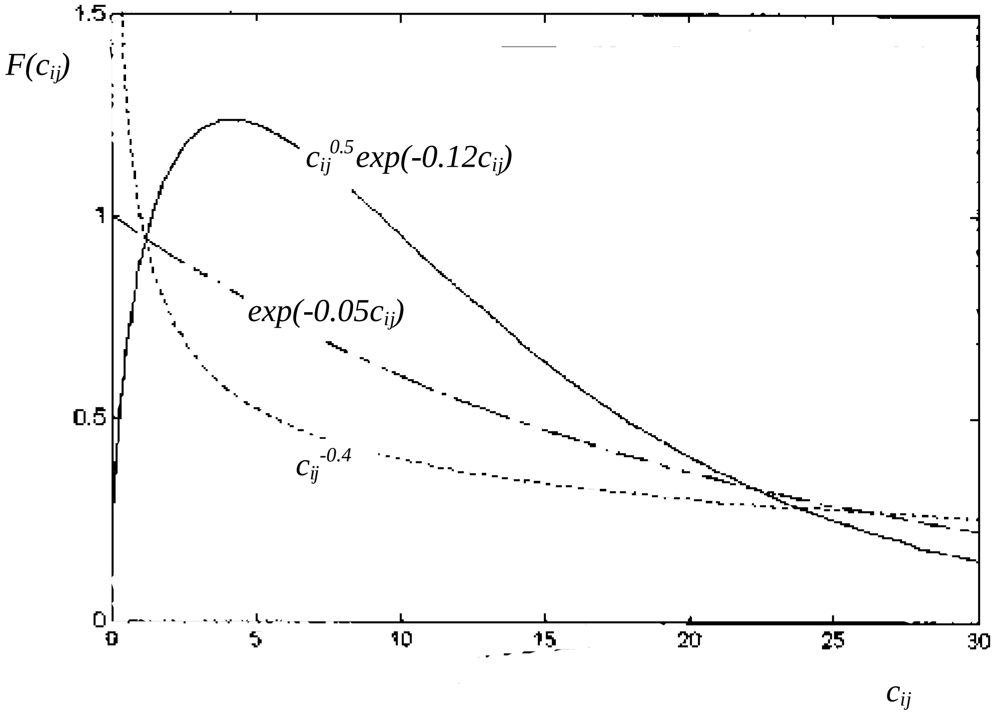


Figure 6-1 Some analytical deterrence functions.

Although often used, the negative exponential function has a disadvantage. The same absolute increase in travel impedance at low values of the travel impedance has the same relative effect on the trips as at a high travel impedance. This does not confirm to our intuition. One would expect, for example, that an increase in travel impedance from 5 to 10 minutes would have a relatively larger effect than an increase from 120 to 125 minutes. It will, therefore, usually only be possible to fit a negative exponential function to empirical data over a limited range of the travel impedance. (See also example 1 in chapter 3.5.)

Sometimes a deterrence function does not decrease monotonously across the entire range of the travel impedance. In the case of car trips, for example, the value of the deterrence function starts at a low level, reaches a maximum, and only then decreases with increasing travel impedance. The reason is that the tendancy to use the car for very short distances will generally be low. Walking or cycling is then a preferable choice. In such cases it is sometimes possible to use a combination of power function and exponential function to describe the data.

It is of course not at all necessary to write the deterrence function in a closed mathematical form. By enumeration (or table lookup) we can also define a function. In that case the function value associated with each argument of the function is specified. In practice this can be done, for example, by storing the deterrence function value (sometimes called a *friction factor*) for a number of discrete values of the argument in a table (a friction factor table) and to assume that the function value in an interval around a discrete value remains constant. One can also apply an interpolation-technique in the interval around the discrete value.

6.5 Gravity model

Having introduced the concepts of travel impedance and deterrence function, we will now proceed with our examination of distribution models. The gravity model to be discussed in this section is strongly related to the growth factor models we studied in a previous section. In the growth factor models we used an existing a-priori matrix to measure the influence of the travel impedance on the number of trips to the various destinations. In the gravity model we use the value of the deterrence function as a measure for the influence of impedance on the expected number of trips between origins and destinations.

6.5.1 Principle of the gravity model

The best way to explain the gravity model is by way of an example. Table 6-6 is an OD-table where the margins show the productions and attractions that have been predicted using a production and attraction model. The problem is to enter trips in the OD-table in order to meet the constraints.

Boundary constraints					
	1	2	3	4	predicted O_i
1					400
2					460
3					400
4					702
predicted D_j	260	400	500	802	1962

Table 6-6 Boundary constraints gravity model example.[2]

The following table contains the travel impedance c_{ij} between all origins and destinations, expressed for example in minutes of generalized time.

Impedance c_{ij} (minutes)				
	1	2	3	4
1	3	11	18	22
2	12	3	13	19
3	15.5	13	5	7
4	24	18	8	5

Table 6-7 Impedance table gravity model example.[2]

Assume that, based on data from a base-year matrix, we calibrated the following deterrence function.

$F(c_{ij}) = e^{-0.1 c_{ij}}$

For every cell in the matrix we calculate the value of the deterrence function, obtaining a table of so-called friction factors. We now start the balancing process as shown in Table 6-8. This table gives the relative proportions between the number of trips in each cell of the OD-table that is to be estimated

Starting matrix = Friction factor table *F(c_{ij})* = exp (- 0.1 c_{ij})						
	1	2	3	4	$\sum_j$	predicted O_i
1	0.74	0.33	0.17	0.11	1.35	400
2	0.30	0.74	0.27	0.15	1.49	460
3	0.21	0.27	0.61	0.50	1.59	400
4	0.09	0.17	0.45	0.61	1.32	702
$\sum_i$	1.34	1.51	1.53	1.37	5.75	
predicted D_j	260	400	500	802		1962

Table 6-8 Friction factor table gravity model example.[2]

In order to achieve an OD-table that complies with the constraints we apply the Furness iteration process to the starting matrix in Table 6-8, in the same way as described in the growth factor method. The end-result is given in Table 6-9.

Verplaatsingen T_{ij} berekend met het zwaartekrachtmodel						
	1	2	3	4	$\sum_j$	a_i
1	157	98	69	76	400	410.0
2	59	204	101	96	460	379.5
3	25	45	138	192	400	229.0
4	19	53	192	438	702	428.7
$\sum_i$	260	400	500	802	1962	
b_j	0.52	0.73	0.99	1.68		

Table 6-9 Results gravity model example.[2]

If we summarise all successive multiplication factors of the Furness process in the factor a_i for the rows and b_j for the columns, we can write the results of the iteration as follows:

$$T_{ij} = a_i b_j F(c_{ij})$$

This is the usual formulation of the doubly constrained gravity model. Note that it resembles the growth factor model. Again, we call a_i and b_j balancing factors and *F(c_{ij})* is the deterrence function, obtained by calibration from the data in the base-year matrix. Completely analogous to the singly constrained growth factor model, we can also deduce the singly constrained gravity model. In that case a_{ij} = 1 ór b_{ij} = 1, and a Furness iteration process is not necessary.

6.5.2 Observations regarding the gravity model

6.5.2.1 *The values of the deterrence function and the balancing factors*

Only the relative magnitudes of the values in the deterrence function matter, not the absolute values. This also applies to the balancing factors a_i and b_j. If we start the Furness iteration process by balancing the columns instead of the rows, we will get different balancing factors. The ratios between the a_i as well as between the b_j, however, remain the same.

Another way to express this is as follows: the balancing factors and the value of the deterrence function are determined up to a constant factor; multiplying, for example, the a_i with a constant factor changes nothing in the result provided that we divide the b_j, the $F(c_{ij})$ or the product of both, by the same constant factor.

6.5.2.2 *Alternative formulations for the gravity model*

In the literature we might find different mathematical expressions for the gravity model. Examples are:

$$T_{ij} = A_i O_i B_j D_j F(c_{ij}) \qquad \text{and} \qquad T_{ij} = l_i Q_i m_j X_j F(c_{ij})$$

In the first expression, O_i and D_j represent departures from zone *i*, respectively arrivals in zone *j*. The formula arises by writing a_i as A_i O_i and b_j as B_j D_j. The original balancing factors a_i and b_j have been replaced by the new balancing factors A_i and B_j. This is not particularly useful. But it is understandable. One would like to interpret the gravity model as a model that describes the trips between an origin and a destination as a function of the characteristics of that origin and destination and the travel impedance between them. Because the balancing factors a_i and b_j are difficult to interpret as characteristics of origins and destinations, one resorts to departures and arrivals. This does not help much because new balancing factors A_i and B_j are now required, and their interpretation is as difficult as the original a_i and b_j

In the second expression one does not use departures or arrivals but other variables Q_i and X_j that are designated as the "polarities" of the zones of origin and destination. The idea is that the polarity is a characteristic for a zone and that it describes a zone's capacity to generate trips, respectively to attract them. We still need the balancing factors l_i and m_j .

6.5.3 Theoretical derivations of the gravity model

We have seen that the gravity model can be considered as an extension of the growth factor model with the associated Furness procedure. Although at first nothing seems to be wrong with the Furness procedure, there are come conceptual problems. This can be understood as follows.

The starting point of the entire procedure was that trip behaviour could be described by a deterrence function that gives the influence of travel impedance on the number of trips. In the first iteration we distributed the total number of departures from a zone over the destination zones proportional to the values of the deterrence function. This proportionality is disturbed in the next iteration, when the column totals must be matched to the calculated attractions. After convergence of the Furness procedure proportionality of the trips with the values of the deterrence function over the entire OD-table is gone. This can be checked in the examples in Table 6-8 and Table 6-9. We cannot harmonise the trips in the entire OD-table with the deterrence function value for each cell and comply with the constraints at the same time.

This shows that the Furness procedure can, at best, be seen as a compromise between, on the one side, optimum proportionality to a deterrence function value and, on the other side, complying with the constraints posed by the calculated productions and attractions.

Due to the unstable theoretical basis of the gravity model as based on the Furness procedure, one has sought other arguments to prove the validity of the gravity model.

We will very briefly discuss two alternative arguments for the validity of the gravity model, namely a derivation based on the theory of entropy, known from physics and information theory, and a derivation based on discrete choice theory.

6.5.3.1 Derivation gravity model from the principle of maximum entropy

Assume that we have to estimate an OD-table based on given marginal sums of the table (sums of origins and destinations). The problem is that very many (in fact an infinite number of) OD-tables will match the given marginal sums. Figure 6-2 shows an example.

5	0	5
1	2	3
6	2	8

4	1	5
2	1	3
6	2	8

3	2	5
3	0	3
6	2	8

Figure 6-2 Different OD tables with the same marginal sums.

All these possible OD-tables, however, do not have the same probability of occurrence. We now assume that the particular OD-table will arise that has the greatest probability of being realised. Using terminology from physics and information theory we can say that we are looking for the OD-table with maximum entropy.

The number of ways in which we can divide T trips over an OD-table in such a way that the distribution over the cells is T_{ij}, is:

$$W(T_{ij}) = \frac{T!}{\prod_{i,j} T_{ij}!}$$

(This is the multi-nominal formula known from probability theory)

Now our task is to find the OD-table with T_{ij} such that:

> $W(T_{ij})$ is maximal (the OD-table can arise in the maximal number of ways; i.e. has a maximal probability to arise) under the constraints:
>
> $$\sum_j T_{ij} = O_i$$
> $$\sum_i T_{ij} = D_j$$

If we solve this maximisation problem we do not distinguish between tables T_{ij} with many trips on a relation with great travel impedance or many trips on a relation with a small travel impedance. To make all OD-tables comparable we add a condition concerning total trip costs across the entire matrix:

> $\sum_{ij} T_{ij} c_{ij} = C$ where C is a constant

Solving this maximisation problem (using Lagrange multipliers) leads to:

$T_{ij} = A_i O_i B_j D_j \ e^{-b\,c_{ij}}$ (where b is related to C)

This is the gravity model with a negative-exponential deterrence function. When calculating the total trip costs, the long-distance trips in the condition $\sum_{ij} T_{ij} c_{ij} = C$ are as important as short-distance trips; if we write the condition in the more general form $\sum_{ij} T_{ij} g(c_{ij}) = C$, where g is a monotonously increasing function of c_{ij}, we obtain one of the common formulations of the gravity model:

$T_{ij} = A_i \ O_i \ B_j \ D_j \ f(c_{ij})$

6.5.3.2 Derivation gravity model from discrete choice theory

Because the distribution problem is essentially a problem of choice of origin and destination, we can apply the discrete choice theory that was dealt with in chapter 3.

The individual utility $U_{i,j}{}^p$ for person p of the choice of i as origin, j as destination and making a trip from i to j consists of a number of components:

$U_{i,j}{}^p = V_i^p + V_j^p + V_{ij}{}^p + e_{ij}{}^p$

where:

V_i^p	the observable utility for person p of an activity in i
V_j^p	the observable utility for person p of an activity in j
$V_{ij}{}^p$	the observable utility for person p of making the journey from i to j
$e_{ij}{}^p$	an error term; accounts for the effect of non-observed attributes

The utilities V_i^p and V_j^p are the results or positive effects that arise out of the choice of places i and j as origin and destination. The utilities are related to the activities that are carried out in i and j (for example, living in i and working in j). The term V_{ij}^p on the other hand represents the effort or the negative effect of the journey from i to j. The term V_{ij}^p, therefore, will have a negative value and this value will decrease further as the effort to be overcome in the journey increases.

We now apply an aggregation step, and assume that the individual utilities may be replaced by an average utility per person across the entire zone:

$$U_{i,j} = V_i + V_j + V_{ij} + e_{ij}$$

If we assume that the e_{ij} are identical and independently Gumbel-distributed, a logit model applies for the probability that a trip from i to j will be made:

$$\Pr(ij) = e^{V_i+V_j+V_{ij}} \ / \sum_{alle\, i'j'} e^{V_{i'}+V_{j'}+V_{i'j'}}$$

If we assume that T represents the total number of trips in the entire OD-table, then the expected number of trips from i to j is:

$$T_{ij} = \Pr(ij) \cdot T = e^{V_i} \cdot e^{V_j} \cdot e^{V_{ij}} \cdot \frac{T}{\sum_{i'j'} e^{V_{i'}+V_{j'}+V_{i'j'}}}$$

The last factor in this expression equals a constant, say K. The constant factor has the effect of a scaling factor that ensures that the calculated trips agree with the total number of trips in the OD-table. After the substitution of variables:

$$exp\ (V_i) = a¢_i \qquad\qquad exp\ (V_j) = b¢_j$$

follows:

$$T_{ij} = K\ a'_i b'_j\ e^{V_{ij}}$$

The scaling factor K can be included in the factors $a¢_i$ en $b_i¢$ without loss of generality. This leads to:

$$T_{ij} = a_i b_j\ e^{V_{ij}}$$

This, in essence, is the gravity model in its most general formulation. We can further elaborate the formula above, by taking a closer look at the utility of a trip V_{ij}.

The utility V_{ij} will be a (generally monotonously decreasing) function of the travel impedance between i and j.

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

After substitution of exp($f\ (c_{ij})$) by $F(c_{ij})$, we obtain the well-known form of the gravity model:

$$T_{ij} = a_i b_j F(c_{ij})$$

If the negative valuation of the trip is one of “costs” and if the travel impedance c_{ij} is more or less proportional to the distance or the travel time between i and j, the function for the disutility will approximate a linearly decreasing function:

$$V_{ij} \approx -b\ c_{ij} \qquad (b > 0)$$

In that case the gravity model with an exponential deterrence function arises:

$$T_{ij} \approx a_i b_j\ e^{-b\ c_{ij}}$$

Given the substantial amount of simplifying assumptions made (namely the assumptions inherent in the logit model and the aggregation step) it is not surprising that the gravity model merely gives an approximate description of the distribution. Substantial differences can, indeed, arise between observations and values calculated by the gravity model, particularly when detailed comparisons are made at the level of particular origins and destinations. In general, however, and taken across the entire OD-table, the observations and model values match reasonably well in practice.

6.5.3.3 Conclusions theoretical derivations

We have seen that the gravity model is essentially based on a Furness iteration procedure where the deterrence function values are used as start values. In addition, the gravity model can be derived using the principle of entropy or from discrete choice theory. These derivations do not prove the “correctness” of the gravity model as such, but they increase our confidence in its correctness. The predictive value of the gravity model should really be established by strictly controlled experiments, as is done in the natural sciences. Unfortunately, such a procedure is hardly ever possible in the socio-economic sciences. (Traffic science, traditionally seen as an engineering discipline, has much common ground with the socio-economic sciences.)

6.5.4 Calibration of the deterrence function

When we discussed the deterrence function, we noted that determining the parameters in that function must be done through calibration with available observations for the study area. In this course, we will confine ourselves to a few introductory remarks about the calibration process. For those interested in a more in-depth treatment of this subject, we refer to a future follow-up course.

Assume that an OD-table with observations is available for a base-year. Assume that a distribution model is applicable to this OD-table, for example the gravity model discussed in this chapter: $T_{ij} = a_i\, b_j\, f\,(c_{ij})$

Parameters in this distribution model are a_i, b_j and the parameters in the deterrence function f. To calibrate the deterrence function now means: determine the parameters in the distribution model in such a way that maximum agreement is achieved between the observed OD-table and the OD-table that has been calculated by the distribution model. A much-used term in this context is "best fit" of distribution model and observations.

How can the parameters be determined in order to achieve a best fit? An obvious method is "trial and error". We make an initial estimation of the parameters based on insight and experience. We then carefully try to adjust the parameters to improve the match. In practice, however, this is a very time-consuming method. We need a more systematic way to carry out the calibration.

Many calibration methods have been proposed over the years. One of the most efficient methods is based on the maximum likelihood principle, a well-known statistical estimation method. An example of this method is the so-called Poisson-estimator that will be dealt with in a future course.

In the previous section we assumed that we had access to an observed OD-table for the calibration. In many cases, however, we have no observations for the cells in the OD-table, although we do have counts on the links of the network; vehicle counts, for example, on several road sections. Depending on the route choice of the travellers a number of OD-relations can use the same link in the network. We must then reconstruct the OD-table as it were, using a route choice model. Route choice models are dealt with in the chapter about assignment. Calibration using counts on network links is dealt with in a follow-up course.

6.6 Application aspects of distribution models

We conclude this chapter about distribution models with a short discussion of a number of practical application aspects.

6.6.1 Intra-zonal traffic

Intra-zonal traffic consists of trips that have their origin and destination inside the same zone. This contrasts with inter-zonal traffic with origin and destination in different zones. We use the travel impedance between the centroids of the zones involved for the inter-zonal trips, whereby traffic is assumed to travel via the model network. The traffic enters and leaves the model network via the connectors. Intra-zone trips, on the other hand, do not use the network. Thus travel impedance between origins and destinations are not known. *This can lead to problems because, due to their short distance, the volume of intra-zonal trips tends to be large.* There are two approaches to address these problems:

- Use small zones and possibly do not include the intra-zonal traffic in the distribution model.
- Approximate the intra-zonal impedance levels, for example as a function of the zone area or as a function of a characteristic diameter of the zone.

6.6.2 External zones

When applying a distribution model, the levels of impedance between the zones must be known. This generally applies to the zones in the study area, the internal trips. A substantial part of the traffic will, however, have its origin or destination outside the study area. It is also possible that both origin and destination lie outside the study area, though these trips traverse the study area. The travel impedance to and from these external zones (also called the area of influence or external area) is harder to define.

A possible approach to this problem is to calculate the internal trips by a synthetic distribution model (the gravity model, for example) and to use a growth factor model for the remaining trips. The data necessary for this purpose can be obtained from counts on a cordon (a closed circle) situated around the study area.

6.6.3 Conversion of the OD-table to suitable units

The results of the distribution calculation serve as an input into the next phase of the traffic demand model, namely assignment. Distribution models usually deal with personal trips. These can be trips per day, peak-period trips or trips aggregated over some other time unit.

In an assignment model for car traffic one will need, for example, data on vehicle trips/hour during the peak period. Factors such as degrees of vehicle-occupation and peak-hour coefficients are used to carry out a simple conversion to suitable units.

7 Transport mode choice

The different phases in the traditional traffic demand model show a logical sequence. First the production- and attraction model is used to determine the number of trips that can be expected. Then the distribution model is used to establish where these trips go, in other words origins and destinations are determined. The assignment phase, which will be dealt with in chapter 8, calculates how the trips take place in practice, with the emphasis on the route choice process.

Going somewhere not only involves a choice of route but also a choice of transport mode. The distribution of trips over the various transport modes is called the *modal split*. Modelling transport mode choice is one of the classical problems in traffic engineering.

The various modalities include walking, cycling, driving, public transport and perhaps some other possibilities depending on local customs. Due to the high social relevance of this theme, the emphasis usually is on the choice problem between car and public transport. Because of road network congestion and environmental degradation caused by road traffic, the relevant policy-making agencies usually aim to encourage use of public transport. It is very important, therefore, to have models that are sensitive to the attributes that influence the choice of transport mode.

We begin this chapter with a short overview of those factors that have a determining effect on transport mode choice.

Over the years, the place of the transport mode choice model within the traditional traffic demand model has changed quite a bit. It was originally thought that transport mode choice had to be modelled as part of production and attraction. In doing so, however, the network characteristics are unable to influence the transport mode choice. Thus, it became preferable to incorporate the transport mode choice in the distribution phase. We will discuss these so-called simultaneous and sequential choice models for distribution and transport mode choice at length.

7.1 Factors that influence transport mode choice

Many factors influence transport mode choice. First and foremost there is the availability of the various means of transportation. People who have no choice but to use one or other transport mode are called *captives* of that transport mode.

The word captives is most often used in connection to public transport. When a household has no access to a car while the destination is too far away to cycle or walk, and when family income does not stretch to car hire or taxi, the family member is said to be a public transport captive.

However, it is also possible that there is no provision of public transport to a destination, or the nature of a job is such that public transport does not apply. In such cases, people necessarily depend on cars and become car-captives.

Those people who are not captive to one or other form of transport are called *choice-travellers*. It is assumed that these travellers base their choice of transport mode on rational considerations. The factors that play a role in this process can be divided into three groups:

- *Traveller characteristics* It appears that there is a connection between transport mode choice and socio-economic characteristics such as profession, income, age, etc. The most significant characteristic is car availability. This characteristic is closely connected to the above mentioned socio-economic characteristics.
- *Transport mode characteristics* In this group of characteristics the differences in travel time and costs between the transport modes are particularly important. However, factors such as parking opportunities and comfort, safety, and reliability also feature.
- *Trip characteristics* The purpose of the trip plays a role here. People might use public transport for the recurrent home-based work trip, and use the car to go shopping. The point in time at which the trip is undertaken is also of significance.

Any traffic model that is intended to represent the transport mode choice accurately should really be sensitive to most of the factors above. In practice however, most applications confine themselves to travel impedance in the mode choice model. We refer to the chapter on distribution for a discussion of the concept of travel impedance.

7.2 Transport mode choice as part of production/attraction calculations

In the past, especially in the US where the first developments of the traditional traffic demand model happened, the calculation of the transport mode choice was a part of the production- and attraction model. Productions and attractions are also known as *trip-ends*. This is why these modal split models are also called trip-end modal split models.

The reason that the calculations were carried out in the production-attraction phase was because the transport mode choice, which was mainly considered to be a choice between car and public transport, was primarily determined by personal characteristics such as income. In some cases, the extent of the availability of public transport services was added in the form of a kind of accessibility index.

The disadvantage of the calculation of the modal split during or directly after the production- and attraction phase is that the destinations of the trips are not yet known at this stage. This means that the network characteristics cannot yet be included in the model. As a consequence, these models do not respond to policy decisions, such as, for example, improvements to the public transport network. This is the main reason why trip-end modal split models are no longer used.

7.3 Transport mode choice as part of the distribution calculation

When the calculation of the transport mode choice is done as part of the distribution calculation, one speaks of trip-interchange models. Carrying out the modal split calculation in combination with the distribution is a logical option. The disadvantage of trip-end models is absent, since trip characteristics such as travel time can now be included in the calculation of the transport mode choice. Destination choice and transport mode choice are, moreover, closely connected. These kinds of models have, therefore, become widely used.

7.3.1 Simultaneous model distribution/transport mode using multi-modal gravity model

The gravity model that we used to calculate the distribution can be adapted so that assignment over the transport modes is simultaneously calculated.

In chapter 6 about distribution we learned that the gravity model for one mode of transport (a *uni-modal gravity model*) can be derived using the concept of entropy or using the logit model that originates from discrete choice theory. When there is a choice of transport modes between the same origin and destination it is possible to derive a *multi-modal* version of the *gravity model*. We will illustrate this by a derivation using the logit model.

The choice set for the logit model in the uni-modal gravity model consisted of all pairs of origins and destinations. When there is a choice between several transport modes we get an enlarged choice set because one or more transport modes must now be included per origin-destination pair.

If we assume, as we did in the derivation of the uni-modal gravity model, that the individual utilities can be replaced by a mean utility per person across the entire zone, the following expression arises for the utility of the choice of *i* as origin, *j* as destination and the undertaking of a trip from *i* to *j* using transport mode *m*:

$$U_{i,j}^{m} = V_i + V_j + V_{ij}^{m} + e_{ij}^{m}$$

With

V_i the observable utility for an activity in *i* (living, for example)

V_j the observable utility for an activity in *j* (working, for example)

V_{ij}^{m} the observable utility of travel from *i* to *j* by modality *m*

e_{ij}^{m} a stochastic error term (effect of unobserved attributes)

The difference with the derivation for a single transport mode is the addition of the index *m* for the various transport modes between *i* and *j*. If we accept that the error terms e_{ij}^{m} are identical and independently Gumbel-distributed (unavoidable assumptions for the logit model), we can derive as we did in the uni-modal case:

$$T_{ij}^{m} = a_i b_j \; e^{V_{ij}^{m}}$$

Similar to the uni-modal case, the utility V_{ij}^{m} represents the “effort” involved in making a trip from *i* to *j* using transport mode *m*. The greater the effort, the smaller the value of V_{ij}^{m}. The effort will be a function of the travel impedance experienced by using transport mode *m* between *i* and *j*. It is very likely, moreover, that the functional form will depend on the transport mode, hence the addition of index *m* to the function *f*:

$$V_{ij}^{m} = f^{\,m}(c_{ij}^{m})$$

After substitution of $\exp(f^{m}(c_{ij}^{m})$ by $F^{m}(c_{ij}^{m})$ we eventually get:

$$T_{ij}^{m} = a_i b_j F^{m}(c_{ij}^{m})$$

thus:

$$T_{ij} = \sum_{m' \in ij} T_{ij}^{m'} = a_i b_j \sum_{m' \in ij} F^{m'}(c_{ij}^{m'})$$

In words: the distribution of the transport flows across the various transport modes m between an origin i and a destination j is proportional to the values of the deterrence function $F^{m}(c_{ij}^{m})$ for those transport modes.

A critical point in this derivation is the assumption regarding the independence of the error terms e_{ij}^{m}. Since a number of alternatives in the choice set apply to trips between identical origins and destinations, this independence may be doubtful. On the other hand, if one accepts that the variance in the total utility largely arises from the differences in perception between the trip impedances, the assumption of independence in the error terms may be justified after all.

Thus far, we assumed that all the alternatives in the choice set are simultaneously available. There is a simultaneous choice for destination and transport mode. This is why this model is called a *simultaneous choice model for distribution and transport mode.*

Another option is to assume a choice hierarchy. This could be a choice of destination first and only when this choice has been made, a choice for transport mode. This leads to an hierarchical or *sequential choice model for distribution and transport mode*, which is the subject of the next paragraph 7.3.2.

The simultaneous multi-modal gravity model assumes the availability of separate deterrence functions per transport mode. Since deterrence functions usually need to be further differentiated according to trip purpose and personal characteristics, a set of deterrence functions is required before a calculation with the gravity model can be carried out. If one distinguishes, for example, between the personal characteristic “car-available” and “car-not-availabe”, the trip purposes “work” and “other" and the modalities “car”, “public transport” and “bicycle”, one needs a set of 12 deterrence functions.

Example of a calculation using the multi-modal gravity model:

Assume that there are three zones A, B, and C. The available transport modes for all relations are car, bicycle and public transport. We do not differentiate as to purpose or personal characteristics.

The marginal constraints, calculated by a production- and attraction model, are given in Table 7-1. The task is to insert trips per transport mode in the Table.

Marginal constraints (car, bicycle and public transport combined!)				
	A	B	C	predicted O_i
A				100
B				100
C				200
predicted D_j	200	150	50	400

Table 7-1 Marginal constraints multi-modal gravity model example

We need the deterrence functions per transport mode. The values of the deterrence function per transport mode (friction factors) are given in the OD-Table 7-2.

Friction factors $F_{ij}^{m}(c_{ij}^{m})$				
		A	B	C
A	car	20	10	2
	bicycle	10	5	1
	public transport	4	3	1
B	car	10	20	5
	bicycle	5	10	2
	public transport	3	4	2
C	car	2	5	20
	bicycle	1	2	10
	public transport	1	2	4

Table 7-2 OD-table with friction factors per transport mode

We now have all data needed to calculate the OD-table per transport mode. It is more correct to say that we will determine three OD-tables: one per transport mode. We must first aggregate the values of the deterrence function per OD-pair This gives the results in Table 7-3.

Now, as was done in chapter 6.5.1, the table is used as a starting matrix for the Furness-process.

Aggregated friction factors $\sum_m F_{ij}^m(c_{ij}^m)$					
	A	B	C	$\sum_j$	predicted O_i
A	34	18	4	56	100
B	18	34	9	61	100
C	4	9	34	47	200
$\sum_i$	56	61	47	164	
predicted D_j	200	150	50		400

Table 7-3 Friction factors aggregated by transport mode

After a number of iterations of the Furness-process we then find the total number of trips calculated by the gravity model. The trips in Table 7-4 are the sums of the trips across all transport modes.

Trips (all modes) by gravity model					
	A	B	C	$\sum_j$	a_i
A	78	22	0	100	1.01
B	50	48	2	100	1.23
C	72	80	48	200	7.85
$\sum_i$	200	150	50	400	
b_j	2.27	1.14	0.18		

Table 7-4 Total number of trips multi-modal gravity model example

Lastly we distribut the total number of trips across all transport modes, in proportion to the value of the friction factors The result is shown in Table 7-5.

Trips by transport mode						
		A	B	C	Total O_i^m	Total O_i
A	car	46	12	0	58	100
	bike	23	6	0	29	
	publ tr	9	4	0	13	
B	car	28	28	2	58	100
	bike	14	14	0	28	
	publ tr	8	6	0	14	
C	car	36	44	28	108	200
	bike	18	18	14	50	
	public tr	18	18	6	42	
Total D_j^m	car	110	84	30	224	
	bike	55	38	14	107	
	publ tr	35	28	6	69	
Totaal D_j		200	150	50		400

Table 7-5 Final results multi-modal gravity model example

7.3.2 Sequential model distribution/transport mode via the logsum

Assume that we want to do the modal split calculation *after* the distribution calculation. We first determine the *total* of all transport flows (without regard to transport modes) between each origin and destination using a distribution model. If there are several transport modes between an origin and a destination, car and train for example, we distribute the complete transport flow across the transport modes concerned *after* the distribution calculation, using a logit model.

This, however, presents a methodological problem. We need the travel impedances between all origins and destinations and a deterrence function to carry out the distribution calculation, using a gravity model for example. But when we start the distribution calculation we have no information yet as to how the flows are distributed across the transport modes. It is unclear, therefore, what travel impedance and deterrence function we should use for a specific origin-destination pair. Should we use the car-specific values, the train-specific values or a kind of average of the two?

Let's look at travel impedance. Assume that there are two possible transport modes between a particular origin and destination. It would seem acceptable to take the average travel impedance of these two transport modes as representing the travel impedance for the relation. But this is incorrect. Assume that the traveller estimates the impedance by car between A and B to be 30 minutes for example (generalised travel time). In that case the introduction of a train connection between A and B with an estimated impedance of 40 minutes of generalised travel time will not make the overall impedance on the relation (30 + 40)/2 = 35 minutes. In general one could say that the overall experienced travel impedance on a relation decreases when an additional travel option is introduced. At most, it will remain equal to the impedance that prevailed befóre the introduction of the alternative transport mode. Nor is it entirely satisfactory to equalise the overall impedance to the minimal impedance of all transport modes on a relation. This does not do justice to the fact that the objectively worst connection, for example in terms of travel time, will still be preferable in the subjective perception of some travellers. As we have shown when discussing discrete choice theory, we simply do not know all the considerations that help travellers decide on a particular transport mode.

The problem above can be solved by using the train of thought expounded in the calculation method of the hierarchical logit model in chapter 3.6.2.

We distribute the total transport flow (aggregated over all transport modes) across all origins and destinations using the following gravity model (see chapter 6.5.3.2):

$$T_{ij} = a_i b_j \ e^{V_{ij}}$$

Here, V_{ij} is the utility related to the trip between i and j taken over all transport modes serving the connection between i and j. Remember that V_{ij} has a negative value and that the value decreases as the effort involved in making the trip increases. If there are several transport modes between i and j, we calculate the *combined* (or *replacing*) *utility* by the following formula:

$$V_{ij} = q \ \ LS_{ij}$$

in which LS_{ij} is the *logsum* (over all transport modes between i and j) given by:

$$LS_{ij} = \ln \sum_{m' \in ij} e^{V_{ij}^{m'}}$$

and where, moreover: $0 < q \ \ \leq 1$

We next find the transport flows per modality by applying the logit model, where the various transport modes represent the choice alternatives between the same origin and destination:

$$T_{ij}^{m} = T_{ij} * \frac{e^{V_{ij}^{m}}}{\sum_{m' \in ij} e^{V_{ij}^{m'}}}$$

If q = 1, then the model discussed here is algebraically equivalent to the simultaneous model we discussed in the previous paragraph 7.3.1. (See also the discussion of the hierarchic logit model in chapter 3.6.2.)

It is sometimes argued that the choices of destination and transport mode are made simultaneously. When choosing a destination one also reflects on the available transport options to that destination. Others argue that the choice of destination comes before the choice as to the kind of transport mode (i.e. a sequential choice process). Both observations bear some truth. Things also depend on the purpose of the trip and other factors.

Because it appears that distribution and transport mode are simultaneously calculated in the multi-modal gravity model, while when using the logsum method the transport mode is calculated subsequent to distribution, the impression is created that both calculation methods reflect different points of view. As we noted above, this only applies when $q < 1$ is in the logsum method. When $q = 1$ both methods are equivalent and can be derived from one another.

To conclude, we point out that the application of sequential choice models using hierarchical logit models has been put forward in the literature (see for example Ben Akiva and Lermand (1985)[3]), but that the sequential choice model is rarely used. This is probably because, in practice, the determination of the correct choice structure and the estimation of the parameters may lead to problems.

8 Traffic assignment

The traditional traffic demand model consists of the following sub-models:

- A production model and an attraction model with which the number of departures and arrivals is assessed.
- A model for distribution- and transport mode choice to determine OD-tables with trips per transport mode.
- A traffic assignment model that is used to convert the data from the OD-tables to flows on the links of the network for the various transport modes.

This chapter deals with traffic assignment models, the third phase of the traditional traffic demand model. We confine ourselves to the basic principles of traffic assignment models. Those wishing more information are referred to the standard text on this topic by Sheffi (1985)[5]**.**

The primary concern in traffic assignment models is *route choice*. It would appear self-evident that a traveller would, in principle, choose the shortest route to his point of destination. This is why shortest route algorithms play an important role in traffic assignment models.

Since there are such large differences between networks for private transport (car, bicycle, etc.) and public transport, they will be dealt with separately.

8.1 Traffic assignment models for road networks

8.1.1 Shortest route in a road network

In the traffic assignment model discussed in this chapter, the shortest route (also called the shortest path) in the network must be repeatedly determined between an origin and a destination. There are many shortest path algorithms. We will discuss one on the best known and most used, namely *Dijkstra's algorithm*.

Assume we are looking for the shortest path between nodes s and t in a network. To that end, Dijkstra's algorithm builds a tree (a shortest route tree) starting from s. This is why this algorithm is also called a "tree builder algorithm".

Starting out from s we travel through the network and label every node u on the way with $L(u)$. This label $L(u)$ indicates the length of the provisionally discovered shortest path from s to u. The labels are initially provisional and can be changed in the course of the algorithm as we find a path that is shorter than the current value of $L(u)$. When a label can no longer be changed (node no longer present in set T in the algorithm below) it becomes a definitive label.

In the algorithm below the following notation is used:

V	the set of all nodes
T	the set of all nodes with a provisional label
s	the node of origin
t	the node of destination

u, v	general indication for a node
e	general indication of a link

The algorithm reads:

```
step 1:    $L(s) := 0$ and for all $v \in V,\ v \neq s : L(v) = \infty$
step 2:    $T := V$
step 3:    Let $u \in T$ for which $L(u)$ is minimal:
               If $L(u) = \infty$ then stop; there is no solution.
               If $u = t$ then $T := T - \{u\}$ and stop;
                   ($L(t)$ is the shortest route from $s$ to $t$)
step 4:    For each link $e$ from $u$ to $v \in T$ :
               If $L(v) > L(u) + length(e)$ then $L(v) := L(u) + length(e)$
Step 5:    $T := T - \{u\}$ and go to step 3.
```

Example:

We illustrate the algorithm by determining the shortest path between *A* and *H* in the network shown in Figure 8-1.

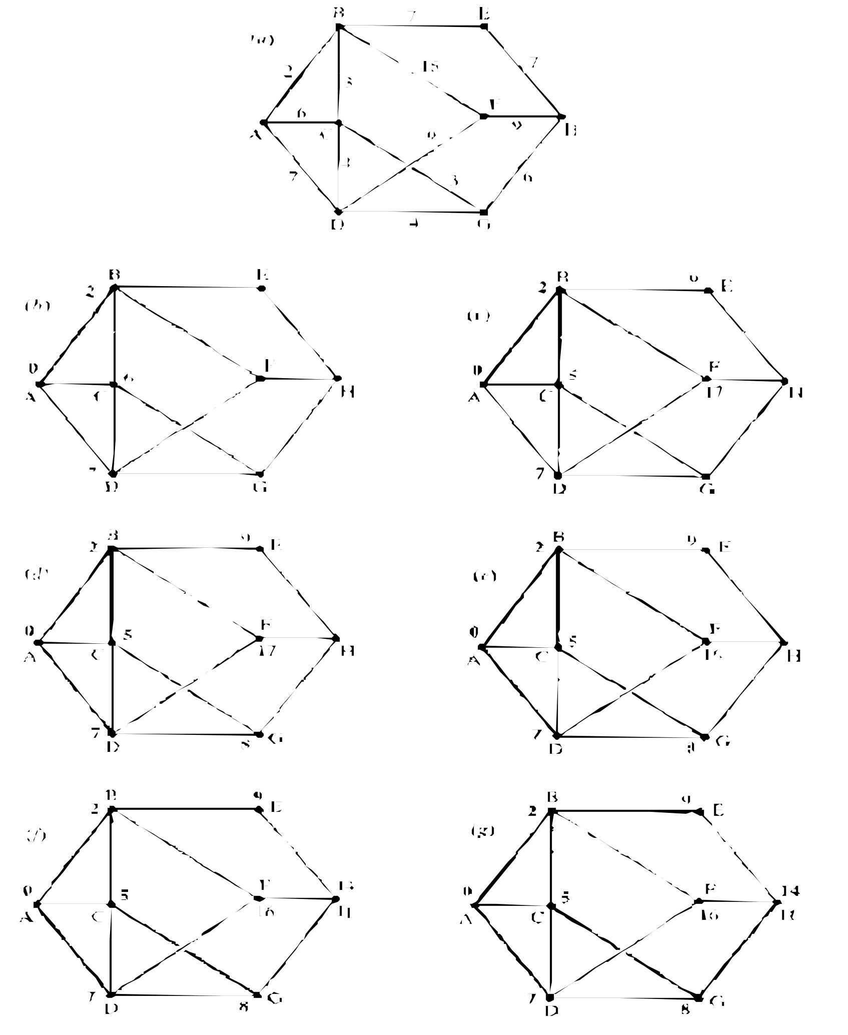


Figure 8-1 Example Dijkstra's shortest path algorithm.[6]

Node *A* receives the label 0, all others the label 'infinite'. Nodes *B, C* and *D* receive the provisional labels 2, 6 and 7. Node *A* is now made definitive, that means that it is removed from the set *T.*

Now *L(B)* = 2 is minimal. Nodes *E, F* and *C* are directly connected to *B.* We now set *L(E)* = 9 and *L(F)* = 17. Node *C* already has the provisional label *L(C)* = 6. Since this provisional label exceeds *L(B)* + 3, *L(C)* is replaced by the new provisional value of 5. Node *B* now becomes definitive with *L(B)* = 2 and is removed from set *T.*

Node *C* from set *T* now has the smallest label. We check the links coming from *C*, where necessary adjust the labels of the nodes that are connected to *C* and make the label of *C* definitive.

This process is repeated until *L(H)* achieves a definitive label value. This label value (14) indicates the length of the shortest path from *A* to *H.* Table 8-1 shows the course of the algorithm. Definitive labels are underlined.

A	*B*	*C*	*D*	*E*	*F*	*G*	*H*	Figure 8-1
$\underline{0}$	∞	∞	∞	∞	∞	∞	∞	a
$\underline{0}$	$\underline{2}$	6	7	∞	∞	∞	∞	b
$\underline{0}$	$\underline{2}$	$\underline{5}$	7	9	17	∞	∞	c
$\underline{0}$	$\underline{2}$	$\underline{5}$	$\underline{7}$	9	17	8	∞	d
$\underline{0}$	$\underline{2}$	$\underline{5}$	$\underline{7}$	9	16	$\underline{8}$	∞	e
$\underline{0}$	$\underline{2}$	$\underline{5}$	$\underline{7}$	$\underline{9}$	16	$\underline{8}$	14	f
$\underline{0}$	$\underline{2}$	$\underline{5}$	$\underline{7}$	$\underline{9}$	16	$\underline{8}$	$\underline{14}$	g

Table 8-1 Solution of the shortest path example.

If, when labelling a node, we remember via which link this node was most recently labelled, we also know at the finish of the algorithm which links constitute the shortest path to this node. The algorithm above stops as soon as we have reached the destination node. If we pursue the algorithm until all nodes have received a definitive label, in other words, when we take T = 'empty' as the criterion at which we halt the algorithm, a shortest-route tree will be built up from the start-point to each node in the network.

8.1.2 Classification of traffic assignment models

The demand for transport, given as trips in the OD-table, varies with time. Similarly, network characteristics may vary over time, be it as a function of transport demand or not. Traffic assignment models can in the first instance be classified according to this aspect of time:

- *Static traffic assignment models* assume that transport demand and supply are time-independent. The traffic flows in the network that are calculated using these static models, therefore, do not change over time and are, in fact, the flows that would emerge if the transport demand remained constant over a sufficiently long time-span. We may express this by saying that traffic is assigned to the entire route between an origin and a destination. Other commonly used terms for static models are steady state or 2-dimensional (2D) models. A steady-state flow is a flow that does not change over time. The term 2D indicates an assignment in the 2-dimensional area of the network, and that the dimension of time has been left out.
- *Dynamic traffic assignment models* do take account of variation in transport demand and with possible changes in the characteristics of the network. As a result, flows on the links in the network are calculated that vary over time. Another name for dynamic models is 3D or 3-dimensional models.

The dynamic models are still in the research stage and are not much used yet, in practice. In this text, we confine ourselves to static traffic assignment models, which have been subject to long experience.

The simplest assignment method is the *all or nothing traffic assignment model*. This method assigns all trips to one route, namely the shortest. No account is taken of the changes in travel resistance due to network loading. This model assumes that each road user is fully aware of the travel impedance of all possible routes and that they weigh them in an equal manner.

In reality, however, several routes between an origin and a destination are used, even when the network is not heavily congested. This is due to the fact that not every road user is fully acquainted with all travel impedances in the network. Travel impedance of the links, moreover, is judged differently by different traffic participants.

This effect can be accounted for by assuming that the perception of traffic impedance varies according to some statistical probability distribution in the population of traffic participants. This leads to the group of *stochastic traffic assignment models*. Incomplete and varying levels of knowledge amongst the road users regarding travel impedances in the network are taken into account. This can be done in two ways. The first method uses theoretical probability curves to lead traffic along alternative routes; the second uses simulation.

As in the all-or-nothing method, the stochastic model does not take into account the changes in traffic impedance that are due to network congestion.

The models that do take changes in travel impedance of a link due to congestion into account are *equilibrium models*. An equilibrium model is used to make assignments in road networks experiencing congestion. The prime characteristic is that the travel impedance (more particularly the time component of the impedance) of a link is a function of the traffic load. If the load on the links that initially constituted the shortest route increases, traffic will search for alternative routes. This will eventually lead to an equilibrium in which travellers will be unable to improve travel time by unilaterally choosing another route. Unilateral action means: without communication, co-operation or agreement with others. This is the equilibrium formulated by Wardrop in 1952 (1st principle of Wardrop), that equilibrium models try to calculate. Wardrop formulated a second type of equilibrium, the so-called system-optimal equilibrium that will be dealt with later on in the chapter.

Finally, the stochastic traffic assignment model and the equilibrium model can be combined. *Stochastic equilibrium models* take account of the effect of congestion and of the differences in perception of traffic impedance by the road users.

		Stochastic effects taken into account?	
		no	yes
Capacity effects taken into account?	no	all-or-nothing assignment	stochastic assignment
	yes	equilibrium assignment	stochastic equilibrium assignment

Table 8-2 Classification of static assigment models

8.1.3 Notation

Indexes

i = origin

j = destination

a = link in the network

r = route (successive links)

Variables

T_{ij} = number of trips per time-unit from i to j

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

c_{ij} = travel impedance from i to j

c_a = travel impedance for link a

$\hat{c}_a$ = system impedance (= $q_a.c_a$) for link a

$\tilde{c}_a$ = marginal system impedance (= $d\hat{c}_a \,/\, dq_a$) for link a

t_{ij} = travel time from i to j

t_a = travel time for link a

q_a = number of trips per time-unit on link a (intensity)

The concept of travel impedance has been discussed in the chapter on distribution. The travel impedance of a trip is written as a linear combination of the duration and costs experienced by the traveller. In a network link one often uses a linear combination of travel time and length of the link. The length of the link then becomes a measure for the travel costs over the link.

8.1.4 All-or-nothing assignment

The all-or-nothing traffic assignment model is a very simple assignment method. Figure 8-2 shows an example.

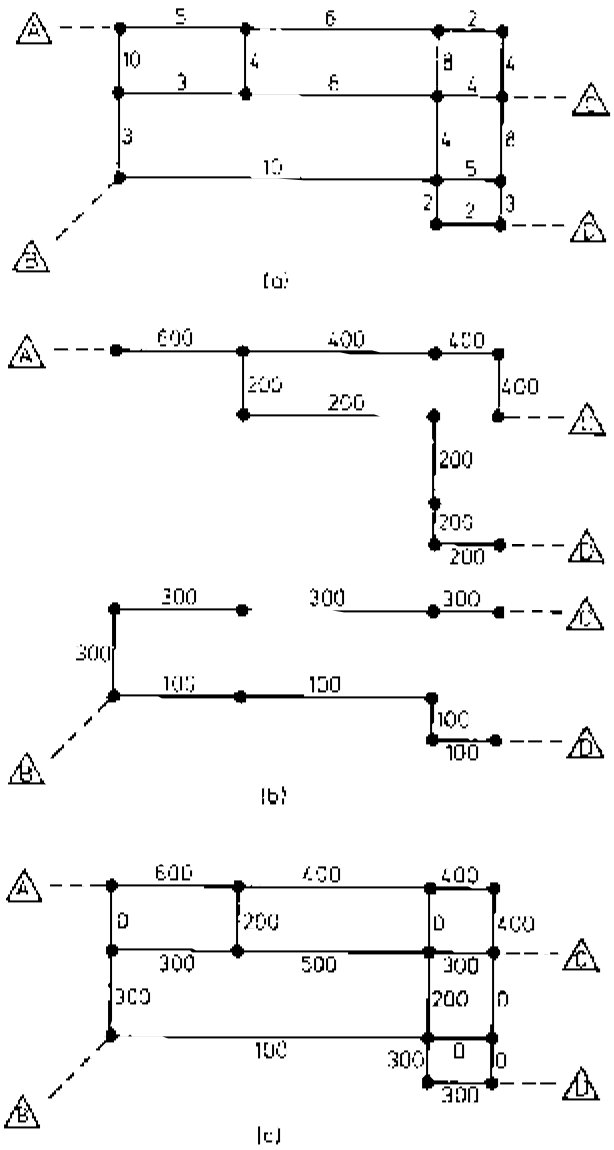


Figure 8-2 Example all-or-nothing assignment.[2]

Figure 8-2a shows the network with the travel impedances along the links. The OD-table that needs to be assigned to this model is as follows:

	C	D
A	400	200
B	300	100

The trips in the OD-table are successively assigned per OD-pair to the shortest route for that OD-pair Journeys that end up on the same link in the network are added. Figure 8-2b shows the shortest route-trees coming from the nodes A and B, with the trips assigned. Figure 8-2c shows the end-result of the assignment.

All-or-nothing traffic assignment models can give reasonable results in a network where there is no congestion and which offers few alternative routes between an origin and a destination, and when, moreover, these few alternative routes show great differences in traffic impedance. The prime function of an all-or-nothing traffic assignment model, however, is its function as a building block in more advanced traffic assignment methods, as will become clear further on in this chapter.

8.1.5 Stochastic assignment

Consider a group of travellers who want to move from a given origin to a given destination. Assume that they can choose between a large number of routes. Each route is characterised by a specific total traffic impedance that can be measured objectively. In an all-or-nothing traffic assignment model we assume that everyone will take the shortest route. However, due to differences in perception, in knowledge and in other factors there will be a divergence of opinion within the group regarding the various levels of traffic impedance. This will cause a divergence in the routes that are taken. It is this effect that a stochastic traffic assignment model tries to account for. In principle, there are two groups of methods that can do this:

8.1.5.1 Stochastic traffic assignment with theoretical probability functions.

These methods use a theoretical probability model, such as a logit model for example, to distribute the travellers across the alternative routes. These methods have a number of disadvantages.

The first drawback concerns the definition of the alternative routes. Realistic networks generally show a very large number of routes between an origin and a destination. It is hardly possible to include all these alternatives in a theoretical probability model. In some of these methods, therefore, so-called *reasonable routes* are determined. Again, there are a number of ways in which this can be done. It is possible, for example, to determine not only the shortest route, but also the second shortest route, the third-shortest, etc. Another method to determine a reasonable route is as follows: a reasonable route is defined as a sequence of nodes, whereby each subsequent node is located further away from the origin and closer to the destination. These methods are both rather arbitrary and the necessary algorithms tend to be complex.

The second drawback is that the methods that depend on theoretical probability models are sensitive to the way in which the network has been defined. The problems that can be encountered when one applies the logit model to route choice have been illustrated in chapter 3.5. These are connected to the assumed identical and independent error terms in the derivation of the logit model.

Because of the drawbacks mentioned above, stochastic traffic assignment models based on theoretical probability curves are now rarely used. We will, therefore, not discuss them further in our text.

8.1.5.2 Stochastic traffic assignment models based on simulation

The prime characteristic of these models is that the traffic impedance of a link is determined using a random number generator (Monte Carlo simulation).

The limited knowledge held by road users regarding the traffic system can be modelled by defining the levels of link traffic impedance as stochastic variables.

$$C_a = c_a + e_a$$

The subjective travel resistance C_a of a link is a stochastic variable and is equal to the objectively measurable traffic impedance c_a increased by a stochastic error term e_a. $E(e_a) = 0$ applies to the mean value of the error term, so that c_a is the mean value of the subjective travel impedance C_a. The objective travel impedance c_a can be interpreted as the travel impedance used in the all-or-nothing traffic assignment.

A Normal distribution is usually taken for the probability distributions of the error terms. More important than the shape of the probability distribution, however, is the specification of the variance of the error term. In order to ensure that the variance at route level remains independent of network coding, it is necessary that the variance is proportional to the link impedance. In other words, the dispersion must be proportional to the square root of the link impedance.

Keeping in mind the remarks above regarding shape and variance of the probability distribution of the error term, one arrives at the following, often used formula for the link impedance:

$$C_a = c_a + z\sqrt{j\ c_a}$$

with:

c_a = the objectively measurable link impedance
z = random number from a (pseudo) normal $N(0,1)$-distribution
j = a factor determining the size of the variance
C_a = subjective link impedance

8.1.5.3 Algorithm for the stochastic traffic assignment model

The algorithm for the stochastic traffic assignment model based on simulation comes down to the following. Applying the formula above we draw lots using a random number generator (Monte Carlo simulation) to determine the subjective link impedance for each link in the network. We carry out an all-or-nothing assignment on this network with subjective link impedances, and we repeat this a number of times. Due to the element of chance, the shortest routes through the network will vary with each draw. This causes a distribution of the traffic flows across a number of routes.

We obtain the desired stochastic traffic assignment by taking the average of a number of all-or-nothing traffic assignments carried out in this way. The number of iterations can be fixed to a previously agreed number *N*. In that case, after each draw of subjective link impedances, *1/N* of the total travel demand from the OD-table is assigned to the network using an all-or-nothing assignment and finally all partial assignments are added.

The algorithm below, however, is a better choice. After each draw of the subjective link impedances the average is taken of the all-or-nothing assignments of the entire OD-table up to that point. This process is repeated until the stopping criterion has been reached. The stopping criterion could apply when the flows achieved after an iteration no longer diverge much from the flows calculated in the previous iteration. This is when convergence is said to have been achieved.

```
$i = 0$
$q_a^{(i)} = 0$
repeat
        $i = i + 1$
        Determine $C_a$ by drawing lots
        Detrmine flows $Q_a$ by an all-or-nothing assignment using impedances $C_a$
        $f = 1 / i$
        $q_a^{(i)} = (1 - f)\, q_a^{(i-1)} + f\, Q_a$
until $stopping\ criterion = true$
```

The outcome of using $f = 1 / i$ in the algorithm above is that each newly calculated $q_a^{(i)}$ is the average of all flows Q_a drawn until that point.

8.1.6 Equilibrium assignment

In the section above, we incorporated the differences between individual road users in our calculation by means of a stochastic traffic assignment model. This can explain why road users choose different routes in otherwise equal circumstances. The reason is that not everyone shares the same opinion as to what constitutes the shortest route.

There is yet another reason why traffic between a specific origin-destination pair distributes itself across several routes. As soon as the traffic volume of this initially shortest route increases, travel time, and thus traffic impedance, increases. Travel impedance on this originally shortest route can increase to such levels that other routes become options for the journey. This is the subject of this section. Note that this section sets out from the assumption that there are no differences between road users! So there are no stochastic elements involved. This is why the subject to be discussed is also called a deterministic equilibrium assignment.

8.1.6.1 User-optimal and system-optimal equilibrium assignment

If the load on a link in the network increases, then the travel impedance of a link also increases. How this can lead to the idea of equilibrium in transport networks can be shown by an example.

Assume that the number of road users who want to move from a given origin to a given destination is known. Also assume that origin and destination are connected by a number of routes. How will the road users distribute themselves over these routes?

If all of them were to take the shortest route (calculated over the unloaded network) then this route could become congested. This would lead to an increase in travel impedance along that route to the point where this route would no longer be the shortest one. Some road users would choose an alternative route. Congestion could also develop on the alternative route, etc. Eventually an equilibrium will be reached, as formulated for the first time in 1952 by *Wardrop*. (Wardrop's first principle)

Traffic distributes itself across the links of a network in such a way that an equilibrium occurs in which no individual road user can lower his travel resistance by unilaterally (independently of the other road users) choosing another route.

If all road users are fully aware of the travel impedances on all links (even if they do not use these links themselves), and if they also judge these travel resistances in the same way, Wardrop's first principle implies the following:

In the equilibrium situation all routes used between a given origin and destination have the same travel impedance, while routes not used have a higher level of travel impedance.

Wardrop's first principle describes the traffic flows that occur when each individual user strives to minimise his own travel impedance. It is a *descriptive* principle. The distribution of traffic flows that occurs is called a deterministic user-equilibrium or a deterministic *user-optimal equilibrium assignment.* The term deterministic points to the fact that stochastic link travel resistances have not been used.

When we multiply the travel impedance on a link by the intensity on that link we get the *system impedance* $\hat{c}_a$ on that link. This is the travel impedance experienced by all vehicles together on the link:

$$\hat{c}_a = q_a c_a$$

Summation of the system travel impedance across all links of the network leads to the *total system impedance* of the network. Since travel time and/or travel distance usually are the most important components of travel impedance, the total system travel impedance is a good measure for total fuel use (and the environmental pollution caused) of all vehicles together across the entire network.

$$\hat{c}_{totaal} = \sum_a q_a c_a$$

For any flow distribution or assignment we can determine the total system impedance. We could, for example, calculate the total system impedance for a deterministic user-optimal equilibrium assignment.

Assume that one assigns an OD-table to the network in such a way that the total system impedance is minimised. This kind of assigment of traffic over the links of the network is called a *system-optimal equilibrium assignment*. In this case we also have an equilibrium, but now instead of all used routes between the same origin and destination having the same travel impedance, they have the same marginal system impedance. We will return to this point in section 8.1.6.4.

The system-optimal equilibrium assignment, was also described by Wardrop and is sometimes referred to as Wardrop's second principle. Wardrop's second principle is not a descriptive, but a *normative* principle. An assignment according to Wardrop's second principle complies with a specific imposed norm, namely that of minimising total system impedance.

The term congestion-free networks when the traffic volume on a link does not influence travel time and thus travel impedance. This generally occurs, at least approximately, at low traffic volumes or when all network links have a high capacity. In section 8.1.6.4we will show that the user-optimal assignment and the system-optimal assignment in congestion-free networks are equivalent. Both are then equal to the flow distribution that is calculated with an all-or-nothing assignment.

When congestion does occur, the two traffic flow distributions differ from one another. The total system impedance will be higher in a user-optimal assignment than in the system-optimal assignment. Congested networks will naturally achieve an equilibrium according to Wardrop's first principle, purely because this principle describes usual human behaviour. A distribution of traffic flows according to the system-optimal assignment (desirable from a social view point, in terms of fuel-use restriction and environmental degradation) presents no stable equilibrium. It is possible, for example, that such a flow distribution gives individual traffic participants the chance to reduce their travel impedance, for example by unilaterally (independently) taking another route. A system-optimal distribution of traffic flow will, therefore, require specific enforced traffic measures (a toll levy, for example). To emphasise the contrast between both assignments, the user-optimal assignment is sometimes called a *selfish optimum*, and the system-optimal assignment a *social optimum.*

8.1.6.2 Time-loss functions

An increase in the traffic volume on a link leads to an increase in travel time and thus to an increase in the travel impedance on a link. On urban networks, we are not primarily concerned with congestion-effects on the link itself, but especially with the delays at intersections. The connection between traffic load and travel time is represented by a time-loss function. One of the most widely-used functions is the BPR-function (Bureau of Public Roads):

$$t_a = t_a^{free\ flow}(1 + a\left(\frac{q_a}{cap}\right)^b)$$

In this formula:

t_a = travel time on link a (including intersection delay)
$t_a^{free\ flow}$ = travel time on link a in an non-congested network ("free flow")
cap = "practical" capacity of link a
a, b = empirically determined coefficients

Common values for the coefficients are a = 0.15 en b = 4. Note that the *practical capacity* at this value of a in the formula above represents the level of traffic intensity whereby the travel time on the link is 15% higher than the travel time at free flow. The practical capacity, therefore, is something different from the maximal capacity of a link, that corresponds to the maximal traffic flow that a link can carry.

A time loss function usually is slightly rising at low volmes, and subsequently rises strongly when the practical capacity is exceeded.

Because we are dealing with travel impedance and not with travel time, we need to make another conversion to generalised travel time. This gives the following function that, for traditional reasons, we will also continue to call a time loss function.

$c_a = c_a(q_a)$

8.1.6.3 Algorithm for the user-optimal equilibrium assignment

Algorithms that are used to calculate equilibrium assignments are also called *capacity restraint* algorithms. They use an iterative procedure. A number of all-or-nothing assignments are, in fact, applied, and each new iteration uses the travel impedances acquired in the previous iteration. An *incremental assignment* method was formerly used where fractions of the OD-table were continuously "loaded onto" the network, until the entire table had been assigned. A better procedure is the one below where the entire OD-table is assigned to the network in each separate iteration. To ensure that the algorithm converges (that means that the results of an iteration differs less and less from the results of the previous iteration) it is necessary to use a weighting factor, whereby the results of the previous iteration(s) are incorporated into the next iteration.

```
$i = 0$
$q_a^{(i)} = 0$
repeat
        $i = i + 1$
        Be
        Determine $c_a$ by a time-loss function: $c_a = c_a( q_a^{(i-1)})$
        Determine flows $q_a^+$ by using an all-or-nothing assigment using impedances $c_a$
        Determine weighting factor $f$ $(0 < f < 1)$ (see below)
        $q_a^{(i)} = (1 - f) q_a^{(i-1)} + f q_a^+$
until stoppng criterion = true
```

Weighting factor

The weighting factor f in the algorithm above is used to find the new link volumes as a combination of the flows that were calculated in the previous iteration and the all-or-nothing volumes of the present iteration.

There are a number of options for the weighting factor:

- *Fixed weighting factor*
  For example: $f = 0.5$
  This is the simplest method. Convergence, however, is not guaranteed and even if the algorithm converges it needs a large number of iterations.
- *Diminishing weighting factor*
  For example: $f = 1 / (i + 0.5)$ or $f = 1 / i$
  With $f = 1 / i$ this is called the *Method of Successive Averages* (*MSA*). The effect of

this kind of weighting factor is that the new flows are the mean of the all-or-nothing assignment of all the iterations carried out up till this point. This method gives about the same results as the method with the optimal weighting factor (see below), but usually takes more calculation time than the optimal weighting factor.

- *Optimal weighting factor*
  The weighting factors used above have the following disadvantages. Convergence of the algorithm is not guaranteed when using a fixed weighting factor and the algorithm is inefficient. Although the algorithm does converge when a diminishing weighting factor is used, the method remains inefficient. An optimal weighting factor for which convergence is guaranteed and which is, moreover, very efficient, is based on the *Frank-Wolfe algorithm*. The Frank-Wolfe algorithm is used to solve a minimisation-problem. In the following paragraph we will investigate how solving a minimisation-problem is related to the determination of an equilibrium assignment.

Beckman transformation; equilibrium assignment formulated as a minimisation-problem

The Wardrop conditions for user-equilibrium can be written as follows:

$c_{ij}^{r} = c_{ij}$ when $T_{ij}^{r} > 0$
c_{ij}^{r} ³ c_{ij} when $T_{ij}^{r} = 0$

The link-loads are:

$$q_a = \sum_{ijr} T_{ij}^{r}\, d_{ij}^{ra} \qquad \textit{where:} \qquad d_{ij}^{ra} = 1 \;\; of \;\; d_{ij}^{ra} = 0$$

The parameters d_{ij}^{ra} indicate whether a route r between i and j does or does not use link a.

Now the total impedance between i and j using route r is:

$$c_{ij}^{r} = \sum_{a} c_a(q_a)\, d_{ij}^{ra}$$

As shown above, Wardrop's user-equilibrium can be written in the form of a large number of mathematical equations. Solving these equations gives the solution to the equilibrium problem, and thus finds the value of the flows on all the links. This, however, presents a problem: solving a system comprising of a great number of (non-linear) equations is an extremely difficult problem in numerical mathematics. In this context, "difficult" means that existing iterative solution techniques do not always converge, or if they do converge, they may take a long time.

We now use a "trick" which is very often applied in numerical mathematics. The problem of solving the equations that resulted from Wardrop's conditions is transformed to the process of solving an equivalent optimisation problem. This means that solving the optimisation problem also gives the solutions to the original system of equations. Why do we use such a roundabout way? Because finding the optimum of a function (a maximum or minimum) is a much simpler problem for which numerous different algorithms are available.

Intermezzo

Assume that we want to optimise a function $w = f(x,y)$ This can be done by taking the derivatives relative to x and y and setting them equal to zero.

$$\partial w / \partial x = 0$$
$$\partial w / \partial y = 0$$

We now solve this system of two simultaneous equations and use them to find the values of x and y for which the function w receives an optimal value. This example shows that an optimisation problem can be transformed to solving a system of equations. Alternatively, solving a system of equations can also be transformed to the problem of finding the optimum of an objective function. The objective function is that function of which the derivatives are the equations that have to be solved.

Beckman transformation

In 1956, Beckman proved that, reasoning along the lines given in the intermezzo above, solving the Wardrop equations for user-equilibrium is equivalent to solving the following minimisation-problem:

$$\min_{q_a} \sum_a \int_0^{q_a} c_a(q)dq$$

subject to:

$$\sum_r T_{ij}^r = T_{ij}$$

This expression indicates that the equilibrium flows q_a are achieved by choosing them in such a way that the sum over all links of the areas under the time loss functions from 0 to q_a is minimal.

We now return to determining the optimum weighting factor. When we described the algorithm for the users optimum equilibrium assignment, we saw that the result of an iteration is written as follows:

$$q_a^{(i)} = (1 - f)\, q_a^{(i-1)} + f\, q_a^{+}$$

The Frank-Wolfe algorithm can now be described as an algorithms that finds an optimal weighting factor f in such a way that at each iteration the largest possible decrease in the value of the above mentioned objective function is achieved. We refer to Sheffi (1985)[5] for more details regarding the functioning of the Frank-Wolfe algorithm.

8.1.6.4 Algorithm for the system-optimal equilibrium assignment

As discussed in paragraph 8.1.6.1, the system optimal assignment is the assignment of flows in the network that minimises the total system resistance. Therefore we need to solve the following minimisation problem:

$$\min_{q_a} \sum_a q_a * c_a(q_a)$$

The following is true for a general differentiable function *f(q):*

$$f(q_a) = \int_0^{q_a} f'(q)\,dq$$ where *f¢(q)* is the derivative of *f(q)*.

When we apply this to the objective function above we get the following minimising problem:

$$\min_{q_a} \sum_a \int_0^{q_a} \tilde{c}_a(q)\,dq \quad \text{where:} \qquad \tilde{c}_a(q) = \frac{d}{dq}(q * c_a(q))$$

The function $\tilde{c}_a(q)$ is the so-called *marginal system impedance function* for link *a*, and can be interpreted as the increase in system impedance on link *a* (the travel impedance experienced by all road users together), for an infinitesemal increase in intensity from *q* to *q* + *dq* on link *a*. If a single new traveller arrives on link *a* the travel impedance on the link increases. Not only the new traveller experiences this as a nuisance, but also all the road users who already were on the link.

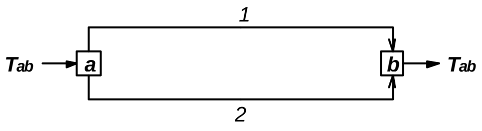
The objective function above shows a remarkable similarity to the objective function that we had to minimise in order to find a user-optimal equilibrium assignment. The only difference is that the time loss function $c_a(q)$ has been replaced by the marginal system impedance function $\tilde{c}_a(q)$). This means that we can apply the algorithm of the user-optimal equilibrium assignment to find a system optimal assignment, as described in section 8.1.6.3, on the understanding that we replace $c_a(q)$ by $\tilde{c}_a(q)$). It also means that in a system-optimal equilibrium assignment the marginal system impedance is equal for all used routes between the same origin and destination, not the normal travel impedance (as is the case in a user-optimal assignment).

It now also becomes obvious that in congestion-free networks (meaning networks where all links experience no influence of traffic volume on the travel impedance) the user-optimal assignment and the system-optimal assignment are equal. For in that case $c_a(q)$ is a constant function, and $c_a(q) = \tilde{c}_a(q)$ applies for all *q*.

8.1.6.5 Numerical example equilibrium assignment

The ideas that play an important role in the equilibrium assignment will be discussed in a detailed example, taken from Ortúzar and Willumsen[2]. Figure 8-3shows a simple network. Locations *a* and *b* are connected by links *1* and *2*. The time-loss functions for the two links are indicated, and shown in the figure. (For the sake of clarity, the time-loss functions have been kept simple in order to make the calculations easier to understand.) The number of trips from *a* to *b* is T_{ab}.

Since the network is so simple, the calculations can be written down analytically. This is not possible for complicated networks, when the iterative algorithm in section 8.1.6.3 needs to be used.



c_1 = 10 + 0.02 q_1

c_2 = 15 + 0.005 q_2

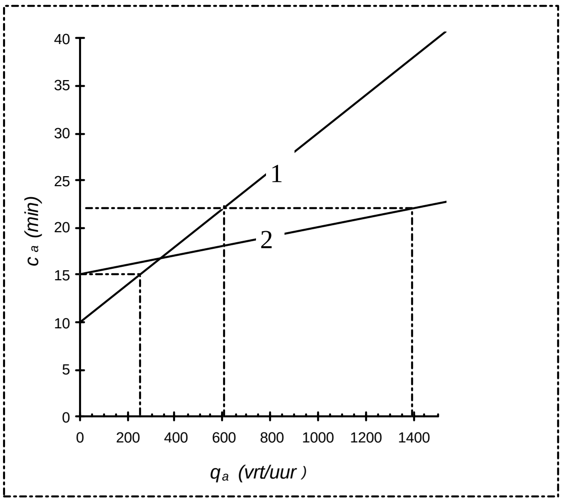


Figure 8-3 Example equilibrium assignment

<u>Question 1:</u>

Calculate the intensities on the two links for various values of $T_{ab.}$

a. T_{ab} < 250

For T_{ab} < 250 it is obvious that all traffic will use link *1*. This is because in that case the travel impedance c_1 < 15, which is smaller than the travel impedance that can ever be achieved on link 2.

b. T_{ab} > 250

As soon as T_{ab} > 250, the traffic will be distributed over both routes in such a way that the travel impedance on both routes is identical (Wardrop's 1st principle) and that the sum of both flows equals T_{ab}.

Assume, for example T_{ab} = 2000
then:

$$10 + 0.02q_1 = 15 + 0.005q_2$$
$$q_1 + q_2 = 2000$$

which leads to: q_1 = 600, q_2 = 1400 and c_1 = c_2 = 22.
This solution is shown in the graph in Figure 8-3.

Question 2:
Find the user-optimal assignment for T_{ab} = 2000 using the Beckman transformation.

We need to solve:

$$\min_{q1,q2}\{ \int_0^{q_1} (10 + 0.02q)dq + \int_0^{q_2} (15 + 0.005q)dq \} =$$
$$\min_{q_1,q_2} \{ 10q_1 + 0.01q_1^2 + 15q_2 + 0.0025q_2^2 \} =$$
$$\min_{q1} \quad \{ 10q_1 + 0.01q_1^2 + 15\,(2000 - q_1) + 0.0025\,(2000 - q_1)^2 \}$$

By taking derivates and setting them to equal to zero, we find that the objective function is minimal for q_1 = 600. Therefore q_2 = 1400 and c_1 = c_2 = 22. This is the same solution that we found in question *b*.

In this simple case we can also use a graph to show that the equilibrium flows are achieved by choosing them in such a way as to minimise the sum of the areas under the time-loss functions, as elaborated in the discussion of the Beckman transformation. See Figure 8-4. Both time-loss functions are depicted in one graph. The horizontal axis has been given two scales in opposite directions in such a way as to make the sum of the flows q_1 and q_2 equal to 2000. It is easy to see that the sum of the areas under the time-loss functions is minimal for c_1 = c_2, with the associated q_1 = 600 and q_2 = 1400. At another combination of flows, q_1 = q_2 = 1000 for example, the sum of the areas increases by the vertically hatched part.

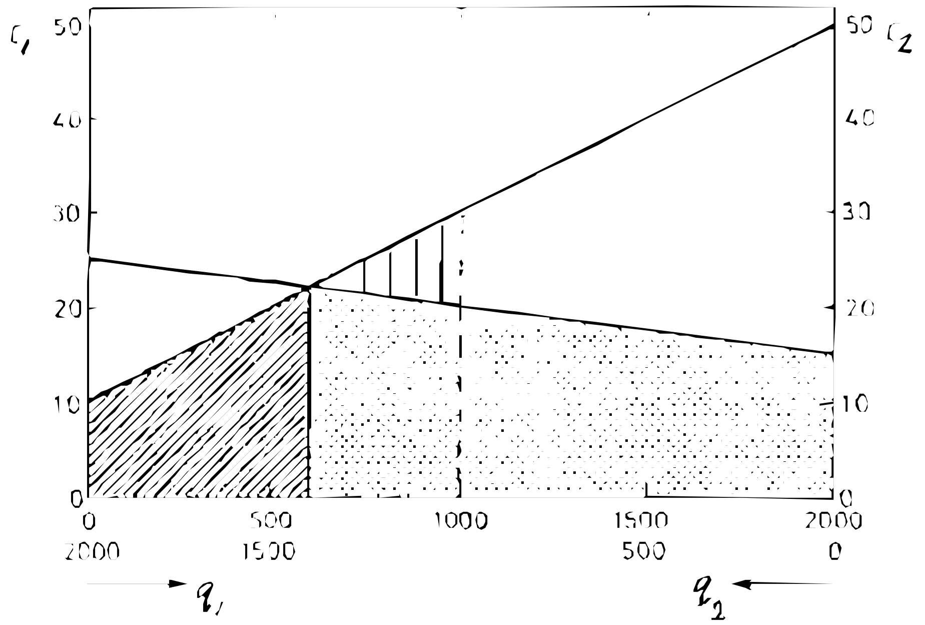


Figure 8-4 Area under time-loss functions.[2]

Question 3:

Calculate the system-optimal assignment for T_{ab} = 2000

The system impedances on the links are:

$$\hat{c}_1 = q_1\,(10 + 0.02 q_1)$$
$$\hat{c}_2 = q_2\,(15 + 0.005 q_2)$$

The marginal system impedances are:

$$\tilde{c}_1 = \frac{d\hat{c}_1}{dq_1} = 10 + 0.04 q_1$$
$$\tilde{c}_2 = \frac{d\hat{c}_2}{dq_2} = 15 + 0.01 q_2$$

In the system-optimal equilibrium assignment the marginal travel impedances along both routes are equal (Wardrop's 2nd priniple).

$$10 + 0.04 q_1 = 15 + 0.01 q_2$$
$$q_1 + q_2 = 2000$$

This leads to: q_1 = 500, q_2 = 1500 en $\tilde{c}_1 = \tilde{c}_2 = 30$

<u>Question 4:</u>

Find the system-optimal assignment for T_{ab} = 2000 using the Beckman transormation.

We need to solve:

$$\min_{q1,q2}\{ \int_0^{q_1}(10+0.04q)dq + \int_0^{q_2}(15+0.01q)dq \} =$$
$$\min_{q_1,q_2}\{ 10q_1 + 0.02q_1^2 + 15q_2 + 0.005q_2^2 \} =$$
$$\min_{q1}\ \{10q_1 + 0.02q_1^2 + 15(2000-q_1) + 0.005(2000-q_1)^2 \}$$

Differentiating and setting the derivative equal to zero we find that the objective function is minimal for q_1 = 500. Therefore q_2 = 1500 and $\tilde{c}_1 = \tilde{c}_2 = 30$. This is the same solution as the we found in question 3.

The results of the calculations for T_{ab} = 2000 are shown in Table 8-3. The travel impedances, marginal travel impedances and system impedances are included for the user-optimal and the system-optimal assignment.

	user-optimal			system-optimal		
	link 1	link 2	total	link 1	link 2	total
Flow	600	1400	2000	500	1500	2000
Impedance	22	22		20	22.5	
Marginale impedance	34	29		30	30	
System impedance	13200	30800	44000	10000	33750	43750

Table 8-3 Results equilibrum assignment example

Note in the system-optimal assignment that the total system impedance is less by 250 vehicle-costminutes per hour compared to the total system impedance in the user-optimal assignment. Although this is socially desirable (saving on fuel costs etc.) the associated flow distribution over the links cannot be forced without additional measures. The travel impedance along link 1 in the system-optimal assignment is, in fact, smaller than that along link 2. This will lead some link 2 users to change to link 1. The equilibrium of the user-optimal assignment eventually would occur, whereby the travel resistance along both routes is equal to 22 minutes of generalised travel time.

8.1.7 Stochastic equilibrium assignment

A variation of Wardrop's first principle applies to the most advanced (and most realistic) static assignment method, namely the stochastic equilibrium assignment (formulated by Daganzo in 1977):

Traffic distributes itself across the links of a network in such a way that an equilibrium occurs in which no individual road user thinks that he can lower his travel resistance by unilaterally (independently of the other road users) choosing another route.

Take note of the underlined word in the above definition. It is about the personal perception of road users. System impedance is higher in a stochastic equilibrium assignment than in a deterministic equilibrium assignment. But it will approach it as the uncertainty of the road users regarding the travel impedances in the network decreases.

The algorithm that we give consists of a combination of the algorithms of the equilibrium assignment and the stochastic assignment. The algorithm is identical to the algorithm of the equilibrium assignment, with one exception: the "do an all-or-nothing assignment" is replaced by the "do a stochastic assignment". We already discussed the way in which a stochastic assignment is done earlier in this chapter.

```
$i = 0$
$q_a^{(i)} = 0$
repeat
    $i = i + 1$
    Determine $c_a$ with time-loss function: $c_a = c_a( q_a^{(i-1)})$
    Determine flows $q_a^+$ by a stochastic assignment using impedances $c_a$
    Determine weighting factor $f$ ( $0 < f < 1$)
    $q_a^{(i)} = (1 - f)\, q_a^{(i-1)} + f\, q_a^+$
until stopping criterion = true
```

The MSA method can be used to determine f in the algorithm above. Faster convergence can be achieved by adaptation of the objective function that is to be minimised using the Frank-Wolfe algorithm. See Sheffi (1985)[5].

8.2 Traffic assignment models for public transport networks

There is an essential difference between a network for public transport and one for private transport, such as a car network. The public transport network is based on *lines*, where services are maintained by a number of vehicles. The capacity of a line is linked to the passenger capacity of the vehicle and the frequency of the service.

8.2.1 Travel impedances in public transport

The travel time from an origin to a destination by public transport comprises the following components:

- feeder transport from the origin to the busstop or station,
- waiting time at the busstop or station,

- travel time in the vehicle,
- possible transfer waiting-time,
- getting from the busstop or station to the destination address.

These components are accounted for in the network by the introduction of links for transport to and from busstops and stations and transfer links. The attributes of these links consist of the travel times (walking, cycling, etc) involved in getting to and from busstops and stations and waiting times.

In the chapter about distribution, we already saw that, though all time components can be measured objectively in minutes, not all travellers perceive these components identically. A minute spent waiting, for example, is perceived as a greater travel impedance than a minute of effective travel time in the vehicle.

To determine the travel impedances, financial costs converted to generalised travel time, must be added to the time components mentioned above.

8.2.2 Waiting times

We use the following notation:

f	=	frequency (vehicles per time-unit)
h	=	interval between two vehicles (=1/f)
t	=	travel time between stops
w	=	waiting time
l	=	travellers per time unit (demand)

The time spent waiting at a stop that is served by one service depends on the interval h between two vehicles (i.e. on the frequency of the service) and on the variance of the intervals.

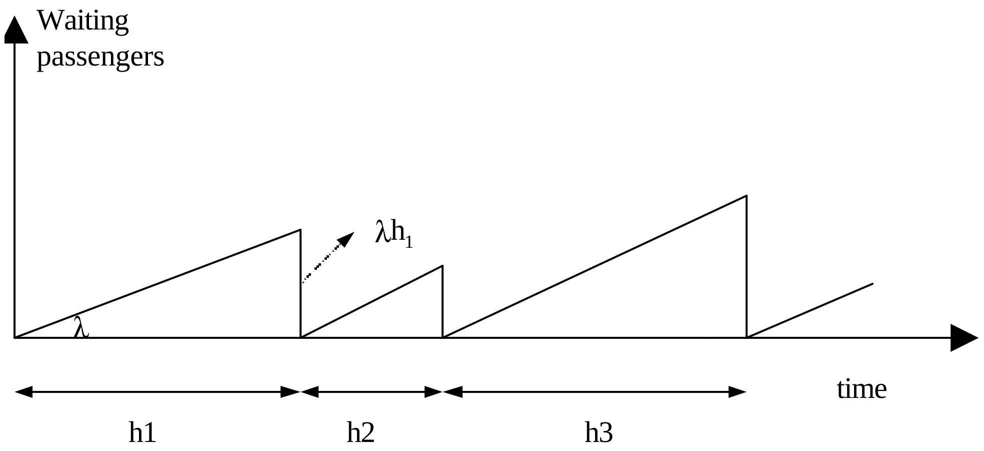


Figure 8-5 Waiting passengers at a public transport stop

Assume that the arrival pattern at a stop is l passengers per time-unit, uniformly distributed over time. Consider *K* intervals. The aggregate waiting time *W* for all travellers equals the surface of the triangles in Figure 8-5.

$$W = \sum_{k=1}^{K} \tfrac{1}{2}\, l \ \ h_k^2$$

The following applies to the number of travellers *N* that arrives at the stop:

$$N = \sum_{k=1}^{K} l \ \ h_k$$

The average waiting time per traveller, therefore, is:

$$\overline{w} = \tfrac{1}{2} \sum_{k=1}^{K} h_k^2 \,/\, \sum_{k=1}^{K} h_k$$

We can now write this as a function of the means of h^2 and h:

$$\overline{w} = \tfrac{1}{2}\,(\overline{h^2} \,/\, \overline{h})$$

A general expression for the sample variance S_h^2 is:

$$S_h^2 = \overline{h^2} - \overline{h}^2$$

So the mean waiting time can be written as:

$$\overline{w} = \tfrac{1}{2}(\overline{h} + S_h^2 \,/\, \overline{h})$$

Half of the interval is often used for the mean waiting time (when travellers arrive uniformly over time). The derivation above shows that this applies, strictly taken, only when the variance of the intervals equals zero, i.e. in the case of a completely regular service. In the case of random arrivals, the average waiting time exceeds half of the mean interval duration!

> Example:
>
> Assume:
>
> mean duration of interval $\overline{h} = 10 \text{ min}$ and standard deviation $S_h = \sqrt{10}$, therefore $S_h^2 = 10$
>
> Then: $\overline{w} = ½ * (10 + 10/10) = 5.5 \text{ min}$

Although this is a most interesting outcome, it can be accepted that most services are fairly regular which leads to a mean waiting period equal to half the duration of the interval period.

If the service frequency is low, the arrival time of travellers at the stop will not be uniformly divided, but will reflect the expected departure time of the bus or train. Observations have shown that there is an upper limit to the mean waiting time of 5 to 10 minutes, say 7.5 minutes.

Waiting times at parallel lines
It is possible that a route between two nodes is serviced by one or more lines. We give an example of the kind of reasoning that can be used to calculate the mean waiting times in that case:

Case 1
The simplest case is the one where the travel times between the two stops are equal for all lines. It is then usual to set the average waiting time for all lines equal to half of the interval time that appertains to a frequency equal to the sum of the frequencies of the individual lines. Passengers, for example for line 1, will on average not have waited longer than that, since it would otherwise have been better to take one of the other lines. However, be aware that the approach above can lead to erroneous results if the lines depart at fixed intervals.

Case 2
Assume that the travel times between the stops differ for the different lines. A lower line frequency, for example, could be compensated by a shorter travel time. It is easy to see that in such cases frequencies can not be added up. Calculating the mean waiting time can now become very complicated.

Assume that there are two lines with $t_1 < t_2$. The travellers will then show a preference for line 1, unless t_2 exceeds t_1 only by a small amount. If $t_2 < t_1 + ½\, h_1$, some travellers will take line 2 all the same. We could now use the following approach:
if $t_1 = t_2$ then add the frequencies, if $t_2 > t_1 + ½\, h_1$ then do not add them at all and apply a special calculation in between these two cases.

One often refrains from this refinement and, in the case of unequal travel times between nodes, calculates waiting times equal to half the interval time of the individual lines.

8.2.3 Shortest route in a public transport network

The usual algorithms that are used to find a shortest route in a road network require adjustment when applied to a public transport network. The option of transfers and associated waiting times cause anomalies, as illustrated in Figure 8-6.

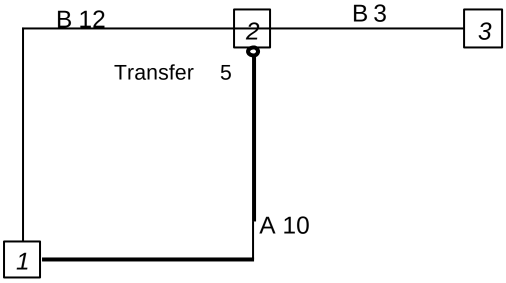


Figure 8-6 Problem finding shortest route by usual algoritms in publ. tr netw.[7]

The journey from node *1* to *3* via line B takes 12 + 3 = 15 minutes. Via line A and B with a transfer at node *2* it takes 10 + 5 + 3 = 18 minutes. Therefore, the shortest route from *1* to *3* follows line B, but the shortest route from *1* to *2* follows line A. This means that, depending on the destination, node *2* can be reached by various routes. This leads to problems in terms of the classical algorithms, because there is no unambiguous shortest route tree from node *1*. One solution would be, for example, to introduce additional links after which the usual shortest route algorithms could be applied. Another solution is to develop specific algorithms for public transport networks.

To give an idea of an algorithm that was specifically developed for public transport we will describe a method that is known under the name *Transitnet* (described in Lamb en Havers (1970)[8].

The transitnet algorithm is illustrated by an example. See Figure 8-7. Say that we want to find/determine the shortest route tree from node 1. For the sake of simplicity we will leave the waiting times out of the consideration.

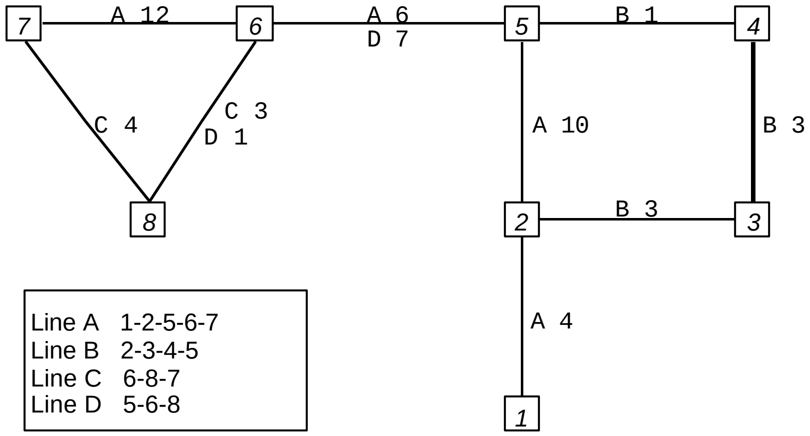


Figure 8-7 Example finding shortest route in publ tr network.[8]

Essentially, the algorithm works as follows:

We first find all the nodes from the starting point that can be reached without transfers and the time required for these journeys. Next we find all the nodes that can be reached by one transfer. If the time involved is smaller that that for a route we found earlier, we adjust the time and route. The same procedure is repeated for two transfers, etc, until the identified routes to all nodes no longer change.

In more detail:

Each node receives a label of the form (L.K.T.) with the following meaning:

L the line last taken to reach this node
K the node where the transfer to this line is made
T the time needed to reach this node

Moreover each node also has a [+] or [-] sign that indicates if there are of lines emanating from this node that still need to be investigated.

		Knooppunten							
Lijnen vanuit knoop	Lijn	1	2	3	4	5	6	7	8
		0.0.0	0.0.∞	0.0.∞	0.0.∞	0.0.∞	0.0.∞	0.0.∞	0.0.∞
		+	-	-	-	-	-	-	-
1	A		A.1.4			A.1.14	A.1.20	A.1.32	
		0.0.0	A.1.4	0.0.∞	0.0.∞	A.1.14	A.1.20	A.1.32	0.0.∞
		-	+	-	-	+	+	+	-
2	B			B.2.7	B.2.10	B.2.11			
		0.0.0	A.1.4	B.2.7	B.2.10	B.2.11	A.1.20	A.1.32	0.0.∞
		-	-	+	+	+	+	+	-
3	--								
		0.0.0	A.1.4	B.2.7	B.2.10	B.2.11	A.1.20	A.1.32	0.0.∞
		-	-	-	+	+	+	+	-
4	--								
		0.0.0	A.1.4	B.2.7	B.2.10	B.2.11	A.1.20	A.1.32	0.0.∞
		-	-	-	-	+	+	+	-
5	A D	A.5.25	A.5.21				A.5.17 D.5.18	A.5.29	D.5.19
		0.0.0	A.1.4	B.2.7	B.2.10	B.2.11	A.5.17	A.5.29	D.5.19
		-	-	-	-	-	+	+	+
6	C D					D.6.24		C.6.24	C.6.20 D.6.18
		0.0.0	A.1.4	B.2.7	B.2.10	B.2.11	A.5.17	C.6.24	D.6.18
		-	-	-	-	-	-	+	+
7	A	A.7.56	A.7.52			A.7.42	A.7.36		
		0.0.0	A.1.4	B.2.7	B.2.10	B.2.11	A.5.17	C.6.24	D.6.18
		-	-	-	-	-	-	-	+
8	C						C.8.21	C.8.22	
		0.0.0	A.1.4	B.2.7	B.2.10	B.2.11	A.5.17	C.8.22	D.6.18
		-	-	-	-	-	-	+	-
7	A	A.7.54	A.7.50			A.7.40	A.7.34		
		0.0.0	A.1.4	B.2.7	B.2.10	B.2.11	A.5.17	C.8.22	D.6.18
		-	-	-	-	-	-	-	-
		0.0.0	A.1.4	B.2.7	B.2.10	B.2.11	A.5.17	C.8.22	D.6.18

Table 8-4 Transitnet algoritm.[8]

The algorithm looks as follows:

Initialisation:
The starting node gets the label [0.0.0] and the indicator [+]. All other nodes get the label [0.0.∞] and the indicator [-].

Repeat as long as there are [+] indicators:

Step 1:
Choose a node with a [+] indicator.
Investigate all the lines that visit this node, excepting the line by which one arrived at this node. This yields new temporary labels for all the nodes that can be reached from this node without further transfers.

Set the indicator for this node to [-].

Step 2:
Update the labels and indicators.
Compare the new labels gained in step 1 to the existing labels. The labels of the nodes concerned are only adjusted and their indicators set at [+] if the new travel time is smaller than the travel time already achieved. The labels and indicators of the other nodes remain unchanged.

Table 8-4 shows the results when this algorithm is applied to the example.

The last line of the table holds all information necessary to reconstruct the shortest route tree. Say that we are looking for the shortest route from node 1 to node 8. Node 8's label (D.6.18) states that the earliest possible time we can arrive at this node is via line D from node 6 and that the journey took 18 minutes. In turn, we can reach node 6 (according to the last line in the table) in 17 minutes at the earliest and we can do so via line A from node 5. Working back like this to node 1, we find the shortest route given in Table 8-5.

Knooppunt	Tijd	Neem lijn
1	0	A
2	4	B
5	11	A
6	17	D
8	18	-

Table 8-5 Shortest route between two nodes in a public transport network.

This example took no account of waiting times. The description of the algorithm shows, however, that the introduction of waiting times and the associated weighting of time components will not be problematic.

8.2.4 Assignment models for public transport

The description of the assignment algorithms for car traffic showed that the methods can be divided into a number of categories:

- All-or-nothing assignment
- Stochastic assignment
- Equilibrium assignment
- Stochastic equilibrium assignment

The same algorithms can be adapted to be used for assignment in public transport.

For the all-or-nothing assignment, the same applies as was already pointed out for the assignment of road traffic: the method can deliver reasonable results in a non-congested network, when there are few alternative routes between an origin and a destination and when these few alternative routes also differ widely in terms of travel impedance.

As in a car network, several routes can exist between two points in a public transport network. A characteristic of the public transport network is that there are several alternative travel routes coinciding in location. Train travel, for example, offers a choice between express trains and local trains. Bus networks also offer different lines which partly follow the same route.

Even when these alternative routes do not share the same travel impedance, travellers will distribute themselves over these routes. This happens, just as is the case in car networks, because not every traveller is fully aware of the travel impedances in the network. Link travel impedances, moreover, are judged differently by different travellers.

Stochastic assignment models are used to find the distribution of travellers over alternative routes. This can be done in two ways. The first method distributes the travellers in function of the frequencies and/or travel times along the alternative routes; the second uses simulation (Monte Carlo method).

The first method has several options: the passenger flows can be distributed proportional to the line frequencies, or they can be distributed proportional to the e-powers of the weighted travel- and waiting times (logit model).

The second method is very similar to the simulation method used in the assignment in car networks. Drawing lots is now used to find the waiting times and the choice of lines with equal length along the same route. The shortest route tree is calculated after each draw of lots.

The all-or-nothing and stochastic assignments above take no account of changes in travel impedances caused by network overload (congestion).

In the assignment in car networks, the models that do take account of changes in travel impedance of a link due to traffic congestion are called equilibrium models.

Congestion in public transport networks can occur when the traveller demand exceeds the capacity of the system. Besides the line frequency, capacity is particularly linked to vehicle capacity. The algorithms for the equilibrium assignment in car networks could, in principle, be adapted for public transport networks. This, however, is hardly ever done. Since real congestion of public transport networks rarely occurs, equilibrium assignments are hardly ever used.

8.2.5 Closing remarks

Some aspects of the route determination process have not been dealt with in this discussion.

- First, there is the influence of tariff structures. In choosing his route, the traveller does not only look at the duration of the journey, but also the costs. Some transport operators apply tariff structures that are not directly in line with the transportation effort. There may not be any relationship between the tariff and the distance travelled, think of season tickets for the entire network, for example. Such situations complicate the application of conventional assignment techniques. Assignment computer programmes that can take complicated tariff structures into account do, however, exist.
- Secondly, there is the interaction between some forms of public transport and road traffic. Busses, for example, that use the car network directly experience the influence of any congestion on that road network. Here too, methods have been developed that can incorporate this interaction.

List of figures and tables with references

The notes refer to the source of figure or tabel.

References

[1] Manheim, M.L. (1979) *Fundamentals of Transportation Systems Analysis*. The MIT Press, Cambridge, Mass

[2] Ortúzar, J. de D. and Willumsen, L.G. (1995) *Modelling Transport*. Second Edition, Wiley.

[3] Ben Akiva, M.E. and Lerman, S.R. (1985) *Discrete Choice Analysis: Theory and Applications to Travel Demand*. The MIT Press, Cambridge Mass.

[4] Caliper Corporation (1996) *Travel Demand Modelling with TransCAD 3.0*

[5] Sheffi, Y. (1985) *Urban Transportation Networks.* Prentice Hall, Englewood Cliffs, N.J.

[6] Garnier, R and Taylor, J. (1992) *Discrete Mathematics for new Technology*. Hilger Bristol

[7] Clercq, F. le (1972) *A public transport assignment method*. Traffic Engineering & Control, June 1972.

[8] Lamb, G.M. and Havers, G. (1970) *Introduction to transportation planning, treatment of networks*. Traffic Engineering & Control, 11 (10), February 1970.