Fabien LEURENT

INRETS
2, avenue Malleret-Joinville
94114 Arcueil, FRANCE
Email leurent@inrets.fr

Transport Demand

and

Choice Models

March 1999

Foreword

This document contains a course on transportation planning models, targeted at engineers and economics. It assumes some basic knowledge of mathematical formulae and statistics.

The overall objective is to provide a clear understanding of transport demand models, from generation models of zonal trip productions and attractions, to assignment models of route choice on a network, passing by trip distribution models of destination choice and also mode choice models. These models simulate the response of demand to various supply actions, of which some instances are: capacity investment, timetable scheduling, traffic scheme, adaptation of fare levels or fuel prices etc.

Our main point is that the microeconomic concepts of demand, supply, consumption and choice are essential to the understanding, which in turn is essential to the design, analysis and usage of transport demand models.

Thus the economic interpretation is continuously emphasized throughout the document. This is accomplished owing to theory, which is the basic way to benefit from scale economies in teaching and learning. The document should also be used in line with its companion course which is dedicated to practical exercises and classroom cases. It is hoped that both courses complement each other in a satisfactory way.

Both of them follow a gradual, bottom-up approach: first concepts for economic analysis of transport, then network models and discrete choice models, next generation and distribution models, lastly the general issues of model composition and selection.

It is my pleasure to thank the European Union in a warmly fashion, for giving me the opportunity and financial support not only to gather this document from previous material mostly in French, but also to share it with the course participants.

Contents

List of figures

List of tables

A Basic transport economics

The economic science pertains to the management of scarce resources. Microeconomic analysis focuses on the following conceptual model: an individual actor, called the decision-maker, is endowed with initial resources and faced to external resources and constraints (eg. prices). The decision-maker can trade initial resources in place of external resources, which results in final individual resources. The decision-maker derives a profit from his final resources: he wants to maximize his profit and to do so he adapts his trade-off between initial and external resources. This process is called choice.

The model of microeconomic choice applies to a variety of areas. In transportation it applies to transport demand as well as to transport supply. We shall first identify the transport market. Next we shall describe in turn the transport demand and the transport supply. Then we shall address the issue of market equilibrium. Lastly we shall introduce the social evaluation of transport, including side effects as well as consumer surplus.

A1 The transport market

A1a Transport as a product

An economic product, or good, is sold on a market by suppliers to consumers. Each supplier, or producer, or firm, sells one or several units of the goods to consumers who also are its customers. Each customer pays his consumption to the supplier with respect to the quantity. The amount of the payment is often equal to the quantity times a unit price.

In transportation, a good is a trip, or carriage, of a given mobile entity from an origin place to a destination place, under some temporal and packaging conditions. The carriage may reduce to the supply of infrastructure or of a packaging. The temporal condition implies that transport cannot be stocked. The spatial condition implies that transport cannot be displaced.

A1b The diversity of transport

The general definition encompasses a variety of cases: passenger or freight, urban or interurban, low cost or high quality. The spatial and temporal circumstances also provide dimensions for analysis: it is often wise to identify several origin-destination pairs (an O-D pair is the couple of an origin zone and a destination zone) and several time periods (eg. heavily trafficked vs. low trafficked).

The analysis may focus on a given market segment, i.e. an intersection point for a combination of analysis criteria. More often it addresses a group of segments, each of which is considered homogeneous with respect to the criteria. In fact each segment may be further subdivided, eg. an origin or destination zone may be divided into sub-zones.

A2 Transport demand

A2a Transport as a derived demand with side features

From the consumer viewpoint, transport serves a primary purpose. In the case of passenger transport, the primary purpose is to perform an activity in the destination place. In the case of

freight transport the primary purpose is to contribute to a process. The basic value of transport pertains to the activity or process, not to the move in itself. Thus the demand for transport is derived from the demand for activities and processes.

However the features of the move influence the value of transport to the consumer: first the price may be deduced from the gross value of the activity, which results in a net activity value. Next there are *side features* such as trip duration (travel time) and comfort which influence the net value of transport to the consumer. The set of side features is known as the quality of service, or the level of service.

A2b On consumers

Transport consumers, or users, may be defined as the individual decision-makers who make the final decision to move. This includes the travellers with private trip purpose, the shippers of freight or of passenger transport with professional purpose. This excludes the intermediate consumers such as fleet operators which consume infrastructure capacity.

In most transport problems, demand is made up of many individual consumers, each of whom has negligible influence: this is referred to as demand *atomicity*, as opposed to supply monopoly or oligopoly.

A2c Demand choice

The decisions to be made by transport consumers include: (i) whether to consume or not, i.e. whether to produce a trip or not; (ii) choice of one transport variant from among several ones, eg. Route choice on a network, mode choice, period choice; (iii) choice of one activity variant from among several ones, eg. Destination choice.

In every case it is assumed that the individual decision-maker makes the choice by himself: this is a user choice, as opposed to a system choice in which a system regulator makes the decisions and assigns each individual consumer to a given variant (or to a given combination of the variants).

A2d Optimal choice and information

The basic principle of economic behaviour is the axiom of *optimal choice*: i.e. that any decision-maker chooses his preferred variant. When the preferences can be represented by a profit function, the decision-maker strives to maximize his own profit, or equivalently to minimize his own cost.

For a transport consumer, the objective is also to maximize the net value of the activity, which is equal to the difference of the gross activity value and the transport cost. The transport cost includes not only the transport price but also the cost of time and comfort. The resulting, aggregate cost is known as the *generalized cost* of transport to the consumer.

The principle of rational choice may be refined in a number of ways, many of which relate to the information available to the consumer. We can assume that each consumer is faced with a specific, restricted choice set, i.e. a subset of variants which he is acquainted with. We can also assume that each consumer is aware of the variant features up to only a given level of accuracy: eg. That he knows the mean travel time up to a given bias, or that he considers both the mean travel time and the variance of travel time.

A3 Transport supply

A3a Systems analysis of transport supply

The production of transport usually involves three components: (i) infrastructure (eg. Footway, river, roadway, railway, station, airport); (ii) mobile entities (eg. Footman, vehicles); (iii) operating rules also known as protocol (eg. Itineraries, driving of public vehicles, timetables, frequencies). Each component may be operated by one or several economic decision-makers, all of which are transport suppliers as opposed to the final consumers.

In the absence of further notice, we consider transport supply in an integrated way. Then a transport firm is an individual decision-maker of relatively large size as compared to the atomic consumers.

A3b Decision variables and supply choice

A transport supplier exerts a control on the service: this control can be described by decision variables. Long-run decision variables pertain to infrastructure or fleet planning: level of capacity, design speed, level of comfort etc. Short-run decision variables can be modified in a relatively short time: level of price (fare level, toll level), design and timetable of public lines, operating speed etc.

The choices of a transport supplier therefore consist in fixing the decision variables: this process is called production planning. As the choice set is usually very large, it is most often reduced to a small number of specimen, called scenarios. The regulator of the transport system may impose service constraints, eventually associated with specific subsidies.

Each supply scenario has a given value for the supplier: the net profit is the difference between revenues (commercial revenues as well as subsidies) and production costs. The basic objective of the supplier is to maximize his profit over the set of scenarios.

A3c Congestion

In transport the level of service may depend on the number of consumers, with little or no control by the supplier. This phenomenon is known as congestion: it may affect the travel time (roadway travel times are increasing functions of mean flows) and also the level of comfort (trip-making under congested traffic is less comfortable than under light traffic, eg. there may be no available seat in a public transport vehicle).

Transport suppliers are not affected by congestion inasmuch as it does not reduce the flow. However there are oversaturated situations in which flow is reduced to a small part of the limit capacity, which results in lost capacity.

A4 Market equilibrium

A4a Tâtonnement on an economic market

An economic market is an abstract meeting place for suppliers and consumers. Suppliers put forward supplied quantities of a good at supply prices, whereas consumers ask for demanded quantities at demand prices. Their meeting induces a set of agreements and of mutual adjustments: this process is known as a tâtonnement. Under abstract assumptions, a tâtonnement process ends in an equilibrium state, at which point the supply price is equal to

the demand price and the sold quantities is equal to the bought quantity (thus it is referred to as the level of consumption).

The adjustment of consumers to the supply price is called the demand function. The adjustment of a supplier to the demand price is called the supply function of the supplier.

A4b Demand function and consumer surplus

Let us differentiate the potential consumers with respect to the individual willingness-to-pay for the product. When only the generalized cost G is likely to change, the number q of serviced consumers depends on it on the basis of the demand function D: $q = D(G)$.

The demand function is related to the cumulative distribution function of the willingness-to-pay, W, on the basis of $D(G) = Q(1 - W(G))$ in which Q is the total number of potential consumers.

The demand function is tightly linked to the concept of consumer surplus. As the willingness-to-pay for a good is akin to a gross profit derived from the consumption by the consumer, the difference between it and the generalized cost is akin to a net benefit, called surplus. The consumer surplus CS related to a demand function is the aggregate of the individual surpluses:

$$CS = \int_G^\infty (g - G)\,\mathrm{d}W(g) = \left(\int_0^{D(G)} D^{-1}(x)\,\mathrm{d}x\right) - D(G).G\,.$$

Consumer surplus then measures the total net profit of the demand.

A4c Supply function

The cost of a production plan is a function $c(y)$ which gives the expense associated with a level of ouput y. This corresponds to the total production of a good by a given supplier. As a result of the firm's behaviour, the cost of producing the level of output y is minimized over the set of decision variables, say x, which can yield that level of output.

Each component x_n of x is a factor demand function $x_n(w, y)$ of the output y and the factor prices w of x. Then $c(y) = \sum_n w_n x_n(w, y)$ is the long-run total cost. The long-run average cost is $LAC = c(y)/y$, while the long-run marginal cost is $MAC = \partial c(y)/\partial y$.

In a competitive market, the firm adjusts its production plan so that the marginal production cost be equal to the output price p_y: thus in the long-run the level of output y is obtained by $\partial c(y)/\partial y = p_y$, which can be inverted into a function $y = S(p_y)$ called the supply function of the firm. Usual supply functions are increasing with respect to the output price.

A5 Cost-benefit analysis

A5a Side vs. primary impacts

The primary impact of transport is to allow users to benefit from several activities located in different places. However there are also side impacts that may affect the transport user (the side features) or other groups of impactees.

Let us mention the following side impacts: (i) congestion, (ii) accidents, (iii) chemical pollution of air and soil, (iv) noise, (v) space occupancy.

Associated with every side impact are several concepts of cost to compensate its effects on the impactees: compensation cost (equal utility by financial transfer), avoidance cost (cost of not

producing the impact) or deterrence cost (cost of a « barrier » between the impact and the impactee).

A5b Social evaluation

From the social viewpoint, each transport plan has a global cost that includes the production cost and the cost of side impacts. It also has a global benefit that includes consumers' surplus, suppliers' surplus, land-owners' surplus.

The purpose of cost-benefit analysis is to evaluate the net social surplus of a transport plan, defined as global benefit minus global cost. Then the social criterion for choice among several transport plans is maximum net social profit, or more often maximum return-on-investment for social funds.

A5c System optimum via control

The system optimum is a state of the transport system that maximizes the net social surplus. It generally differs from the user optimum since users have a selfish behaviour and transport exerts side impacts. However there are decision variables which could enable a system regulator to achieve a system optimum by controlling the user optimum: eg. setting link tolls on a network to make the users' generalized costs correspond to marginal social cost rather than to marginal selfish cost.

A6 Conclusion

A number of economic concepts were introduced which enable one to analyze transport in an economic framework. All of them are important and deserve emphasis. Demand choice and market equilibrium deserve a special mention.

B Models

What is a model? Does an economic model differ from a statistical model? Which confidence should be granted to a model? Is it possible to transfer a model from one case to another?

All of the previous questions are frequent and important ones. They deserve rigorous answers, i.e. answers formulated in a theoretical framework with clear concepts and assumptions. In other words, a model of models is required.

This lesson provides a theoretical framework to analyze a scientific model. There are two main approaches to a model, external versus internal. Section 1 introduces the external description of a model, whereas the internal description is addressed by Section 2. Lastly Section 3 provides external and internal descriptions of transport demand models.

B1 External description of a model

B1a Example

Let us consider the case of a mode choice model. It predicts the number of trips on each of a given set of transport modes (eg. car, public transport) on the basis of the overall number of trips, and of some attributes of the modes and the trip-makers.

As the modes are described in an abstract way by their attributes (eg. price, time, number of transfer points), we may simulate not only an *actual* state with actual attributes, but also *virtual* states with modified attributes. A model may predict the reaction of demand in response to modified fare levels, which is of paramount importance to a network operator who wants to maximize the revenue.

B1b General definition

A scientific model may be described as a machine that turns inputs into outputs: inputs are assumptions, whereas outputs are consequences. A model enables the analyst to deduce the consequences from a given set of assumptions in a systematic way.

Thus a model is a system for knowledge. Its basi aim is to simulate a real-world system by way of analogy: in the analogy between the model and the system under study, the assumptions and consequences pertain to the state of the system.

The external description of a model consists in:

(1) the set of inputs,

(2) the set of outputs,

(3) the function which transforms inputs into outputs.

This description is external since it focuses on inputs and outputs, rather than on the transformation function.

B1c Mathematical formula

We may also describe a model by way of the following mathematical formula:

$$Y = F(\Theta, X)$$

In which Y denotes the output variables, X and Θ denote the input variables and function F denotes the transformation function.

In a typical simulation, the input variables X and Θ are also called exogenous variables since their values are set outside the model (eg. fare levels in a mode choice model). The output variables Y are also called endogenous variables because their values are determined inside the model. As regards the input variables, we denote by Θ, and we call parameters, those with a value which remains constant over a wide range of cases, whereas the residual exogenous variables X may change from case to case. Lastly we have to differentiate control variables, which are inputs with value set by the analyst without constraint, from perturbation variables of which the value is constrained by the environment.

B1d Model estimation

Given X, Θ and F, the simulation viewpoint is to derive the value of Y.

The estimation viewpoint is as follows: given F and observations (X_i, Y_i) of distinct states *i* of the system, estimate the parameters Θ which make the model outputs $\hat{Y}_i = F(\Theta, X_i)$ match the observed Y_i in an optimal way.

In most models indeed, some of the exogenous variables cannot be observed or measured in a straightforward or cost-efficient way: thus the missing values have to be completed by estimation or inference, rather than by guess.

B2 Internal description of a model

B2a Definition

The internal description of a scientific model consists in a scientific interpretation of the transformation function: why and how do the inputs yield the outputs? This may be called the semantic content of the model. It is made up of structural assumptions which characterize the model and remain stable from case to case.

B2b Example

A mode choice model may possess the following internal description:

- The number of trips assigned to a given mode results from the individual choices of trip-makers.
- Each trip-maker selects the mode which he prefers, on the basis of the price and time.
- The trip-makers' preferences are summarized by a generalized cost associated by the trip-maker to each mode.
- The generalized cost of a mode depends on both the modal attributes and the consumer's attributes (eg. individual value-of-time, availability of car).

B2c The four aspects of a model

As a schematic representation of elements and relationships, the model formalizes theoretical knowledge of the system. We shall use the term *semantic content* to describe this formalized knowledge: it includes the elements and the relationships between them. The elements are organized into subsystems. The relationships (eg causality, simultaneity of occurrence) are

explanatory mechanisms, or causal phenomena which fix the values of endogenous variables (explained by the model) on the basis of exogenous variables (fixed outside the model).

The formal content of a model consists in two parts: first a formal image of the semantic content, second the specific formal issues of characterization, existence, stability and uniqueness. The formal image of the semantic content can be obtained by specifying a mathematical notation for each element and then expressing each explanatory mechanism as an elementary relationship which links some elements (variables which are endogenous to the relationship) to others (variables which are exogenous to the relationship). Then the resulting mathematical relationships can be synthesized into a characteristic formula in the form of a standard mathematical program (eg. Convex optimation problem, fixed-point problem).

A simulation is required in order to solve the characteristic formula and this requires a process which is called a *solver*: this is the *technical content* of the model.

Lastly, empiric application requires data concerning the studied case: we shall use the term *empiric content* to describe the data and parameters and the specification of the mathematical functions involved in the characteristic formula.

B2d About model transferability

The net analogy between a system and a scientific model of it cannot be observed. Only a gross analogy between an observed model of the system and the scientific model can be analyzed. This gross analogy applies to the four aspects of the model.

Several philosophical viewpoints are possible on the respective importance of each aspect. From the empiric viewpoint, only the external description is important, hence the empiric aspect and the solver. From the scientific viewpoint the internal description is most important, hence the semantic content and the associated characteristic formula. Only the semantic content provides a rational foundation to transfer a model from one case to another.

B3 Overview of transport demand models

Most transport demand models can be assessed by reference to the so-called four-step model. This consists in a classical sequence of four models. The four modelling blocks are respectively: a generation model, a distribution model, a mode choice model and a network assignment model.

B3a External description

Let us first define a transport zone: a transport system pertains to a specific area which is a set of individual locations. A transport zone is a subset of individual locations: rather than considering every trip on an individual basis from their own origin place to their own destination place, trips are aggregated with respect to origin zones and destination zones, making up origin-destination (O-D) pairs.

A generation model outputs the total number of trips which are either produced by an origin zone (zonal trip production) or attracted by a destination zone (zonal trip attraction). Its inputs consist in zonal attributes (eg. number of inhabitants) and in transport network attributes (eg. accessibility).

A distribution model outputs the O-D trip rates. Its inputs consist in zonal and network attributes: this may include the zonal productions and attractions yielded by a generation model.

A mode choice model outputs the trip rates per mode and O-D pair. Its inputs include O-D trip rates and modal attributes.

A network assignment model outputs the trip rates per route, for a set of routes which all belong to a given modal network. Inputs include O-D modal trip rates and route attributes, mostly derived from link attributes. In fact the results are often aggregated by link, yielding link flows as output rather than route flows. These link flows represent the local traffic and they may be compared with the local capacity to assess the network performance.

In the four-step model, each of the first three models produces outputs which are inputs to the next step model. Network assignment is the conclusion since it supplies the analyst with the network traffic load.

B3b Internal description

The following lessons provide more detailed external and internal descriptions of transport demand models. A prominent internal feature is economic choice: network assignment is interpreted as route choice, mode choice as an economic choice, trip distribution as destination choice, and even generation may be interpreted as trip frequency choice.

Each choice is modelled by describing a supply side (eg. network routes in assignment) as opposed to a demand side (i.e. trip makers): demand preferences are the ultimate rationale for the choice from among the variants.

B4 Conclusion

A number of definitions pertaining to models have been introduced. The external description serves exogenous purposes and is associated with practical application. The internal description serves an endogenous purpose of knowledge and explanation of behaviour.

C Supply side of assignment

A network assignment model deals with the route choice of trips from origins to destinations through a transport network: the network may be a car network, a transit network, a railway network, an airline network, or a combination of these and others (i.e. a multimodal network). The supply side of assignment is comprised of the characterization of network routes (topology and economic attributes) and the congestion phenomena which relate the local travel time to the local traffic.

C1 Paths, routes and strategies

C1a Nodes and arcs

A network consits in a set of nodes and a set of arcs, each of which is a directed link from a tail node to a head node. The traversal of an arc is the local movement (or carriage) from the tail node to the head node.

In the transport of persons or goods, nodes may represent junction places, origin and destination places, loading or unloading points, or merely places at which the local attributes of a line change (eg. number of lanes).

An arc may represent a one-way road, a loading or unloading operation. Two-way roads correspond to two arcs since the local flow of the opposite directions do not compensate each other.

C1b Paths

A path is a positive sequence of directed arcs $(a_1,...,a_n)$, such that the head node of arc a_i coincides with the tail node of arc a_{i+1}. A path is elementary if it does not contain a given arc more than once. Thus an arc is an elementary path from its tail node to its head node.

The traversal of a path is the movement (or carriage) from the tail node of the first arc to the head node of the last arc.

On a transport network, some arcs may not be open to certain classes of trips: eg. hazardous materials. Other restrictions relate to turns: at a given junction it may not be possible to transfer from a given incoming arc to a given outcoming arc.

C1c Lines and routes

In car networks, topological restrictions on paths are the exception rather than the rule. This extends to all individual modes (eg. on foot, bicycle, moto) for which there is no transport vehicle or the vehicle is operated by the user.

In public transport however, most often the vehicles are controlled by an operator and assigned to a given service line: bus lines, railway lines, airlines and so on. These networks possess fewer arcs and fewer nodes (including alighting and transfer points); they usually operate owing to a complementary access network (on foot network in urban transport, car network in urban and interurban transport). The meeting of two lines in a given node may not correspond to an (almost) instantaneous contact: the services are discrete rather than continuous.

The schedule of services is also an important feature which may be described by a service frequency when services follow each other at an even pace.

Two distinct service lines may share more than one node, or even include a same subsequence of network nodes: then the lines operate in parallel between their common nodes. If the fares and quality of service (comfort, travel time) are similar, potential consumers perceive the parallel services in a unified manner and the two frequencies add up to a joint frequency (or combined frequency).

Thus a public route is best described by a sequence of transfer nodes, such that each pair of successive transfer nodes is associated with a set of parallel service lines.

C1d Strategies

The concept of a path (or a route) as a topological structure has a fixed, static significance: it does not vary with time. In a path choice model, the analyst has to make explicit whether the path is chosen prior to the trip, in a static way, or the path may be adapted during the trip, possibly on the basis of dynamic information.

The concept of a strategy extends the concept of a path (or a route) by adding a dynamic dimension to it. Precisely a strategy is defined as a bundle of paths, i.e. a subnetwork from an origin node to a destination node, along which there are choice nodes where dynamic information may determine the next arc (or service line).

In public transport, at a boarding node the dynamic information may consist in the lines which are serviced by the incoming vehicles: the trip-maker may choose the first incoming vehicle which belongs to a line that will take him closer to his destination.

C2 Path attributes

C2a Additive attributes

The overall length of a path k can be decomposed along the arcs a which it traverses: $L_k = \sum_{a \in k} L_a$, provided that the path is elementary.

This extends to the overall travel time of a path, if we include into it the transition times between successive arcs. These transitions are knows as turn penalties (eg. left turn penalties in car assignment).

As regards the fare of a path, it may be established on an overall basis (eg. a cordon fare scheme) or on an additive basis (eg. sum of tolls, sum of driving costs).

C2b Perceived vs. measured attributes

Qualitative attributes such as comfort or to a lesser extent time are perceived on a subjective basis rather than an objective one. For instance, trip-makers perceive a given unit of waiting (or walking) time approximately twice as much intensely as the same unit of in-vehicle time.

This may be modelled by splitting the travel time into several subterms such as in-vehicle time, waiting time, transfer time, walking time etc. and by weighting each subterm by a specific coefficient to obtain a perceived travel time.

A similar method is used to take into account some aspects of comfort: the availability of a seat to a passenger can be evaluated by splitting the in-vehicle or waiting time into time with an available seat and time without any available seat.

C2c Explicit variations

A given attribute may vary in an explicit way with respect to temporal conditions or the category of trip. An arc toll or travel time may depend on the weekday and time-of-day. It may also depend on the category of trip: first-rate vs. second rate in public transport, heavy vehicle vs. light vehicle in private transport.

C2d Other variations

Arc and path attributes are submitted to two types of variations, deterministic vs. random. Deterministic variations with respect to state variables such as flow rate are investigated in the next section. By definition random variations cannot be controlled, even if their effect may be attenuated. Prominent among them are the random variations in roadway travel time: even on a motorway they may represent up to 90% of the total variations which also include deterministic variations owing to the traffic level!

C3 The supply function

C3a Basic definition

In the neoclassical microeconomic model of supply, the supply function is the dependency of the level of output upon the selling price: at a given selling price the firm maximizes its profit by producing an adapted quantity of output. Conversely, the reciprocal supply function determines the selling price with respect to the produced quantity.

C3b Transport supplier behaviour

In a transport choice model, only the demand choices are modelled. The transport supplier choices are taken as exogenous: network structure, infrastructure and fleet capacities, frequencies, free-flow travel time, tolls and fares, are fixed by assumption.

The only endogenous phenomenon is that of congestion, i.e. the dependency of the level of service upon the traffic level. This is somewhat analogous to a reciprocal supply function, although it is not controlled by a supplier.

C3c On travel time functions

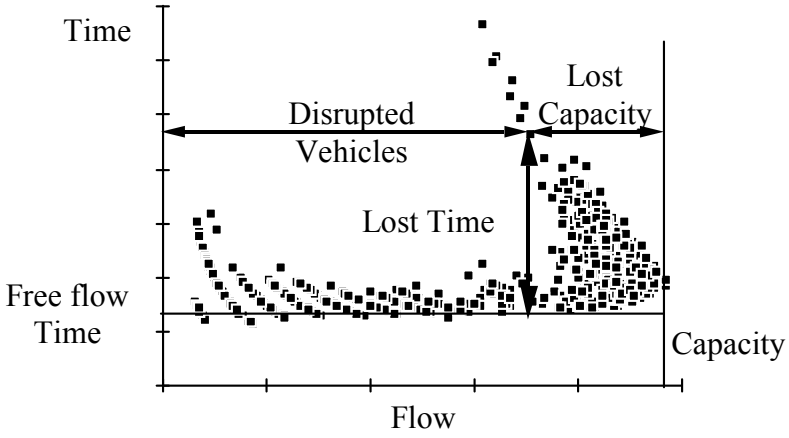
An arc travel time function relates the travel time experienced by mobile units throughout the arc to the level of traffic, which is measured by the flow rate. The time-flow diagram of a roadway arc may be analyzed in three parts. First there is an uncongested part, from null traffic to the vicinity of a limit value called capacity. The travel time keeps increasing from the free-flow travel time at null traffic to higher values. This corresponds to flows in which the mobile units do not influence each other too heavily.

Second a near capacity part, with flow in the vicinity of capacity: the travel time may vary abroad a wide range. This corresponds to flows in which the mobile units are quite close to each other but can still move on without stopping most of the time.

Third an oversaturated part, in which travel time is higher while flow is smaller. This corresponds to flows in which the mobile units are very close to each other and forced to stop during a large proportion of the time.

In a static transport model, only the uncongested part and eventually the near capacity part may be represented in an explicit way. Thus truly oversaturated flows cannot be addressed.

Fig. a. Time versus Flow diagram.

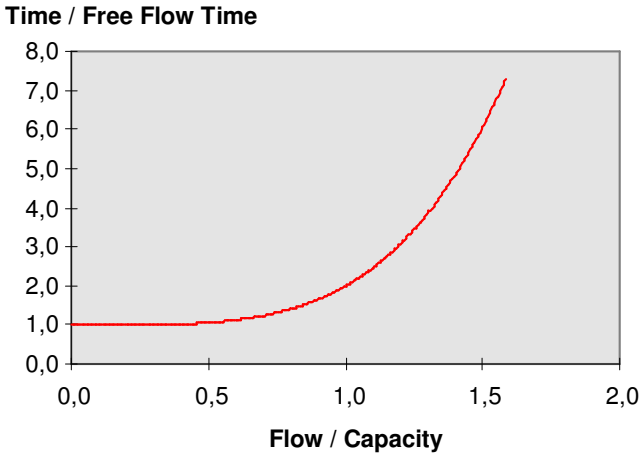


The usual traffic level variable is the vehicular flow, measured in passenger car units (pcu) in the case of roadway traffic: a heavy vehicle is 'converted' into several pcus. The rate of currency depends on the slope and the curving of the road: its value is approximately 2 for even, straight roads.

Lastly the travel time may also depend on the mobile unit which experiences it: the free-flow travel time of a truck is commonly higher than that of a car. However, in the near capacity and oversaturated part where overtaking becomes very difficult, the travel times of the different mobile classes should collapse into a single variable.

The function of time with respect to flow also involves a reference flow, called practical capacity, at which point the time increases more sharply with respect to flow. A standard formula is the BPR function: $t = t_0(1+\alpha(x/C)^{\beta})$ in which t is the travel time, t_0 the free-flow travel time, x the flow and C the practical capacity, while α and β are shape parameters.

Fig. b. A BPR Travel Time function.



C4 Conclusion

The supply side of a network assignment model is an exogenous representation of the transport services. It is not necessary to identify the network operators, which shows that the main analyst's interest does not lie in their behaviour. The only adaptive supply attribute is the travel time, which may be assumed to depend on the traffic level.

D Demand side of assignment

Most of the endogenous variables in an assignment model pertain to its demand side: the routes from origins to destinations with their respective flows for each class of mobile units, the arc and route level of service, and also the origin-destination trip rates when demand functions are considered.

Central to the subject is the economic behaviour of rational, optimal choice: each consumer strives to maximize his own profit, thus to use a shortest path (or route, or strategy). The consumer's preferences are modelled by way of a generalized cost function which integrates supply attributes.

Optimal choice and generalized cost are generic concepts that apply to a variety of demand segments. Apart from the segmentation with respect to origin-destination pair, dimensions of segmentation include the trade-off between quality of service and the price, most notably the value-of-time which determines the trade-off between time and price.

The microscopic behaviour of the consumers entail macroscopic effects such as congestion: the quality of service may depend on the flows, which in turn may depend on the quality of service through the demand function. There are several concepts of a supply-demand equilibrium: user equilibrium with no coordination between the consumers, vs. system equilibrium in which a coordinator fixes the routes to be taken by the consumers.

D1 Generalized cost

In a theoretical economic market, each good is characterized by its nature and its price: the nature may be defined simply as an index, or more explicitly by several attributes of quality (eg. the place and period of availability). Each consumer has his own preferences towards the goods, he has his own quality requirements which determine his willingness-to-pay for a given variant of a product.

The generalized cost of a variant to a consumer is an aggregate measure of the consumer's willingness-to-pay for the variant. In the case of route choice, the generalized cost includes the influence of supply price, travel time, comfort and so on. It is often modelled as a linear function of the path attributes: to consumer *c* the generalized cost of path *k* is

$$G_k(c) = \sum_n \alpha_n(c) X_n(k),$$

in which *n* is an attribute index, $X_n(k)$ is the value of the n-th attribute for path *k*, $\alpha_n(c)$ is the weight of the n-th attribute for consumer *c*. The weights $\alpha_n(c)$ describe the marginal trade-off of the consumer between the product attributes. More precisely, the ratio $\alpha_n(c)/\alpha_m(c)$ measures the rate of currency from attribute *n* to attribute *m*. The ratio $\alpha_T(c)/\alpha_P(c)$ of the time coefficient to the price coefficient is the value-of-time (VoT), i.e. the maximum amount of money which the consumer is willing to pay in order to save one unit of time.

D2 Optimal choice

Like the generalized cost, optimal choice is specific to a given decision-maker. In the absence of any further assumption, the decision-maker in transport demand choices is the individual consumer. The trip-maker makes the decision of whether to travel or not, and if positive of which variant to consume.

The basic economic behaviour is to select the available variant which maximizes the profit of the consumer, using a generalized concept of profit which corresponds to generalized cost, up to a minus sign. This generalized profit is also called utility, hence we may consider the utility function of a product (or a variant of a product).

Thus, on a transport network, a trip-maker strives to minimize his generalized cost over the set of available paths (or routes, or strategies) from his origin to his destination. The notion of shortest path depends on the definition of generalized cost.

The rationality assumption may be weakened by specifying the level of information available to the consumer. The available variants may be restricted to a subset of known variants. The path attributes may be known up to some degree of accuracy, or equivalently up to some degree of uncertainty. This is further explained in the context of discrete choice models.

D3 Demand classes

D3a Criteria for segmentation

It is often convenient to make the simplifying assumption of homogeneous demand, i.e. that every consumer behaves in the same way and perceives the products in the same way (same generalized cost function).

However there are models in which it is essential to differentiate several segments within the demand: the basic segmentation in transport choice models also pertains to origin-destination pairs.

Another obvious segmentation pertains to supply attributes: consumer classes which are submitted to different values should be differentiated, eg. cars and trucks because of roadway travel times. In the same way, if exogenous attributes vary across the modelled period, we have to divide it into homogeneous subperiods.

When demand choice involves trade-off between several variant attributes, eg. price and time, it is essential to differentiate demand segments according to the trade-off coefficients, namely the VoT in a price-time trade-off. This segmentation is easy to perform as it involves only the demand side and not the supply side. Specifically, in a price-time model which considers a variety of trade-offs between time and price, the segments of the VoT can be described in a statistical manner on the basis of the cumulative distribution function of the VoT. A distribution of the VoT is specified for a given O-D pair and a given mobile type; it can be reduced to the parameters of a standard distribution (eg. log-normal or gamma).

D3b The binary price-time model

Let us analyze the binary choice between two routes with a price-time model. The first route is a slow route with time T_1 and low price P_1, while the other one has reduced time $T_2 < T_1$ together with higher price $P_2 > P_1$. Let us denote by ΔT the difference in time $T_1 - T_2$ and by P the difference in price $P_2 - P_1$.

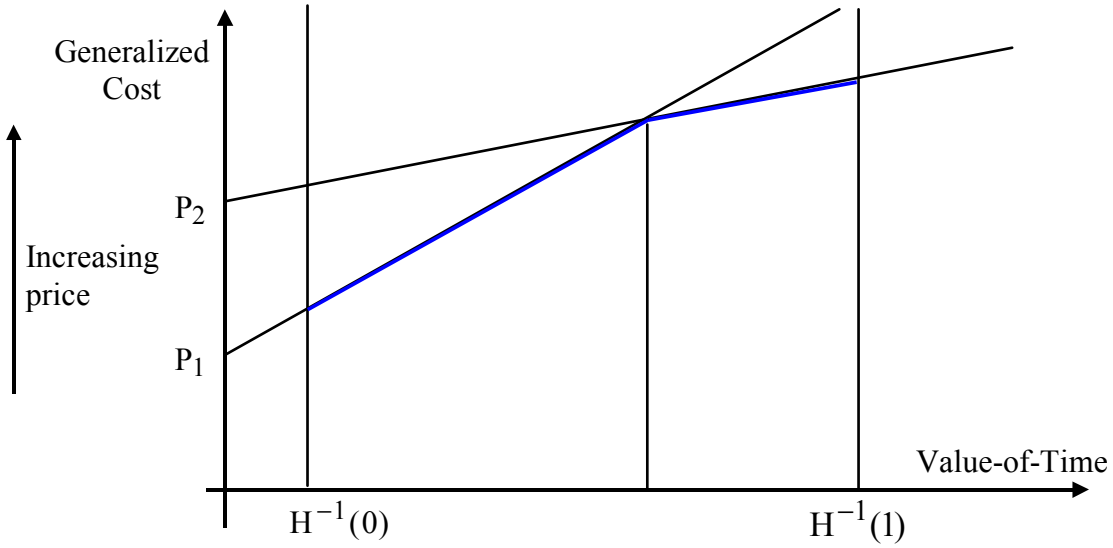
On the demand side, we assume that there are q consumers with distributed value-of-time. Let H be the cumulative distribution function of the VoT: for a given VoT v, there is a proportion $H(v) = \Pr(\text{VoT} \le v)$ of consumers with VoT lower than, or equal to, v.

To a consumer with VoT v, the generalized cost of route k is a function $G_k(v) = P_k + vT_k$. The consumer chooses the cheaper route, i.e. the first route if $G_1(v) \le G_2(v)$ or the second route if $G_1(v) \le G_2(v)$. Thus VoT v is assigned to the first route if and only if $v\Delta T - P \le 0$, or equivalently $v \le P/\Delta T$.

The cut-off value $v_{\text{off}} = P/\Delta T$, called the supply cut-off VoT, separates the lower VoTs which take the slow route from the higher VoTs which take the fast route.

As there are $qH(v_{\text{off}})$ consumers with VoT lower than v_{off}, these make the customers of the slow route, whereas the $q(1 - H(v_{\text{off}}))$ consumers with VoT higher than v_{off} make the customers of the fast route.

Fig. c. Generalized cost functions.



D3c Elementary example

Demand side. Let us assume a total volume q = 10,000 veh/day, values-of-time distributed uniform from A = 0 to B = 20 euros/hour. Thus the cumulative distribution function is $\text{H}(v) = (v - A)/(B - A)$ for $v \in [0, 20]$, $\text{H}(v) = 0$ if $v \le A$ and $\text{H}(v) = 1$ if $v \ge B$.

Supply side. We assume that the first route has travel time T_1 = 2 hours and price P_1 = 0, and that the second route has travel time T_2 = 1 hour and price P_2 = P. Then ΔT = 1 hour.

For a given toll fare of P ≥ 0, the cut-off value v_{off} = P/ΔT must be compared to $A = \text{H}^{-1}(0)$ and $B = \text{H}^{-1}(1)$:

- if $v_{\text{off}} \le A$ then all consumers take the second route.
- If $v_{\text{off}} \ge B$ then all consumers take the first route.

- If $v_{off} \in]A, B[$ then the proportion of lower VoTs is $H(v_{off}) = \frac{P/\Delta T - A}{B - A}$ which makes the market share of the first route, while the proportion of higher VoTs is $1\text{-}H(v_{off}) = \frac{B - P/\Delta T}{B - A}$ which makes the market share of the second route.

To conclude, the flow on the toll route is $f_2 = q(1 - H(v_{off})) = q\frac{B - P/\Delta T}{B - A}$ if $P/\Delta T \in]A, B[$, or q if $P/\Delta T \leq A$ or 0 if $P/\Delta T \geq B$ (figure d). The toll revenue is $qP\frac{B - P/\Delta T}{B - A}$ if $P/\Delta T \in]A, B[$, or 0 otherwise: this (part of) quadratic function attains its maximum value at point $q\frac{B - P/\Delta T}{B - A}$ if $P/\Delta T = (A+B)/2$ (figure e).

Fig. d. Toll route traffic.

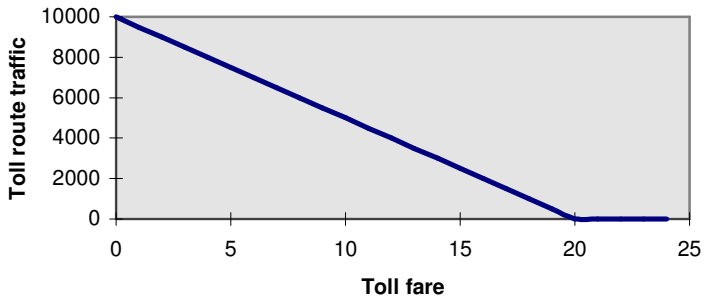
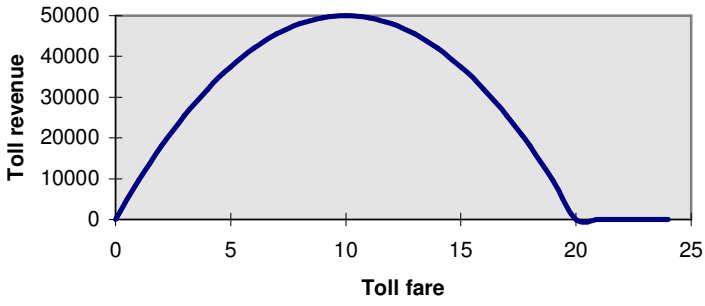


Fig. e. Toll revenue.



D4 Demand function

We can also differentiate the potential consumers with respect to the individual willingness-to-pay for the product. When only the generalized cost G is likely to change, the number q of serviced consumers depends on it on the basis of the demand function D: $q = D(G)$.

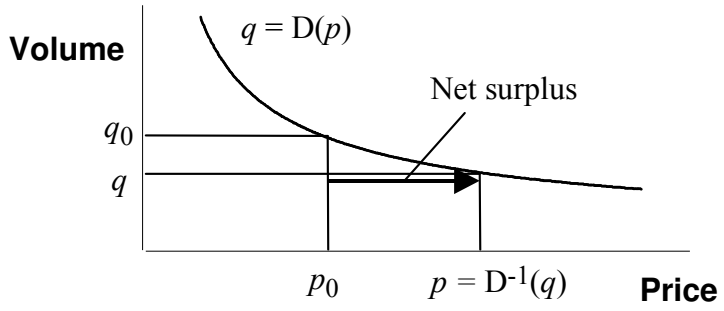
The demand function is related to the cumulative distribution function of the willingness-to-pay, W, on the basis of $D(G) = Q(1 - W(G))$ in which Q is the total number of potential consumers.

The demand function is tightly linked to the concept of consumer surplus. As the willingness-to-pay for a good is akin to a gross profit derived from the consumption by the consumer, the difference between it and the generalized cost is akin to a net benefit, called surplus. The consumer surplus CS related to a demand function is the aggregate of the individual surpluses:

$$CS = \int_G^{\infty}(g - G)\,\mathrm{d}W(g) \;= \left(\int_0^{D(G)} D^{-1}(x)\,\mathrm{d}x\right) - D(G).G\,.$$

Consumer surplus then measures the total net profit of the demand.

Fig. f. A demand function and net surplus.

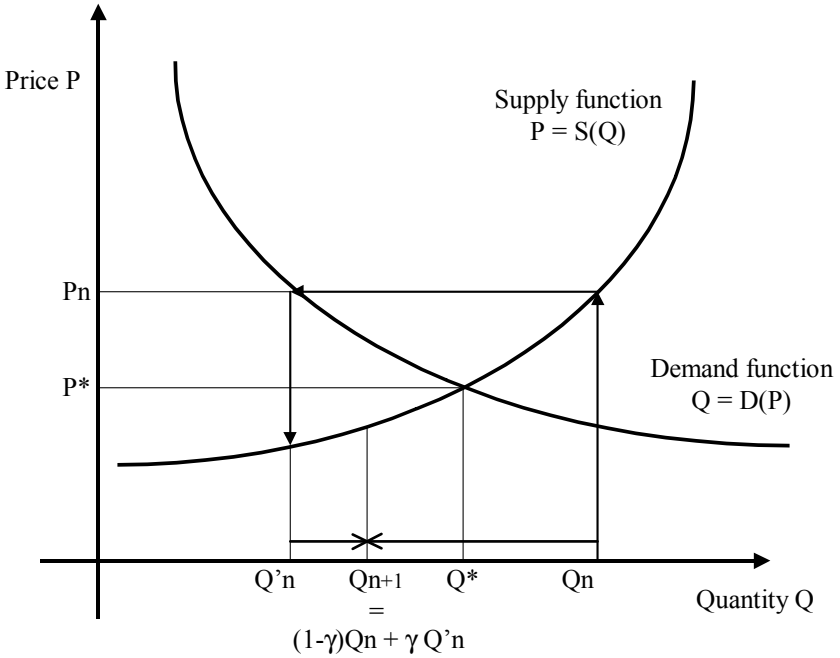


D5 Supply-demand equilibrium

D5a Principle

An economic market is defined as a meeting place for suppliers and consumers. Each decision-maker is driven by his own interests and strives to maximize his own profit. The supply function represents the response of a supplier to a demand price, and similarly the demand function represents the response of the consumers to a supply price. The market establishes a primal equality between the quantities (it equates the supplied quantity to the demanded quantity), and also a dual equality between the prices (it equates the supply price to the demand price). The overall result is known as the supply-demand equilibrium of the market. It can be illustrated as the intersection point of the supply function and the demand function in a quantity-price diagram.

Fig. g. Supply-demand equilibrium.



D5b User equilibrium

The supply-demand equilibrium induces a set of conditions on the variants of the product. The main condition is that, all other things being equal (i.e. under equivalent qualities), only the variants with minimal price are consumed.

This principle is known as the user optimum principle (or Wardrop's first principle) in network assignment theory. It applies to the global consumption of a homogeneous class of consumers, with specific origin, destination, mobile type and VoT. It states that the consumption of the class may only be assigned to variants with a minimum price, or generalized cost: hence at equilibrium all used variants have the same price, which is inferior or equal to that of any unused variant.

The user equilibrium can easily be illustrated in a binary choice with two independent variants and homogeneous consumers. Let us assume that there are two parallel routes k from one origin to one destination, and that the travel time T_k of route k depends on the flow f_k on the basis of a travel time function $T_k = \mathrm{t}_k(f_k)$. The equilibrium state depends on the total flow $q = f_1 + f_2$ in the following way.

There may be some cases in which only one of the two routes is competitive: if $\mathrm{t}_1(0) = \mathrm{t}_2(0)$ both routes are competitive for every value of q but this is the exception rather than the rule. Let us also assume than $\mathrm{t}_1(0) < \mathrm{t}_2(0)$, and let us test whether q may be assigned uniquely to the first road. Then it must hold that $\mathrm{t}_1(q) \le \mathrm{t}_2(0)$. If on the contrary $\mathrm{t}_1(q) > \mathrm{t}_2(0)$, i.e. if $q > \mathrm{t}_1^{-1}(\mathrm{t}_2(0)) \equiv q^*$, then the second route is competitive.

The first conclusion is also that if $q \le \mathrm{t}_1^{-1}(\mathrm{t}_2(0)) \equiv q^*$ then only the first route is loaded and thus $(f_1, f_2) = (q,0)$ and $(T_1, T_2) = (\mathrm{t}_1(q), \mathrm{t}_2(0))$.

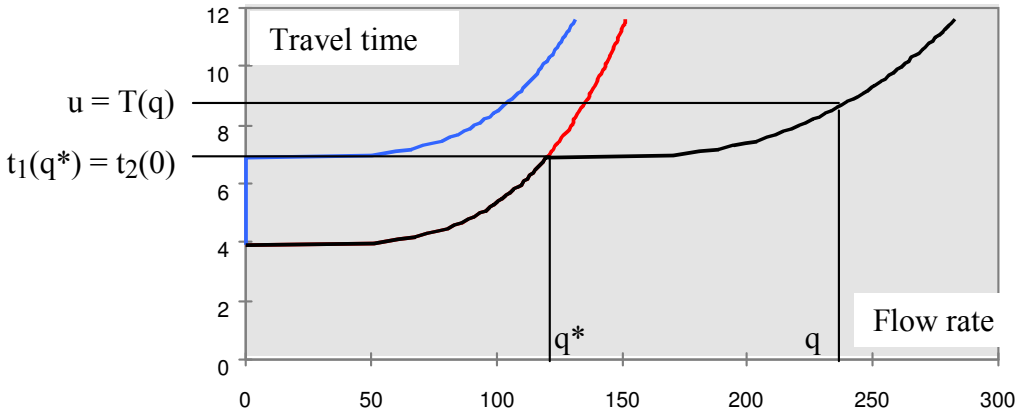
When $q > \mathrm{t}_1^{-1}(\mathrm{t}_2(0)) \equiv q^*$, the total flow is divided in two parts on the basis of

$$\begin{cases} \mathrm{t}_1(f_1) = \mathrm{t}_2(f_2) \\ f_1 + f_2 = q \end{cases}.$$

Let us define an auxiliary variable $u = \mathrm{t}_1(f_1) = \mathrm{t}_2(f_2)$. Then $f_1 = \mathrm{t}_1^{-1}(u)$ and $f_2 = \mathrm{t}_2^{-1}(u)$, hence u is characterized as the solution of the following equation $\mathrm{t}_1^{-1}(u) + \mathrm{t}_2^{-1}(u) - q = 0$. Defining T^{-1} as the function $u \mapsto \mathrm{T}^{-1}(u) \equiv \mathrm{t}_1^{-1}(u) + \mathrm{t}_2^{-1}(u)$ and T as the inverse function of T^{-1}, we obtain that the equation is solved by $u^* = \mathrm{T}(q)$, from which the equilibrium flows can be derived.

Let us illustrate the functions t_1 and t_2 on a Time versus Flow diagram. The curves of t_1^{-1} and t_2^{-1} add up to yield the curve of T^{-1} for $u \ge \mathrm{t}_2(0)$, which is also the curve of function T for $q \ge \mathrm{t}_1^{-1}(\mathrm{t}_2(0)) = q^*$. For values of q lower than q^*, function T is same as function t_1.

Fig. h. Competition between two parallel routes.



The user equilibrium principle can be extended in a number of directions, most notably by partially relaxing the homogeneity assumption (eg. price-time equilibrium, stochastic equilibrium).

D5c System equilibrium

A related, though largely different, concept is that of system equilibrium, in which a coordinator (or system regulator) fixes the productions and consumptions in order to attain a social objective (presumedly settled by a political authority).

A simple form of the social objective is to minimize the total travel time on a transport network. This system equilibrium principle is also known as Wardrop's second principle. A more realistic objective is to minimize a total social cost related to travel on the network, including energy and pollution costs and also consumer surplus.

The main interest of this concept is to serve as a reference point for a user equilibrium model: the social cost of the user equilibrium may be assessed with respect to the minimum social cost.

D6 Conclusion

Assignment models enjoy a deep economic outreach as they include the economic behaviours of optimal choice between variants, demand function and supply-demand equilibrium. This economic content is essential and should prevail upon operations research features (eg. shortest path algorithm, solution of equilibrium).

E Discrete choice models

A network assignment model is a discrete choice model because it represents the choice of entities (the trip-makers) among a finite number of variants (the routes), which are mutually exclusive forms of a product. Many other transport choice models are discrete choice models, from distribution models of destination choice to mode choice models.

The two lessons on assignment followed an economic guideline: we took as granted the demand functions, the travel time functions, the trade-off coefficients etc. in order to focus on the economic interpretation of the choice. An alternative, complementary approach is of a statistical nature: given observed variant market shares and observed supply and demand attributes, how can some unknown parameters be estimated? Such an estimation parameter could be the average value-of-time.

This lesson makes a firm distinction between the two approaches, economic and statistical. Section 1 contains background material on discrete choice models, from definitions to economic or statistical interpretation. In Section 2, we introduce functional relationships between discrete endogenous variables and exogenous variables, either discrete or continuous: this includes the widely-used logit and probit formulae. Section 3 provides guidelines to analyze the economic content of discrete choice models. Section 4 addresses the statistical estimation of a discrete choice formula: it describes the maximum likelihood estimator, its main statistical properties and derived tests. Lastly, Section 5 summarizes a case study of discrete choice models, related to the Prado-Carénage toll, underground, car-only roadway tunnel in the French city of Marseilles.

E1 General background

E1a Terminology

In a discrete choice model, the significance of the words ‘discrete’ and ‘choice’ is as follows. The adjective ‘discrete’ means that there is a finite (or denumerable) number of positions, all of which are mutually exclusive. In route choice, a position is a path (or route, or strategy), whereas in mode choice a position is a transport mode.

In economic discrete choice models, the various positions may be the variants of a given economic good (eg. Travel from an origin to a destination in an assignment model).

The noun ‘choice’ in itself means a decision process conducted by a decision-maker. In the context of general discrete choice models however, it just refers to a process of assigning a number of entities to a set of positions. The assignment may correspond to an economic process; it may alternalively correspond to a stochastic process.

To sum up, a general discrete choice model is a quantitative formula to assign entities to positions, also called variants. The entities can be distinguished from each other up to a limited degree of accuracy: the resulting underdetermination provides a probabilistic foundation to the formula. Of prominent interest are the position market shares, formulated as functions of the attributes of the positions and of the entities.

E1b Economic perspective

A discrete choice model has an economic content if it is possible to provide an economic interpretation to the entities, the positions and the assignment process. The following table summarizes the possible interpretation.

General discrete choice	Economic discrete choice	Assignment model
Entity	Decision-maker	Trip-maker
Position	Variant	Route (or path etc)
Assignment process	Economic choice	Path choice

The degree of accuracy up to which entities can be distinguished corresponds to the explicit assumptions about the demand, especially the segmentation into demand classes.

Let us denote by k a position, c a class of entity, X a vector of attributes (of classes or of position) and Θ a vector of parameters. The market share formula is as follows

$$p_{k/c} = F_{k,c}(\Theta, X)$$

in which $F_{k,c}$ is a given mathematical function.

In the economic perspective, from assumptions about X, Θ and $F_{k,c}$ we derive the market share $p_{k/c}$.

E1c Statistical perspective

In the statistical perspective, the question of economic significance is not relevant. Whether the result of a choice or of a random process, the market share of a position is considered as an endogenous variable, to be analyzed with respect to exogenous variables such as the attributes of entities or of variants (the X vector).

Given observations of $p_{k/c}$ and X and under prior assumtions on $F_{k,c}$, the aim is to estimate the unknown values of the parameter vector Θ, and also to predict the accurracy of this estimation.

E2 Market share formulae

E2a Generic formulae

The discrete choice of a consumer c among variants k involves binary endogenous variables $[y_{ck}]_k$ such that $y_{ck} = 1$ if consumer c chooses variant k or 0 otherwise. Thus $\sum_k y_{ck} = 1$. If we consider a class of N consumers $c(i)$, $i \in \{1,2..N\}$, there are N vectors of endogenous variables $[y_{c(i),k}]_k$. The market share of variant k is also

$$p_{k/c} = \frac{\sum_{i=1}^{N} y_{c(i)k}}{N}.$$

As $p_{k/c}$ is a continuous variable, it may be analyzed more easily than the individual binary variables y_{ck}.

Formally a discrete choice model is defined as a functional relationship between $p_{k/c}$ and exogenous variables X which include class attributes and variant attributes:

$$p_{k/c} = F_{k,c}(\Theta, X), \tag{1}$$

in which Θ is a vector of parameters and $F_{k,c}$ is a given mathematical function submitted to the normalization constraint $\sum_k F_{k,c}(\Theta, X) = 1$.

Formula (1) indeed is an aggregation formula: the aggregation is conducted on a class of consumers, or more precisely on a class of consumption units.

There may be a second aggregation step when several consumer classes are considered simultaneously: this is stated as

$$p_{k/C} = \sum_{c \in C} p_{k/c} \frac{N_c}{N}, \tag{2}$$

in which C is the set of consumer classes, N_c is the number of consumers in class c and $N = \sum_{c \in C} N_c$ is the total number of consumers.

E2b Discrete choice model under quotient form

In a discrete choice model under quotient form, there exist functions $G_{k,c}$ such that

$$p_{k/c} = F_{k,c}(\Theta, X) = \frac{G_{k,c}(\Theta, X)}{\sum_{k'} G_{k',c}(\Theta, X)}. \tag{3}$$

Thus the normalization condition $\sum_k F_{k,c}(\Theta, X) = 1$ is automatically satisfied.

A well-known case is the *linear logit* model, in which $G_{k,c}(\Theta, X) = \exp[\sum_n \theta_n X_{kn}]$ where X_{kn} is the n-th type attribute of the k-th variant and θ_n is the n-th component of the parameter vector Θ. The linear logit model is especially convenient for maximum likelihood estimation.

E2c Discrete choice model under distribution form

In a binary choice model under distribution form, there is a cumulative distribution function H_c and a real-valued function G_c such that

$$F_{1,c}(\Theta, X) = H_c(G_c(\Theta, X)) \text{ and } F_{2,c}(\Theta, X) = 1 - H_c(G_c(\Theta, X)). \tag{4}$$

Thus the normalization condition $F_{1,c}(\Theta, X) + F_{2,c}(\Theta, X) = 1$ is automatically satisfied.

A well-known case is the *linear probit* model, in which H_c is the cumulative distribution function of a reduced gaussian distribution (with null mean and unit variance) and $G_c(\Theta, X) = \sum_n \theta_n X_n$ is a linear combination of the parameters θ_n and the attributes X_n.

The distribution form also includes the binary price-time model, in which H_c is the CDF of the value-of-time in class c and G_c is the supply cut-off VoT.

E2d On exogenous variables

The value of a continuous real variable indicates a position on the real axis. The distance between two positions has a physical or economic significance, let us say a geometric significance.

In the case of a discrete variable or qualitative nature, the relative positions of the modalities have no geometric significance. However a discrete variable D_n may be included in a quantitative function $F_{k,c}$ by way of 'dummy variables' $z_{n,i}$ associated with each modality *i* of D_n by assuming that $z_{n,i}=1$ is modality *i* is satisfied or $z_{n,i}=0$ otherwise. Thus it always holds that $\sum_{i\in I_n} z_{n,i}=1$. Then real coefficients $\theta_{n,i}$ can be associated to the dummy variables $z_{n,i}$ and the product $\theta_{n,i}z_{n,i}$ may contribute to a function $F_{k,c}$.

E3 Economic content

Let us recall that a discrete choice model has an economic content if it is possible to provide an economic interpretation to the entities, the positions and the assignment process.

E3a Deterministic choice models

The binary price-time model viewed as a distribution discrete choice model is a simple deterministic choice model: economic choices underly the variant market shares. It is deterministic because for any given VoT, the associated choice has a unique result (almost surely).

Thus the continuous distribution of VoT across a population of consumers provides a simple DC model. Other simple DC models relate to the distribution of a discrete character such as O-D pair or trip purpose, which yield a DC model under quotient form.

E3b Random utility theory

In the theory of random consumer utility, it is assumed that consumers choose from among variants in a joint economic and random way. Precisely, eacn consumer *c* associates to each variant *k* a random utility function $U_{k,c}$ and he chooses variant *k* on every occasion such that $U_{k,c}\geq U_{k',c}$ for every variant *k'*.

Thus the market share of variant *k* to consumer *c* is $\Pr(U_{k,c}\geq U_{k',c}\ \forall k')$, which is individual with respect to the consumer but aggregate with respect to the random occasions.

Function $U_{k,c}$ may include attributes of the variants and of the consumer. Most often the analyst prefers not to include attributes of other variants. A conventional, simple linear utility function is as follows

$$U_{k,c}=\sum_{n\in I}\theta_n X_n(k)+\sum_{m\in J}\theta_m X_m(c)+\varepsilon_{k,c}(\omega)$$

in which *I* and *J* are disjoint subsets of components of Θ, attributes $X_n(k)$ relate to variant *k*, attributes $X_m(c)$ relate to consumer *c*, ω denotes a probabilistic alea (random occasion) and $\varepsilon_{k,c}$ is a random variable.

This linear formula separates the respective influence of variant, consumer and randomness in an additive way. We can interpret its deterministic part $\overline{U}_{k,c}=\mathrm{E}_{\omega}[U_{k,c}]=$

$\sum_{n\in I}\theta_n X_n(k)+\sum_{m\in J}\theta_m X_m(c)+\bar{\varepsilon}_{k,c}$ as a generalized cost, up to a minus sign. Provided that time T_k and price P_k are included in the first part of $U_{k,c}$, the ratio θ_T/θ_P of their coefficients may be interpreted as the value-of-time of consumer *c*. However in most cases the estimation is based on a segmentation of all consumers into classes and the parameters θ_T and θ_P relate to either the whole set of consumers, or a given class. Then ratio θ_T/θ_P is above all an average (mean or median) value-of-time.

A random variable $\varepsilon_{k,c}$ is comprised of its mean $\bar{\varepsilon}_{k,c}$ and a random residual $\varepsilon'_{k,c}(\omega)=\varepsilon_{k,c}(\omega)-\bar{\varepsilon}_{k,c}$. The mean is called the modal constant of variant *k* and it is usually included in the deterministic part of the utility function by way of a dummy variable: it reflects the aggregate effect of all unobserved factors. The random residual can be interpreted as the analyst's uncertainty on explicit attributes *X*, or as the uncertainty in the consumer's evaluation of the variant. In the case of repeated choice by a given consumer, the latter part may be due to temporal variations (eg. of the variant attributes).

E3c Derivation of logit and probit model

The logit and probit models can be obtained as random utility models under specific assumtions.

In the linear logit case, a linear utility function is assumed in which the residual variables $\varepsilon'_{k,c}$ are independent and identically distributed as a Gumbel variable with variance $\pi^2/6$. Thus their common cumulative distribution function is $G(x)=\exp(-\exp(-x+\gamma))$ where $\gamma\cong 0.577$ is Euler's constant. Then, denoting by $\overline{U}_{k,c}$ the deterministic part of the utility function, it holds that

$$\Pr(U_{k,c}\geq U_{k',c}\ \forall k')=\frac{\exp(\overline{U}_{k,c})}{\sum_{k'}\exp(\overline{U}_{k',c})}.$$

In the linear probit case, a linear utility function is assumed in which the residual variables $\varepsilon'_{k,c}$ are multivariate normal with matrix of covariance M. This enables one to consider joint variations of the $\varepsilon'_{k,c}$. In the binary case with independent residuals $\varepsilon'_{1,c}$ and $\varepsilon'_{2,c}$ of which the variances are σ_1^2 and σ_2^2 respectively, the first variant is chosen with probability $\Pr(U_{1,c}\geq U_{2,c})=\Pr(V_1+\varepsilon'_1\geq V_2+\varepsilon'_2)=\Pr(\varepsilon'_2-\varepsilon'_1\leq V_1-V_2)=\Phi((V_1-V_2)/\sqrt{\sigma_1^2+\sigma_2^2})$ in which Φ is the CDF of a reduced gaussian random variable and $\sigma_1^2+\sigma_2^2$ is the variance of the gaussian variable $\varepsilon'_2-\varepsilon'_1$.

E3d Economic outreach

The economic outreach of a discrete choice model depends on the explicit and accurate description of economic choices. The description is as much explicit as there are explicit dimensions of analysis, eg. O-D pair, trip purpose, mobile type and VoT. Its accuracy depends on whether the dimensions of analysis have straightforward economic significance (eg. VoT more meaningful than trip purpose) and whether the economic parameters correspond to the preferences of an individual decision-maker (particular choice better than the mix of several choices).

E4 Statistical estimation

An estimation problem consists in giving an estimated value to unknown parameters of a probabilistic model, on the basis of joint observations of endogenous and exogenous variables.

E4a Samples and likelihood

The basic concept is that of a random sample: a random sample of one observation is the observation of the result Y_t of a random experiment *t*, such that each modality *y* of the result Y_t has a given probability $\Pr(y|t,\Theta)$ of outcoming.

The probability $\Pr(y|t,\Theta)$ can also be viewed as a function of the parameter Θ: then it is written $g_t(\Theta|y)$ and it is called the likelihood function of Θ given sample *t* and observation *y*.

A random sample of size *T* usually refers to the observation of the *T* results $[Y_t]_{t=1..T}$ of *T* independent random experiments. The probability to observe the joint modality $[y_t]_{t=1..T}$ is $\prod_{t=1..T}\Pr(y_t|t,\Theta)$ under the independence assumption, and it can still be viewed as a likelihood function of Θ given the sample and the observations:

$$L(\Theta|[y_t]_{t=1..T}) = \prod_{t=1}^{T} g_t(\Theta|t,y_t).$$

The set of observations often splits into subsets within which the random experiments are identically distributed. Let T_i be the size of the i-th subset S_i; then

$$L(\Theta|[y_t]_{t=1..T}) = \prod_i L_i(\Theta|S_i,[y_t]_{t\in S_i}) = \prod_i \prod_{t\in S_i} g_i(\Theta|t,y_t)$$

since the function g_t is identical for all the samples in the i-th subset.

E4b Maximum likelihood estimation

Maximum likelihood (ML) is one of the two standard statistical estimation methods, the other one being least squares. ML is particularly well suited to analyze discrete dependent variables.

The principle of ML estimation is to select the value of the parameter Θ which maximizes the likelihood of a set of observations under general constraints such as the normalization condition $\sum_k F_{k,c}(\Theta,X)=1$. The intuitive interpretation is that this value Θ^{ML} yields the highest probability of outcome for the observation. Note that the ML estimator is a function of the observation: thus it is also a random variable.

There is a number of algorithms to compute the ML estimator. Most are based on the necessary first-order condition of maximization applied to the log-likelihood function which is the logarithm of the likelihood function, $\Lambda = \ln L$. In the absence of constraints, this is expressed as

$$\frac{\partial \Lambda}{\partial \theta_n} = 0.$$

E4c Properties of ML estimator

The ML estimator has desirable statistical properties: asymptotically (for very large samples) it is convergent, i.e. it concentrates on the true value. Thus in the limit it has no bias and its

deviation tends to zero. The accurracy of the estimation can be estimated by way of the matrix of second derivatives of the log-likelihood function, $H = [\partial^2 \Lambda / \partial\theta_n \partial\theta_m]_{n,m}$ evaluated at point Θ^{ML}, on the basis of $[\mathrm{cov}(\theta_n^{ML}, \theta_m^{ML}]_{n,m} \cong -H^{-1}$.

E4d Tests

We can avail ourselves of the estimated covariance matrix to perform tests of significance on the estimated parameters. We may test whether the n-th component θ_n^{ML} is different from 0 at the 1-α confidence level by constructing the ratio $t = \theta_n^{ML} / \sqrt{\mathrm{var}(\theta_n^{ML})}$ and computing the probability that a reduced gaussian variable has larger absolute value than *t*.

Another application of ML is to evaluate the likelihood of a restricted model, as compared to a reference model. Let us assume that in the restricted model the vector Θ' of *N'* parameters is a subvector of Θ and corresponds to a value of Θ with the *N-N'* components out of Θ' set at zero value.

Then the log of the squared ML ratio, $2[\ln \Lambda(\Theta^{ML}) - \ln \Lambda(\Theta'_{ML})]$, is distributed as a khi-square random variable with *N-N'* degrees of freedom. We can also evaluate the critical probability of the restriction hypothesis $\Theta = \Theta' \cup \{0\}_{N-N'}$. This is the probability that a khi-square variable with *N-N'* degrees of freedom has value larger than the observed ratio, say *x*: it can be written as $\Pr(\chi^2_{N-N'} \geq x)$.

E5 Case study of the Prado-Carénage tunnel

The Prado-Carénage tunnel in Marseilles was opened in 1993 and is the first experience in France of urban road tolling. This opportunity was used to implement a major survey which yielded a considerable amount of valuable information on the route choices made by drivers in Marseilles (1995).

This information was used to estimate the distribution of the value-of-time (VoT) of car trips in Marseilles: a binary choice model with a varying coefficient (the VoT) was formulated and estimated.

E5a The economic model and its interpretation

The behavioural model is based on the axiom of economic rationality: this states that the decision-maker chooses between the available options on the basis of his or her personal preferences. In the context of route choice on road network the decision-makers are drivers who decide which path to take: they are assumed to select the most advantageous path, i.e. the quickest or the least expensive path or, what is more likely, a path which represents a compromise between the shortest time and the lowest cost.

The hypothesis of transitive preferences states that the preferences of a decision-maker can be expressed by a utility function, i.e. an overall rating which the decision-maker assigns to each available option. This leads us to represent the utility of the tolled road for the driver *i* by $\mathrm{U}_{\mathrm{toll}}(i)$ and that of the toll-free road by $\mathrm{U}_{\mathrm{free}}(i)$. The hypothesis of rationality is expressed as follows: if $\mathrm{U}_{\mathrm{toll}}(i) \geq \mathrm{U}_{\mathrm{free}}(i)$ the driver will choose the tolled road but if $\mathrm{U}_{\mathrm{toll}}(i) < \mathrm{U}_{\mathrm{free}}(i)$ the driver will choose the toll-free road. Everything therefore depends on how the utility difference function $\Delta\mathrm{U}(i) = \mathrm{U}_{\mathrm{free}}(i) - \mathrm{U}_{\mathrm{toll}}(i)$ is specified.

We shall represent the difference in utility between the toll-free road and the tolled road by a model of the following form, in which v_i is the VoT for driver i (for this particular trip), P_{free} and P_{toll} are the prices of the two routes, T_{toll} and T_{free} are their travel times and $\Delta\varepsilon = \varepsilon_{free} - \varepsilon_{toll}$ is the difference between the random errors for each route:

$$\Delta U(i) = -(P_{free} - P_{toll}) - v_i(T_{free} - T_{toll}) + \Delta\varepsilon . \qquad (A)$$

Formula (A) takes account of the diverse trade-offs between price and journey time as the VoT v_i depends on the consumer. The random variate $\Delta\varepsilon$ encompasses the additional elements which are not taken into account in an explicit way.

E5b Survey results

Measurements of journey times and objective knowledge about paths. Travel times were surveyed on a set of road routes which pass through the tunnel or its main competing roads. Each was subjected to several types of periodic measurement (morning peak, evening peak or the rest of the day), so that the mean journey time on these routes be known to within two minutes at the 95% confidence level. Remark that the routes in the survey are the "main legs" of the true routes, and *they ignore the end legs of these*, which are specific to each trip. The distance and price of paths can also be found. The distance is measured in the field or on a map. The price is estimated from the toll if the route goes through the tunnel and by estimating the costs of car travel on the basis of the product of the distance and a ratio per unit distance. Note that in 1995 the toll fare was 11 F, for a tunnel 2.5 km long.

The origin-destination (O-D) surveys and the revelation of preferences. The O-D surveys intercepted the vehicles in the tunnel or on the competing roads and revealed not only certain characteristics of the drivers (age, sex) but also some of the circumstances surrounding the trip (purpose, location of departure and arrival and, of course, the time of passage through the survey station). The O-D surveys also *reveal the route choices* made by drivers because a driver who is intercepted at a station has chosen the route he or she is on in preference to alternative routes. This provides indirect information on the VoT distribution.

Additional information from a network assignment model. As data about the routes which are not chosen is also required, a network assignment model was used to compute the "shortest" alternative path (i.e. that with the shortest journey time) on the basis of the modelled journey times.

E5c Econometric processing

We shall replace (A) by the following:

$$\Delta U'(i) = P - v_i \Delta T' + \Delta\varepsilon' \qquad (B)$$

where $P = P_{toll} - P_{free}$, $\Delta T'$ denotes the difference between the estimated mean main leg journey times between the surface route and the competing route through the tunnel while $\Delta\varepsilon'$ is a new disruption which includes both the disruption $\Delta\varepsilon$ in the economic model and the difference $\Delta T - \Delta T'$ between the true time saving, ΔT, and what is substituted for it, $\Delta T'$.

Thus the disruption $\Delta\varepsilon'$ represents (i) the uncertainty in the economic behaviour model, (ii) the fluctuation affecting the sampling of observations, (iii) the uncertainty affecting main leg journey times and end leg journey times,, (iv) the uncertainty affecting computation of the alternative route, (v) the uncertainty affecting perception of times, (vi) the uncertainty affecting the aggregation of O-D pairs. The combined effect of these sources of variability and uncertainty is to interfere with the exploitation of observations with a view to finding

important unknown quantities such as the distribution parameters for VoT. The ultimate aim of the econometric model is to free us as far as possible from uncertainties because it frees us from the variabilities which we are able to make explicit. Here the purpose is to separate the variability which is associated with values-of-time from the residual uncertainty $\Delta\varepsilon'$.

Modal market share. Given P, ΔT' and a VoT of v, the individual toll route market share is

$$\Pr(\text{Toll} \mid v) = \Pr\{\Delta U'(v) \le 0 \mid v\} = \Pr\{\Delta\varepsilon' \le v\Delta T' - P \mid v\} = Z(v\Delta T' - P) \qquad \text{(C)}$$

In which we let Z denote the cumulative density function of Δε'. This is the *individual toll route market share*, specific to a given VoT v. Under the additional assumption that Z does not depend on v, we obtain the following *aggregate market share* of the toll route

$$\Pr(\text{toll}) = \Pr\left(\left\{\Delta U'(i) \le 0\right\}\right) = \int_v \Pr\left(\left\{\Delta U'(i) \le 0 \middle| v\right\}\right) d\Pr(v)$$

$$= \int_v \Pr\left(\left\{\Delta\varepsilon' \le v\Delta T' - P \middle| v\right\}\right) d\mathrm{H}(v) \qquad \text{(D)}$$

in which function H is the cumulative distribution function of the VoT, so $d\mathrm{H}(v) = d\Pr(v)$. By changing variables $\alpha = \mathrm{H}(v)$ and letting H^{-1} denote the reciprocal of H, this is transformed into

$$\Pr(\text{toll}) = \int_0^1 Z\left(\mathrm{H}^{-1}(\alpha)\Delta T' - P\right) d\alpha \,. \qquad \text{(E)}$$

Parametric estimation problem. Formula (C) gives the toll route market share as a function of the distributions H and Z. We shall formulate an estimation problem in which H and Z are the unknowns; more precisely, we assume that H and Z belong to certain functional classes within which they are identified by the numeric values of certain parameters. We assume that $Z(x) = Z_0((x - \Delta\overline{\varepsilon'}) / \sigma_{\Delta\varepsilon'})$ where the two parameters $\Delta\overline{\varepsilon'}$ and $\sigma_{\Delta\varepsilon'}$ are respectively the mean and standard deviation of $\Delta\varepsilon'$, and Z_0 is the (known) CDF of a reduced variate. For example, we can assume that Z_0 is the distribution function Φ of a reduced normal variate, which gives a probit model; if in addition we specify that H is a log-normal distribution we construct a probit model with a log-normal varying coefficient.

The issue of identifiability. As P does not vary in the observations, the parameters are related to each other. Letting μ and σ be the mean and standard deviation of the log VoT, only the scalar terms $(\Delta\overline{\varepsilon'} + P) / \sigma_{\Delta\varepsilon'}$, $\mu - \ln(\sigma_{\Delta\varepsilon'})$ and σ can be identified. Thus the mean VoT cannot be found, as we can only know $\mu - \ln(\sigma_{\Delta\varepsilon'})$. On the other hand σ can be identified. If we assume that the VoT are log-normally distributed, the ratio between the standard deviation and the mean depends only on σ on the basis of $(\exp(\sigma^2) - 1)^{1/2}$.

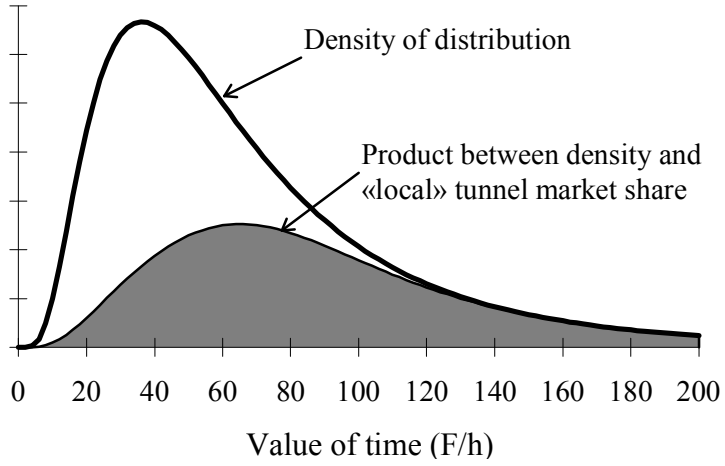
E5d Results

Estimated parameters. We estimated a logit model with log-normal VoT by the maximum likelihood method, yielding that $(\Delta\overline{\varepsilon'} + P) / \sigma_{\Delta\varepsilon'}$ = 1.535, $\mu - \ln(\sigma_{\Delta\varepsilon'})$ = -2.03, σ = 0.662. We applied the likelihood ratio test to the assumption «there is only one VoT», i.e. «σ = 0»: as compared to a logit model with a single VoT, there is a 12 point log-likelihood gain, which means that the probability of a unique VoT is less than 10^{-4}!

Inference of the mean VoT. Assuming that the mean random disruption is zero, we obtain that $\sigma_{\Delta\varepsilon'} = 11/1.535 \approx 7.17$ F, from which we can compute a median VoT of 56.5 F/h, a mean of 70.3 F/h and a standard deviation of 52.2 F/h. The uncertainty relating to the VoT parameters is of the order of 2 or 3 F/h.

The share of variability. We can compare the variability caused by the dispersion of VoT to the residual variability $\text{var}(\Delta\varepsilon')$. For a fixed time saving of ΔT' and no distance saving, the variance of ΔU' can be broken down into $\Delta T'^2 \text{var}(v) + \text{var}(\Delta\varepsilon')$. If we consider a median time saving ΔT' = 11 mn and a lognormal VoT, we obtain $\Delta T'^2 \text{var}(v) \approx 91 F^2$ in comparison to $\text{var}(\Delta\varepsilon') \approx 52 F^2$. Thus the variability due to the VoT is almost twice the residual variability, which means that it is thoroughly justified to make it explicit!

Fig. i. Modal share of the tunnel according to the value-of-time (for $\Delta T'$ = 11 mn).



Alternative specifications. Estimation of fixed coefficient stochastic models, distributed coefficient models with no residual random variate, logit models with a log-logistic or rectangular coefficient, confirms firstly the dispersion of values-of-time and the clear statistical superiority of the models which make it explicit, and secondly the importance of the role played by the residual random variable.

In a model with a varying, uniform coefficient and a reference CDF Z_0 which is integrable in closed form, we can give a closed-form formula for the modal market share, which facilitates the computation of the likelihood function. Let [A ;B] be the supporting interval of the uniform coefficient and $\tilde{Z}_0$ be a primitive function of the reference CDF Z_0. Then

$$\Pr(\text{toll}) = \frac{\sigma_{\Delta\varepsilon'}}{(B-A)\Delta T'}\left[\tilde{Z}_0\left(\frac{B\Delta T' - P - \Delta\overline{\varepsilon'}}{\sigma_{\Delta\varepsilon'}}\right) - \tilde{Z}_0\left(\frac{A\Delta T' - P - \Delta\overline{\varepsilon'}}{\sigma_{\Delta\varepsilon'}}\right)\right].$$

E5e Conclusion

The results we shall conserve are as follows: firstly, the dispersion of VoT for car journeys in Marseilles. The importance of residual causes other than the mean time saving or the distance saving should also be noted - these are responsible for one third of the variability of choices (through the residual random variate). We should also not forget the supplementary hypothesis which we have used to determine a mean VoT: the absence of a mean difference (other than that relating to time, distances or the toll) between the tunnel and competing

routes. To free ourselves from this hypothesis the problem would need to be examined with several levels of tolls.

E6 Conclusion

We must be careful to clearly distinguish two different probabilistic settings. The first one pertains to the statistical viewpoint and results from the random fluctuation in the observations. The second one pertains to the economic viewpoint and corresponds to aggregation assumptions. The two settings are linked together by a single fact of paramount importance: the variables of the economic setting can be considered as estimation parameters in the statistical setting. A basic form of this fact is that the variant market shares may be estimation parameters.

F Generation models

A trip generation model outputs the number of trips produced by origin zones or attracted by destination zones. The production of a given trip may be analyzed as the result of a choice, in a trip frequency choice problem. However most generation models follow another approach, which is more statistical than economic. In these models, the zonal productions and attractions are mathematical functions of exogenous variables. There are zonal models in which the exogenous variables are zonal attributes, as opposed to category models in which the exogenous variables pertain to segments of individual trip-makers.

We shall first describe zonal generation models. Next we shall outline the linear regression model which is useful to estimate a linear generation model. Then we shall introduce category generation models.

F1 Zonal models

F1a Principle

The principle of zonal generation models is made up of two parts, each of which makes sense in an intuitive manner:

1- the more activities in a transport zone, the more trips the zone produces or attracts.

2- The more people, or employment, or commercial area etc, in a transport zone, the more activities take place within it.

The resulting models output the number of produced trips, called production, of each origin zone, and also the number of attracted trips, called attraction, of each destination zone.

Let us denote by P_o the production of an origine zone o and by A_d the attraction of a destination zone d. A typical zonal generation model is formulated by the following equations:

$$P_o = F_{ori}(X_{on})$$

$$A_d = F_{des}(X'_{dm})$$

in which the input variables X_{on} are attributes of the origin zone, the X'_{dm} are attributes of the destination zone, F_{ori} and F_{des} are mathematical functions which may include parameters.

F1b Linear models

The basic usage of a generation model is to derive the zonal productions and attractions from given inputs X_{on}, X'_{dm}, F_{ori} and F_{des}. Linear relationships are of special interest for the sake of analytical simplicity. This assumption implies that

$$P_o = F_{ori}(X_{on}) = \sum_n \alpha_n X_{on}$$

$$A_d = F_{des}(X'_{dm}) = \sum_m \alpha'_m X'_{dm}$$

in which the variables α_n and α'_m are parameters.

F1c Design and estimation

The construction of a zonal generation model involves the following steps.

1- the identification of transport zones,

2- the selection of output variables, eg. productions and attractions for a given period and a given trip purpose,

3- the selection of the input variables X_{on} and X'_{dm},

4- the measurement or forecast of the input variables X_{on} and X'_{dm},

5- the estimation of functions F_{ori} and F_{des}, i.e. of the parameters α_n and α'_m in the linear case.

Steps 2 and 3 determine the economic content of the model, whereas Steps 4 and 5 determine the statistical significance.

F1d Case study: the Paris regional traffic model

Most transport planning studies in the Paris area are based on the traffic model MODUS of the DREIF (state agency responsible for regional transport). MODUS is a five-step model which includes generation, distribution, time-of-day split, modal choice and assignment (either car or public transport).

The generation component of MODUS is a zonal generation model. The study area is made up of 1,277 demand zones. Up to 62 O-D purposes are considered (table below); these are grouped into 8 classes for sake of simplicity.

Tab. B. Matrix of O-D trip purposes.

<table>
<tr><th>Destination purpose
Origin purpose</th><th>Home</th><th>Usual work-place</th><th>Professional</th><th>Shop-ping</th><th>Leisure</th><th>Private</th><th>School, Univer-sity</th><th>Other</th></tr>
<tr><td>Home</td><td></td><td>1</td><td>3</td><td colspan="5">5</td></tr>
<tr><td>Usual workplace</td><td>2</td><td></td><td rowspan="2">7</td><td colspan="5" rowspan="3">8</td></tr>
<tr><td>Professional</td><td>4</td><td colspan="2"></td></tr>
<tr><td>Shopping</td><td rowspan="5">6</td><td colspan="3" rowspan="5"></td></tr>
<tr><td>Leisure</td></tr>
<tr><td>Private</td></tr>
<tr><td>School, University</td></tr>
<tr><td>Other</td></tr>
</table>

Each zonal trip production (or attraction) is formulated as a linear combination of seven land-use attributes:

- number of inhabitants,
- number of employed people,
- number of working positions,

- number of commercial working positions,
- number of leisure-related working positions,
- number of service-related working positions,
- number of student positions.

F2 Linear regression

Linear regression is the basic statistical procedure to estimate the parameters variables α_n and α'_m in linear generation models.

F2a Basic assumptions

Let us assume that

$$P_o = a + bX_o + e_o$$

in which the index *o* identifies an observation, P_o is an observed zonal production, X_o is an observed zonal attribute and e_o is a statistical error that includes design error as well as measurement error.

We denote by O the set of observations and we further assume that the errors e_o are independent samples of identically distributed random variables ε_o, with null mean and variance of σ^2.

F2b Estimators

Then the least squares estimates of *a* and *b* are respectively

$\hat{a} = \overline{P} = \frac{\sum_{o \in O} P_o}{|O|}$ where $|O|$ denotes the number of observations in *O*, and

$\hat{b} = \frac{\sum_{o \in O} P_o (X_o - \overline{X})}{\sum_{o \in O} (X_o - \overline{X})^2}$ where $\overline{X} = \frac{\sum_{o \in O} X_o}{|O|}$.

These estimators are unbiased (i.e. the means of $\hat{a}$ and $\hat{b}$ over all possible samples are the true values of *a* and *b* respectively). The variances are formulated in the following way:

$$\text{var}(\hat{a}) = \sigma^2 / |O|$$

$$\text{var}(\hat{b}) = \sigma^2 / \sum_{o \in O} (X_o - \overline{X})^2 .$$

When the residual variance σ^2 is unknown, it can be estimated by the following estimator

$s^2 = \frac{1}{|O|-2} \sum_{o \in O} (P_o - \hat{P}_o)^2$, in which $\hat{P}_o = \hat{a} + \hat{b} X_o$.

The standardized estimator $t_b = (\hat{b} - b) / \sigma_b = \frac{\hat{b} - b}{s / \sqrt{\sum_{o \in O} (X_o - \overline{X})}}$ is a Student random variable with $|O|$-2 degrees of freedom. It enables one to test a null hypothesis {b = 0} by

measuring the critical probability of observing the sample under this assumption. This is the probability that a Student variable with $|O|$-2 degrees of freedom has greater absolute value than t_b.

F2c Extension

These formulae readily extend to the multidimensional case, in which both b and X_o are vectors of several numbers.

F3 Category models

In category models, the individual trip-makers are segmented into mobility groups, each of which is endowed with its own mobility rate per trip purpose.

F3a Mobility segments

We shall denote by p a trip purpose; it may be a compound index including several index variables such as destination purpose and time period. Let c be a mobility segment, i.e. a class of trip-makers who exhibit homogeneous segmentation variables (eg. household type, household size).

F3b Trip generation rates

It is assumed that trip generation rates are relatively stable for given segments. Let us denote by $r_p(c)$ the trip rate of segment c for purpose p, and by $S_i(c)$ the size of segment c (eg. number of persons) in zone i. Then the zonal trip production of segment c for purpose p is

$$P_{cp}(i) = S_i(c) r_p(c) .$$

The trip production with purpose p in zone i is $P_p(i) = \sum_c S_i(c) r_p(c)$.

The zone may of course be one of the segmentation variables.

F3c Usage

Like zonal generation models, category generation models possess only an empiric outreach. As regards passenger traffic the segmentation with respect to exogenous variables such as income, car ownership, household structure, household size, category of home, value of land, density and modal accessibility looks more sensible than zonal disaggregation. As regards freight traffic, segmentation variables include commercial area for stores, number of sales, number of employees and of course product type.

The estimation of a category model involves statistical procedures such as linear regression to estimate the trip rates. However there is no statistical goodness-of-fit measure to assess the relevance of the segmentation.

On forecasting with a category model, the segment trip rates are parameters whereas the segment sizes are exogenous variables. Thus it is necessary to forecast the segment sizes in order to apply the category model. This can be achieved by a demographic model which forecasts the evolution of a given population from stage to stage on the life (or family) cycle. The stages may be defined by way of the following break points:

- the appearance of children,
- the time when the youngest child reaches school age,

- the time when a youth leaves home and either lives alone, with other young adults, or marries,
- the time when all the children of a couple have left home but the couple has not yet retired,
- the time when all members of a household have reached retirement age.

F3d Case study: the French National Household Travel Survey (1993)

This is a nation-wide survey repeated every tenth year or so (1966, 1973, ,1981, 1993). In 1993, 14,200 households were surveyed and gave the following pieces of information:

- Socio-economic attributes, including number of persons, age, sex, profession, homeplace, workplace.
- Private means of transport: cars, parking places, motor cycles, bicycles, subscriptions to public transport networks.
- Daily local mobility.
- Long distance trips over siw months.

The following table indicates some results, either aggregate of disaggregate, with respect to persons (aged of six years or more).

Tab. C. Aggregate vs. Personal motorized travel consumption in France, 1993.

Indicateur	Local trips	Other short trips	Long trips	Total
M trips per week	847	21	14	882
Nb of trips per person and per week	15.9	0.4	0.3	16.6
Average length (km)	9.8	21.4	383	16.1
Km per person and per week	157	8	102	267
Average duration (mn)	19	32	254	23
Average speed (km/h)	31.4	40.1	90.5	42.0
M Voy.km, per week	8 300	450	5 410	14 160

The darkened line corresponds to category trip rates for the three purposes of local trips (i.e. with length less than 100 km, trip ends less than 80 km away from home), other short trips (with length less than 100 km), or long trips.

F4 Conclusion

Most current generation models have no explicit economic content. Frequency choice models hardly constitute an alternative, since trip-makers probably do not think of trip frequency as the issue of a choice. Land-use models may be valuable: they deserve and require further development.

G Economic distribution models

The distribution of a commodity from warehouses to customers consists in retrieving quantities stocked in the warehouses and transporting them to the customer's locations, so as to meet the customer's demand, eventually at a least total cost. In the spatial distribution of trips, the warehouses are origin zones whereas the customers are destination zones.

We shall now introduce the main economic models of trip distribution, and give their economic interpretation. The gravity model may be the best-known demand function: however it has little economic significance, unless it is derived from destination choice models. There are two families of destination choice models, zone choice models versus activity choice models. In zone choice models, each destination zone attracts trips because of aggregate attributes (eg. number of inhabitants, number of employment). In activity choice model, each individual activity may attract trips.

G1 Activities, zones and trips

G1a From activities to trips

Each trip is a movement from its origin point to its destination point. Its extreme points are distinguished from their intermediate points because they correspond to sufficiently long stops, that enable one to perform an activity.

These activities appear as the underlying reason for trips.

A transport model in which activities are explicit and attract or produce the trips is a causal model of trips. In economic theory, the demand for trips is said to derive from the demand for activities.

G1b From individual activities to zone attractiveness

Rather than considering the activities on an individual basis, we can consider them in an aggregate way by grouping them according to their location. Each group of locations may coincide with a transport zone, origin and destination. The activities located in a given zone make up a statistical population, with aggregate attributes such as the total number of them: thus the aggregate attributes are associated with the zone. Some of them may serve as measures of the zone attractiveness with respect to the potential activity consumers.

In fact, activities can also be grouped according to criteria other than location. The main criterion is the nature of activity, usually defined as the purpose of activity: home, work, education, shopping and leisure, are possible purpose classes for passenger transport. Each of them may be subdivided, eg. work between white-collar work or blue-collar work. In freight transport an analogous distinction of purpose applies to final and intermediate goods. Final goods can be subdivided into residential usage, industrial usage, commercial usage or services; whereas intermediate goods correspond to an industrial purpose.

To sum up, activities are usually grouped together with respect to both location and purpose. The total number of activities for a given purpose is an important zonal attribute.

G1c Supplied activities versus serviced activities

Presumably there are many more supplied activities than serviced activities. An urban instance pertains to theatre or cinema, when only part of the available seats are booked. There may also be unsatisfied work proposals. Thus a distinction should apply between the number of supplied activities and the number of serviced activities, but this would be impractical.

G2 The gravity model

This model is based on a physical analogy with Newton's law of gravity in mechanics.

G2a Principle

In the basic gravity model of trip distribution, every O-D flow is proportional to the production of the origin zone, to the attraction of the destination zone and to the reciprocal of the squared distance:

$$q_i = KP_o A_d /(d_i)^2 ,$$

in which K is a constant of proportionality, P_o is the production of origin zone *o(i)*, A_d the attraction of destination zone *d(i)* and d_i the distance from zone *o* to zone *d*.

By extension, a formulation of O-D flows as a product of variables related to the origin and destination zones or to the O-D movement is also called a gravity model. This includes the following log-linear model

$$q_i = K \exp(\sum_n \alpha_n X_{n,o} + \sum_m \alpha'_m X'_{m,d}) f(G_i)$$

in which the variables $X_{n,o}$ are origin attributes, $X'_{m,d}$ destination attributes, α_n and α'_m are parameters, G_i is the generalized cost from origin to destination and f is a deterrence function (often a decreasing one).

G2b Application

The usage of a gravity model requires the following data:

- Production variables associated with the origin zones (ex. number of inhabitants),
- Attraction variables associated with the destination zones (ex. number of working positions),
- Deterrence variables from origin to destination (e.g. travel time and price),
- A deterrence function (e.g. the negative exponential function); this includes the coefficients of the production and attraction variables.

The outreach of the gravity model is essentially descriptive. There is little knowledge about how parameters change in time or space. The main application, if not backed by a destination choice interpretation, is to disaggregate an O-D matrix with respect to sub-zones.

G2c Margin constraints

A deterrence function is often used together with margin constraints, such as zonal productions and attractions. In a singly constrained model, only the zonal production (or attraction) margins are fixed, whereas in a doubly constrained model both production and attraction are fixed.

The margin constraints determine the coefficients π_o and α_d in the O-D flow formula:

$$q_i = P_o \, \pi_o \, A_d \, \alpha_d \, f(G_i) \, .$$

G3 Zone choice models

G3a Discrete choice from among destination zones

Each destination zone may be considered as a variant in an economic choice, of which the decision-maker is the destination-user and also the trip-maker. This fits into the framework of random utility theory, and we may associate each destination zone *d* with a random utility function U_d .

By assuming that there is a decision-maker for the destination of a trip, we also assume there is an origin point from which the available destinations are evaluated: thus we consider utility functions U_{od} rather than U_d to mark the dependency upon the origin zone.

The utility function takes into account attributes of the destination, of the movement from origin to destination and also of the origin zone if there are specific O-D pair effects.

When in linear form, the influence of destination attributes is separated from that of movement attributes, which contribute to the utility of the same way as in mode choice or route choice (the trip generalized cost, up to a minus sign, is included in the destination utility function).

G3b Derivation of the gravity model

A logit choice model with linear utility function $U_{od} = \sum_n \alpha_{on} X_{n,d} - \theta G_{od} + \varepsilon_{od}(\omega)$ yields the following market-share formula:

$$\Pr(d \mid o) = \frac{\exp(\sum_n \alpha_{on} X_{n,d} - \theta G_{od})}{\sum_{d'} \exp(\sum_n \alpha_{on} X_{n,d'} - \theta G_{od'})} \, .$$

As the denominator does not vary with respect to the destination zone, this expression is identical to a gravity model with an origin margin constraint. This provides a random utility foundation to the gravity model, or more precisely to certain gravity models.

G3c Discussion

Discrete choice theory provides both an economic interpretation and statistical methods to distribution models. However several strong assumptions deserve emphasis. First assumption is that the utility function of a destination represents its attractiveness: there is no guarantee since the underlying activities are not made explicit.

Second assumption is that the comparison of utility functions represents the choice among destinations: the individual choices are grouped with respect to the origin zones, thus aggregate choice is modelled rather than individual choice.

Third assumption is specific to the logit model: the similarities or dissimilarities across zones are not taken in account, except in the zonal attributes.

G4 Activity choice models

We shall now describe an economic model of choice among activities: it is assumed that each activity has a given value for every potential customer, and that each consumer chooses the best vacant activity, i.e. the one with maximum net value after subtraction of the transport price.

G4a Zonal distribution of activity utility

For each destination zone *d*, let us order the supplied activities with expect to increasing utility (e.g. earnings for work purpose). The resulting statistical population has size A_d and a structure described by a cumulative distribution function, F_d. For a given utility *v*, there are $A_d F_d(v)$ activities with utility lower than v in zone *d*.

Activity customers, when supplied with these activities of zone *d*, choose the best available ones, those with highest utility. Under the assumption that no activity can be serviced to more than one customer, a number of T_d customers corresponds to the T_d activities with highest utility, hence a proportion T_d / A_d of the population. Thus the minimum utility of the serviced activities is $\tilde{v}$ such that $1 - F_d(\tilde{v}) = T_d / A_d$, or equivalently $\tilde{v} = F_d^{-1}(1 - T_d / A_d)$.

The cut-off value $\tilde{v}$ represents the minimum utility of serviced activities, as well as the maximum utility of vacant activities in the zone (provided that F_d is strictly increasing at $\tilde{v}$). Thus it measures the value of zone *d* for a marginal customer (before subtracting the cost).

G4b Destination zones as competitors

A marginal customer located in an origin zone *o* demands the best available activity, the one which maximises the net utility equal to gross utility minus transport cost.

Among the vacant activities of zone *d*, the best net utility is $\tilde{v}_d - G_{od}$. The marginal customer chooses an activity, hence a zone *d*, which maximises $\tilde{v}_d - G_{od}$. This destination zone is efficient with respect to the origin zone O.

The greater the number T_d of consumers that choose zone *d*, the less attractive the zone *d* is for marginal consumers since $\tilde{v}_d$ is a decreasing function of T_d. Thus there may be several efficient destination zones with respect to the origin zone: the equilibrium condition between zone *d* and zone *d′* is:

$$\tilde{v}_d - G_{od} = \tilde{v}_{d'} - G_{od'} = \pi_o$$

since the common value is specific to the origin zone. The value π_o represents the best net utility available from zone *o*. It measures the accessibility of zone *o*, or more precisely the accessibility to vacant activities from zone *o*.

G4c Binary model

In the binary case of a single origin zone *o* and two destination zones *d* and *d'* there is a one-to-one correspondence between the number of customers originating from zone *o*, and the accessibility π_o.

First, we should determine whether only one destination is efficient or both. In the absence of customers, the more valuable destination is the one with higher net value $\tilde{v}_d - G_{od} =$

$F_d^{-1}(1) - G_{od}$. Assuming it is zone *d*, we must check whether $F_d^{-1}(1 - E_o / A_d) - G_{od} \geq F_{d'}^{-1}(1) - G_{od'}$.

If positive, zone *d* gets all the demand, hence we obtain the first form of the one-to-one correspondence:

$$\pi_o = F_d^{-1}(1 - E_o / A_d).$$

If negative, the other destination also gets customers. Thus, if both destinations are efficient, we obtain the second form of the one-to-one correspondence:

$$E_o = T_d + T_{d'} = A_d(1 - F_d(\pi_o + G_{od})) + A_{d'}(1 - F_{d'}(\pi_o + G_{od'})).$$

G4d Derivation of the gravity model

Under the empiric assumption that the utilities of activities follow an exponential distribution, the cumulative distribution function of zone *d* is formulate as $F_d(v) = 1 - \exp(-\lambda_d v)$.

From the single origin zone *o*, an efficient destination *d* has cut-off value $\tilde{v}_d$ such that $\tilde{v}_d - G_{od} = \pi_o$, hence $T_d = A_d \exp(-\lambda_d (G_{od} + \pi_o))$. All destinations are efficient from zone *o* since all of them have some activities with very large utility. By summing over destinations, we obtain the following gravity formula

$$\frac{T_d}{E_o} = \frac{A_d \exp(-\lambda_d G_{od})}{\sum_{d'} A_{d'} \exp(-\lambda_{d'} G_{od'})}.$$

G5 Conclusion

Economic distribution model provide a foundation to transport economics. We described three families of models with increasing economic outreach. The gravity formula can be derived within each family, as well as by a maximum entropy argument: same formal model for different semantic models.

H Empiric distribution models

After studying economic distribution models related to destination choice, we shall focus on the empiric, or statistical, side of the distribution-problem: the goal is to estimate O-D flows, or parameters of a distribution model, on the basis of observations.

We shall first describe the estimation problem in terms of inputs and outputs. Then we shall indicate standard estimation methods for distribution parameters, based on maximum likelihood or least squares. Lastly we shall introduce the maximum entropy method to infer O-D flows from data such as link traffic counts.

H1 The estimation problem

H1a Notation

Let *i* denote an O-D pair with O-D trip rate q_i, generalized cost G_i and specific attributes X_{in}. Let *o(i)* be the origin zone of O-D pair *i* and *d(i)* its destination zone.

Let $\Theta = (\theta_n)_n$ be a vector of parameters.

We may also denote by $q_i = D_i(\Theta, X)$ a functional dependency of q_i on parameter Θ and exogenous variables $X = (X_{in})_{i,n}$.

Sums of O-D flows may also be of interest: let *a* be an index for a such count, x_a be the related total and π_{ai} the proportion of O-D pair *i* in count *a*. Thus $x_a = \sum_i \pi_{ai} q_i$.

H1b Modelled versus observed variables

A model variable, say x_a, corresponds to specific conditions: for example to the average weekday in interurban traffic or to the average morning peak hour in urban traffic.

An observation of x_a, denoted by x_a^t in which *t* is the index of observation, will in general differ from the true value because of random sample error. If the error is additive this may be formulated as $x_a^t = x_a + \varepsilon_a$ in which ε_a^t is a random error. However in distribution problems other formulations of error may be more realistic.

Most observations available for an empiric distribution model are points (X_a^t, x_a^t), where X_a^t is the subvector of X related to the O-D pairs that contribute to x_a^t (those with $\pi_{ai} > 0$).

H1c Inputs and outputs

The input variables include observation points (X_a^t, x_a^t), proportions π_{ai} and functions D_i.

There may be two kinds of outputs. On one hand, the O-D flows q_i are direct outputs, obtained by inference rather than by estimation because the number of them usually exceeds by far the number of observations. On the other hand, model parameters θ_n are 'indirect' outputs which may obtained by standard estimation methods.

H2 Classical estimation methods

Several estimation methods have been proposed to estimate the parameters of a distribution model: maximum likelihood estimator with Poisson-distributed flows or normal flows, least squares or generalized least squares. We shall indicate a simple maximum likelihood method.

H2a Assumptions

Each O-D flow q_i is assumed to depend on exogenous variables X and parameters Θ on the basis of the following demand function, $q_i = D_i(\Theta, X)$. Thus every aggregate flow x_a depends on parameter Θ on the basis of $\hat{x}_a(\Theta) = \sum_i \pi_{ai} D_i(\Theta, X)$.

We hereafter assume that modelled flow variables are trip rates, i.e. ratios of number of trips to a given duration.

Observations consist in link counts indexed by t, with duration d_t and a count of n_t trips. It is assumed that each observation is a Poisson random variable with mean intensity $\hat{x}_t = \hat{x}_{a(t)}(\Theta)$. Thus the probability of counting n_t trips during d_t units of time is

$$\Pr(n_t; d_t; \hat{x}_t) = \exp(-\hat{x}_t d_t)(\hat{x}_t d_t)^{n_t} / n_t!.$$

H2b Solution by maximum likelihood

The likelihood of an observation is also $L_t(\Theta \mid d_t, n_t) = \exp(-\hat{x}_t d_t)(\hat{x}_t d_t)^{n_t} / n_t!$. Under the simplifying assumption that observations are mutually independent, the joint log-likelihood of all observations is as follows

$$\Lambda(\Theta) = \sum_t \ln L_t(\Theta) \approx \sum_t n_t \ln(\hat{x}_t d_t) - \hat{x}_t d_t - n_t(\ln n_t - 1)$$

$$\Lambda(\Theta) \approx \sum_t n_t (\ln(\frac{\hat{x}_t d_t}{n_t}) - 1) - \hat{x}_t d_t .$$

By maximizing $\Lambda(\Theta)$ with respect to Θ, we obtain the maximum likelihood estimator Θ^{ML}.

H2c Comments

Several sources of uncertainty are absent from this estimation process: there is no specific O-D pair error on q_i, no measurement error on n_t. Thus the results may not be realistic in problems with specific O-D pair relationships.

H3 Maximum entropy inference

H3a Definition of entropy

Entropy measures the uncertainty associated with a set of disjoint events. To clarify this definition we shall consider a textbook case. Let E be a set with N elements e_n (e.g. trips) which are to be distributed between M disjoint sets E_m (e.g. O-D pairs). In order to describe fully a distribution it is necessary to know the identity of all the elements in each subset E_m. A more cursory description involves stating N_m the number of elements in each subset E_m. Each cursory description, $(N_m)_{m=1..M}$, satisfies the constraints $N_m \in \{0,1..N\}$ and

$\sum_{m=1}^{M} N_m = N$. It corresponds to several complete descriptions $(y_{nm})_{n=1..N,m=1..M}$ where $y_{nm} \in \{0,1\}$ is a binary variable equal to 1 when e_n belongs to E_m or 0 otherwise (therefore $\sum_{m=1}^{M} y_{nm} = 1$). Precisely, there are $W = N!/\prod_{m=1}^{M} N_m!$ full descriptions which are compatible with a cursory description (this is the number of permutations of N elements between M subsets each of size N_m).

H3b Maximum entropy and minimum cross entropy

With no information other than the total number of elements N and the number of subsets M what is the most probable cursory description? If we assume that each of the elementary permutations is of equal probability, the cursory description which maximizes W corresponds to the largest number of permutations. From Stirling's formula, when x is large $\ln(x!) \approx x(\ln x - 1)$, thus $\ln W \approx N \ln N - \sum_m N_m \ln N_m$ or alternatively, by using $p_m = N_m / N$ the frequency with which a randomly selected element belongs to the subset E_m, $\ln W \approx \sum_m p_m \ln p_m$. Thus maximizing W is also a question of maximizing the *entropy function* $H(p) = -\sum_m p_m \ln p_m$ where $p = [p_m]_m$.

By maximizing H subject to the constraints $p_m \geq 0$ and $\sum_m p_m = 1$, $p_m = 1/M$ is obtained, that is to say that elements are equally distributed between subsets. In this case the disjoint events are macroscopically indiscernable and our uncertainty about them is maximum. With more data (e.g. link traffic counts), it is possible to find the most probable cursory description within the constraints which apply to the additional information, by still maximizing H subject to the additional constraints. The *principle of maximum entropy* consists of selecting the probability distribution which the available information (observations) is just sufficient to calculate: thus we have the greatest chance of being right.

When prior information is available about the unknown distribution p, in the form of a prior distribution q, a related principle consists of selecting the distribution which matches the available information while being as similar as possible from the prior distribution, i.e. while minimizing the cross-entropy function $I(p:q) = \sum_m p_m \ln \frac{p_m}{q_m}$. When q is the uniform distribution, minimizing $I(p:q)$ amounts to maximizing $H(p)$. This second principle is known as the principle of *minimum cross-entropy*.

H3c Instances

The simplest instance of maximum-entropy matrix estimation is the singly-constrained problem in which only the row (or column) totals are imposed: in this case the problem can be decomposed into origin-based sub-problems and solved in a straightforward way. The most well-known instance is the doubly-constrained model in which both the row and column totals are imposed. This can be further refined by taking into account an old (or prior) matrix and minimizing the cross-entropy, subject to the rows and columns constraints; and alternatively or further by enforcing other constraints such as link traffic counts, provided that the trip rate proportions are available on the counted links (eg. on the basis of a network assignment). Another sophistication involves the network level-of-service, G_i: the O-D trip rate q_i may be assumed to depend on it on the basis of some demand function $q_i \propto g_i(G_i)$.

H3d Model formulation and solution

Let us consider the generic problem of estimating a trip matrix subject to destination (column) constraints and other, "external" constraints which may include origin margins as well as link traffic counts; we also assume that the trip rate q_i of the i-th O-D pair depends on the level of service, G_i, on the basis of $q_i \propto g_i(G_i)$. Let $q = [q_i]_i$ be the vector of the O-D trip rates. Let also $g_i = g_i(G_i)$ be the i-th demand function. The entropy objective function is stated as:

$$J(q) = -\sum_i q_i \ln \frac{q_i}{g_i},$$

which is to be maximized under the following two sets of constraints:

$$\sum_{i \in s} q_i - A_s = 0$$

$$\sum_i \pi_{ti} q_i - B_t = 0,$$

in which *s* denotes a destination zone, A_s is the total number of trips attracted by destination zone *s*, *t* denotes a traffic count, π_{ti} is the coefficient of the i-th O-D pair in the t-th count and B_t is the known number of trips in count *t*. In the case of an origin margin, B_t is the total number of trips generated by origin zone *t*, whereas π_{ti} equals 1 if *t* is the origin of the i-th pair or 0 otherwise. In the case of a link traffic count, π_{ti} is the proportion of the i-th trip rate which traverses the t-th link.

The primal maximum entropy program is also $\max_{q \geq 0} J(q)$, subject to $\sum_{i \in s} q_i - A_s = 0$ and $\sum_i \pi_{ti} q_i - B_t = 0$.

An old trip matrix $\tilde{q} = [\tilde{q}_i]_i$ may be taken into account by replacing g_i with $\hat{g}_i = \tilde{q}_i g_i$ or better with $\hat{g}_i = \tilde{q}_i g_i(G_i) / g_i(\tilde{G}_i)$ if the ancient level of service $\tilde{G}_i$ is known.

Let us associate multipliers α_s and β_t to constraints $\sum_{i \in s} q_i - A_s = 0$ and $\sum_i \pi_{ti} q_i - B_t = 0$ respectively. At the solution of the optimization program, for each O-D pair *i* with $q_i > 0$ it holds that

$$q_i = g_i \exp(\alpha_{s(i)} - 1) \exp(\sum_t \beta_t \pi_{ti})$$

When all of the external constraints correspond to origin margins, the second exponential term reduces to $\exp(\beta_{t(i)})$, yielding the doubly-constrained distribution model.

In the general case, summing over all O-D pairs *i* with a given destination *s* and letting $\Psi_s(\beta) = \sum_{i \in s} g_i \exp(\sum_t \beta_t \pi_{ti})$, we obtain that $\sum_{i \in s} q_i = \exp(\alpha_{s(i)} - 1) \Psi_s$, hence

$$q_i = g_i \exp(\sum_t \beta_t \pi_{ti}) / \Psi_{s(i)}(\beta).$$

H3e Computation by Bregman's method

The following multiproportional algorithm is also known as Bregman's method:

Step 0, Initialization. Set iteration counter $n = 0$. Fix a convergence level ε > 0. For all *i*, set $q_i^{(0)} = g_i$.

Step 1, External constraint processing. Increment iteration counter $n = n + 1$. For each O-D pair *i*, set $q_i^{(n)} = q_i^{(n-1)}$. For each external constraint *t*, solve for h_t the following equation: $\sum_i \pi_{ti} q_i^{(n)} \exp(\pi_{ti} h_t) = B_t$, then for every O-D pair *i* with $\pi_{ti} \neq 0$ update $q_i^{(n)} = q_i^{(n)} \exp(\pi_{ti} h_t)$.

Step 2, Column constraint processing. For each column constraint *s*, perform the following sequence: {compute $\rho_s = \sum_{i \in s} q_i^{(n)}$, then set $\sigma_s = 1/\rho_s$ and for all O-D pairs *i* with destination *s* update $q_i^{(n+1)} = q_i^{(n)} \sigma_s$}.

Step 3, Convergence test. If $(\sum_t (B_t - \sum_i \pi_{ti} q_i^{(n)})^2)^{1/2} < \varepsilon$, then terminate, else go to Step 1.

At the end of Step 2 indeed, all column constraints are satisfied. Step 1 vanishes when the singly-constrained model is considered. An external constraint *t* in which π_{ti} can only assume the values 0 or 1 is processed in Step 1 in the same way as a column constraint in Step 2, since the equation is reduced to $\exp(h_t) \sum_i \pi_{ti} q_i^{(n)} = B_t$. In the case of a link count, the corresponding equation may be solved by a one-dimensional Newton method: the problem amounts to minimizing the function $\sum_i q_i^{(n)} \exp(\pi_{ti} h_t) - B_t h_t$ which is strictly convex provided that there exists an O-D pair *i* with $q_i^{(n)} > 0$ and $\pi_{ti} > 0$.

H3f Example

We first considered the biproportional problem of estimating a 3x3 matrix with column totals $[368 \quad 533 \quad 254]$, row totals $[406 \quad 384 \quad 311]'$ and prior matrix $\begin{bmatrix} 107 & 160 & 100 \\ 160 & 210 & 107 \\ 88 & 123 & 100 \end{bmatrix}$.

The solution matrix was attained after three iterations with an accuracy of two digits after decimal point. The iterations are reported in the following table.

Tab. D. Successive iterations of Bregman's method.

Iteration	End of Step 1	End of Step 2	Convergence measure
1	$\begin{bmatrix} 134.11 & 200.54 & 125.34 \\ 128.81 & 169.06 & 86.14 \\ 88.00 & 123.00 & 100.00 \end{bmatrix}$	$\begin{bmatrix} 140.64 & 216.99 & 102.21 \\ 135.07 & 182.92 & 70.24 \\ 92.28 & 133.09 & 81.55 \end{bmatrix}$	5.89
2	$\begin{bmatrix} 140.69 & 217.06 & 102.25 \\ 133.60 & 180.92 & 69.48 \\ 93.51 & 134.86 & 82.63 \end{bmatrix}$	$\begin{bmatrix} 140.77 & 217.13 & 102.10 \\ 133.67 & 180.98 & 69.38 \\ 93.56 & 134.90 & 82.52 \end{bmatrix}$	0.038
3	$\begin{bmatrix} 140.77 & 217.13 & 102.10 \\ 133.66 & 180.96 & 69.37 \\ 93.57 & 134.91 & 82.52 \end{bmatrix}$	$\begin{bmatrix} 140.77 & 217.13 & 102.10 \\ 133.66 & 180.96 & 69.37 \\ 93.57 & 134.91 & 82.52 \end{bmatrix}$	0.00024

Next we considered the triproportional problem with same inputs, except for two additional constraints associated with trip length classes. The first of these constraints is $q_{11} + q_{23} = 214$. Then the solution matrix is $\begin{bmatrix} 143.00 & 216.26 & 100.74 \\ 132.01 & 180.99 & 71.00 \\ 92.98 & 135.75 & 82.26 \end{bmatrix}$, which is attained after eight iterations with an accuracy of two digits after decimal point.

H3g Comments

Maximum entropy inference of O-D matrix is widely used because it requires relatively little data and little computational effort. Its results are robust in that they differ as little as possible from the prior knowledge included in the model.

Data inconsistency may arise, eg. redundant yet inconsistent link traffic counts. This may be removed by taking into account observation error in an explicit way.

H4 Conclusion

There exists a variety of inference or estimation methods for the distribution problem. Inference by maximum entropy is the simplest method. However standard estimation methods are in command to estimate the parameters of a demand function.

I Systems of models

The complexity of a transport system is obvious as it is made up of several suppliers, many consumers, physical and economic phenomena with spatial and temporal heterogeneity and variability. This complexity pushes the analyst to select some of the elements and some of the relationships in order to examine them in a specific way, on the basis of a specific model.

Specific models of separate parts of the transport system can then be grouped together into a system of models, or better a structure of models. The resulting, compound model is purported to analyze a wide range of issues.

The objective of this lesson is to describe the basic operations of model composition in a systems analysis framework. First we introduce in turn bundle models which integrate a number of model components, sequences of models, parallel models, and complex models including feedback. Then we take again our four-part analysis framework to examine the main issues of composition, theoretical and practical.

I1 Integrated models

An assignment model integrates one or more of the following components:

- Choice from among variants (the paths),
- Rational choice (shortest path),
- Heterogeneity of consumers (demand classes),
- Demand function,
- Supply function (travel time functions).

The components that are linked together share some variables (i.e. demand volume, generalized cost) of which the interpretation is unique whatever the component. For instance the generalized cost in a car assignment model refers to travel along a car path and does not include parking charges. But in a mode choice model the generalized cost of the car should include parking charges, making the definition of the variable different.

We define an integrated model as a bundle of model components with same set of inputs and same set of outputs. This bundle can be considered as a black box, described by means of a set of input variables, a set of output variables and a quantitative transfer function that derive outputs from inputs.

I2 Serial combination

The integration of several model components induces a very strong relationship between them. Models may also be linked in a variety of weaker ways, including serial (or tandem) combination. This results in an oriented sequence of models.

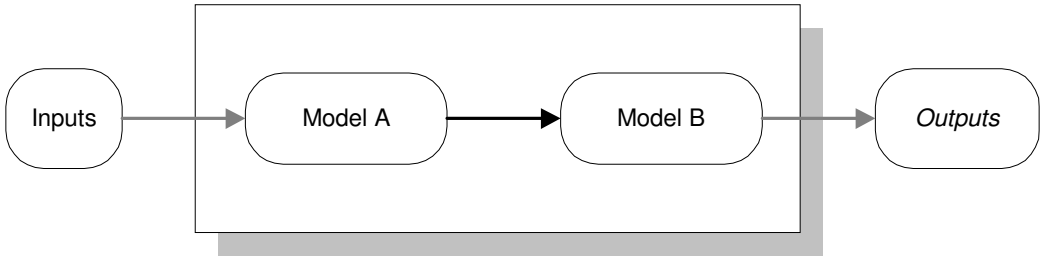
We define an oriented serial combination of two models *A* and *B* as the sequence of *A* and *B*, such that some outputs of model *A* are inputs to model *B*. Thus the former is phrased as an input model of the latter, which is an output model of the former.

For instance the car trip rate outputted by a mode-choice model may be input to a car assignment model.

An oriented sequence of models is a successive serial combination of several models. The most well-known instance in transport is the four-step model, i.e. a sequence of four models which address in turn generation, distribution, mode choice and route choice.

A caution note is in order here: the analyst must be aware that the interpretation of a variable, which is output of one model and input to another, may depend on the model. Eg. the generalized cost of car that is considered in a route choice model may differ from that in a mode choice model, which may also include parking charges.

Fig. j. Serial combination of two models.

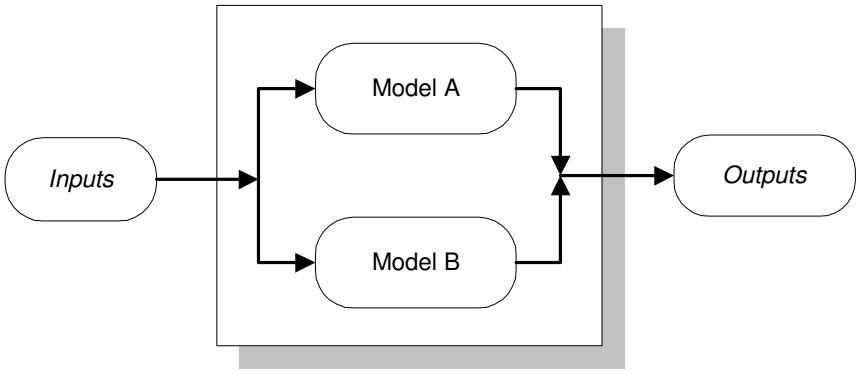


I3 Parallel combination

Two models are combined in a parallel way if their inputs include outputs from a third model (not necessarily the same variables, but produced by the same model), and/or some of their outputs serve as inputs to another model.

Eg. a car route choice model and a transit route choice model, of which the modal trip rates result from the same mode choice model, are parallel-combined models.

Fig. k. Parallel combination of two models.



I4 Feedback and complex structures

A complex structure of models is a graph-oriented structure, of which the arcs are bundle models and the nodes are sets of variables, also called information nodes.

This allows for serial and parallel combinations. A more complex relationship, though weak altogether, is the feedback combination, either direct or indirect.

There is a direct feedback between model *A* and model *B* is some outputs of *A* are inputs to *B* and simultaneously some outputs of *B* are inputs to *A*.

For instance, in a combined distribution and assignment model the distribution model outputs O-D trip rates which are inputs to the assignment model, whereas the assignment model outputs O-D trip costs which are inputs to the distribution model.

There is an indirect, or compound, feedback between model *A* and model *B* if some outputs of *A* contribute to the inputs of *B* either directly or through an oriented sequence of other models, and conversely some outputs of *B* contribute to the inputs of *A* either directly or through an oriented sequence of models.

In transport choice models, the main feedback is when model *A* supplies model *B* with primal, quantity variables, whereas model *B* supplies model *A* with dual, price variables.

I5 Composition issues

These can be decomposed with respect to their nature: semantic, formal, technical or empiric.

I5a Semantic issues

Semantic issues of model composition pertain to the significance of the variables and the phenomena (or behaviours): Is the meaning of a variable global (i.e. common to all models in the system), or relative to a given model? In the latter case there may be conflicts between the models that share this variable.

Are the phenomena consistent with each other? For instance, should the destination choice be addressed prior or posterior to the mode choice?

I5b Formal issues

Assuming that semantic issues have been dealt with in a satisfactory way, there remain formal issues such as:

Is there a global equilibrium state? If positive, is a global equilibrium unique? Is it stable?

I5c Technical issues

Technical issues pertain to the simulation of the global system, or equivalently to the solution of the whole structure of models.

Is there an integrated solver, or a system of solvers?

In the latter case, what about the technical uncertainty of each individual solver? How does it contribute to the overall technical uncertainty?

I5d Empiric issues

Is there an integrated database for all the ‘information nodes’, or several databases dedicated each to specific nodes?

In the latter case, what about the empiric uncertainty of each database?

I6 Conclusion

Structures of models enable the analyst to model a complex system by way of decomposition. Apart from bundle models, which are subsystems of models, the modelling blocks are linked together by weak relationships: serial combination, parallel combination and feedback. We provided a framework to analyze these relationships in a systems theoretic way. We also identified the main issues of decomposition.

J Appropriate level of analysis

A given transport planning study may be addressed by way of several alternative models. In the case of the Prado-Carénage tunnel, a first model was based on a Stated Preferences survey, whereas a second one was based on a Revealed Preferences survey together with a network assignment model. The cost of the study depends on the model to a large extent, as the model design and validation are costly operations. However the total study cost also includes other costs, related to scenario design and evaluation, study report, call for tenders, and eventually the level of quality which may induce a high cost, though an hidden one.

In general the customer of the study is willing to get high value for money. Under given moneay and time resources, he has to make the economic choice of which model to use. The objective of this lesson is to deliver some basic rules for model selection. The lesson contains three parts: Section 1 introduces minimal requirements for relevance. In Section 2, we describe the various trade-offs which relate respectively to the semantic, formal, technical and empiric part of the model; this is illustrated by continuous reference to the Prado-Carénage study. In Section 3, we put forward the foundation of a fair deal between the customer and the consulting firm: this should be considered prior to any call for tenders.

J1 Minimal requirements for relevance

J1a Relevance depends on semantic content

The overall objective of this course is to provide an economic understanding of transport demand models: our basic point is that model relevance can only be assessed with respect to the semantic content, whatever the mathematical formulae or the empiric performance.

Of course a rigorous mathematical formula and a good empiric performance are desirable features; but this is of little use in a planning study unless backed by a relevant semantic content, which provides the only scientific basis for transfer to the planning horizon.

J1b Assessment of model relevance

Model relevance can be assessed using the following guidelines, formulated as questions:

- Which elements of the system under study are modelled in an explicit way? For instance, do we explicitly represent the consumer, the available variants, the quality, the price?
- Which relationships within the studied system are modelled in an explicit way? For instance, does a model of microeconomic choice explicitly represent the reason for choice?

As a transport system involves many elements and many relationships, it is useful to distinguish between in-depth extension and in-breadth extension.

In-depth extension pertains to classes of elements and relationships: customers and suppliers are considered in an abstract, generic way. At this level of analysis, we distinguish between consumer groups, e.g. with respect to O-D pair, trip purpose or value-of-time.

In-breadth extension pertains to the objects within each class: eg. the transport zones and the associated O-D pairs, or the network arcs. The level of disaggregation determines the

accuracy of the results: if we model N categories of network arcs according to decreasing importance, only the flows of the N-1 first categories may be considered for further analysis.

J2 Cost-benefit analysis

Is a finer model more relevant than a coarser one? Do the additional benefits match the additional costs? These trade-off issues are central to the choice of an appropriate level of analysis. In order to address them, we shall discuss in turn each aspect in a model: semantic, formal, technical and empiric.

J2a Semantic content

A model may be finer than another one owing to semantic extensions, either in-depth or in-breadth.

In-depth semantic extension. This consists in basic concepts or behaviours captured in the model. This includes description (eg. More supply attributes in the generalized cost function), segmentation (eg. Choice-makers differentiated with respect to value-of-time), enriched behaviours (eg. Elastic demand).

In the Prado-Carénage study, the basic assignment model with one VoT and deterministic costs was enriched along two directions: first the disaggregation of VoT, second the stochastic generalized cost.

Generalized cost Value-of-Time	Deterministic	Stochastic
Single	**Basic**	**Plain logit**
Distributed	**Price-time**	**Distributed logit**

In-breadth semantic extension. This consists in additional objects encompassed in the model: these objects are analogous to objects already captured and do not imply in-depth extension. Instances include refinement (eg. Splitting a transport zone into subzones, or a trip purpose into subpurposes; consideration of less important roads or smaller parts of roads) and spatial or temporal extension (eg. Larger area, longer planning horizon).

In the Prado-Carénage study, the in-breadth dimension was purposedly reduced by aggregating O-D pairs with respect to mean travel time difference and mean network distance differences.

Advantages. In-depth and in-breadth semantic extensions enable the analyst to deal with more complex problems. Thus the scope for analysis is enlarged. The influence of additional factors may be analyzed, which makes the model more sensitive.

J2b Formal content

The formal content of a model consists in two parts: first the formal image of the semantic content, second the specific formal issues of characterization, existence, stability and uniqueness.

Formal image of semantic content. A refined semantic content has a finer formal image than the basic semantic content. However the formal extensions are quite easy owing to the genericity of mathematical formulation. This involves little additional cost.

In the Prado-Carénage case, the formulae have been available prior to the study; there was only an adaptation cost.

Specific formal issues. The state characterization of a refined model by way of a standard formula may be much more difficult than that of the basic model. Instances include segmented and stochastic assignment models. The resulting characteristic formula may be of the same kind as that of the basic formula, eg. A convex optimization program; in this case the investment cost of formal extension is readily amortized and may even be balanced by additional formal properties (eg. Uniqueness of solution in stochastic assignment models).

However the extended formula may be of a different, more complex kind than the basic formula: eg. Multiclass assignment with asymetric interaction between classes has no convex programming formulation and few properties of uniqueness and stability.

In the Prado-Carénage case, there was no equilibrium issue. Had such an issue arisen, the mathematical formula was available and there would have been little additional cost.

J2c Technical

The technical content of a simulation model involves a solver which produces a solution to the characteristic formula.

Availabillity of robust algorithm. In case of numerical simulation, the additional technical cost of a finer model depends on whether the characteristic formula is of the same kind as that of the basic formula or not. If positive, the basic algorithms can be extended at little cost. If negative the basic algorithms require further development and may loose many of their computational properties.

In the Prado-Carénage case, a specific integration method was used to compute the aggregate tunnel market share in the stochastic model with distributed VoT. The resulting numerical integration formula did not increase the computation cost by much; however the log-likelihood function was not concave and its maximization involved considerable effort.

On computation cost. The additional computation cost of a refined model may be assessed in terms of space or time complexity. As regards space complexity, more features imply more variables hence more storage space, but storage space has become quite unexpensive.

As regards time complexity, the additional complexity depends on whether an equilibrium is involved or not. In the equilibrium case, the refinements may greatly enhance the stability and then reduce the number of iterations to reach a given level of accurracy. However each extended iteration may be much more expensive than a basic one, same as the computation in the absence of an equilibrium.

J2d Empiric

The empiric part of modeling may be divided into representation, estimation and prediction.

Representation. This includes coding and editing the model attributes which are available or measurable in a straightforward way, eg. When the required information is public.

An extended model implies an extended representation: this may result in more expensive coding and editing. However some aggregate models are more expensive than their refined counterparts since they involve not only a refined representation but also aggregation of this at a specific aggregation cost: eg. Strategic transport models in which the network attributes are aggregated.

Estimation. Standard statistical methods require more detailed observation data to estimate a finer model, eg. More observation groups. The computation cost of estimation also becomes heavier, particularly so when the method looses some formal properties (eg. The concavity of the log-likelihood function in the linear logit model is lost in the probit model).

An alternative paradigm to standard statistical theory relates to information theory and encompasses inference by maximum entropy. In this alternative framework, it makes sense to 'estimate' the values of disaggregate parameters on the basis of aggregate information used in a specific way. This specific way consists in adding minimal implicit information to the data, implicit information being the opposite of the explicit model formula.

Prediction. This includes the simulation and the representation of scenarios. A finer model involves finer scenarios, which may be both an advantage where it is easier to represent detailed exogenous variables or a disadvantage where it is easier to represent aggregate exogenous variables.

In the Prado-Carénage case, the same data set was used to estimate all models, either with distributed VoT or not, either deterministic or stochastic. Each of the two in-depth extensions provided a natural framework for statistical estimation, which was an advantage over a single-VoT, deterministic model.

J3 Towards a fair deal?

J3a Elements of cost

A study customer should consider the following elements of cost:

- Semantic design, in-depth and in-breadth.
- Mathematical formula.
- Solver and computation costs.
- Empiric content: data coding and editing, estimation and inference of model parameters.
- Model validation by application to an unobserved situation.
- Scenario design.
- Scenario assessment.
- Report, communication and file-keeping.

When preparing a call for tenders, the study customer should invite the consulting firms to decompose the total costs with respect to those elements.

J3b Investment versus amortized costs

Most of the cost elements may be amortized over time. When the study customer is a transport agency, the study should be part of a continuous planning process and benefit from previous contributions.

Let us make the opportunities for cost amortisement more precise:

- Semantic design, by transfer.
- Mathematical formula, by transfer.
- Solver and computation costs, by use of compatible computers, same packages and permanent skilled operators.
- Data coding and editing, by re-use of validated databases.
- Estimation and inference of model parameters, by re-use of model and inputs.

- Model validation, by re-use of validated parameters.
- Scenario design, by re-use of dedicated databases.

The study customer should be careful to decompose the planning process into successive steps (and iterations), in order to divide the truly investment cost over time and also to maintain control over the steps. When a survey is required to feed a model and a study, it may be wise to consider the survey as a study of its own, and model feeding as part of a second, properly forecasting study.

J3c Internal versus external validation

The internal validation of a model consists in assessing its relevance, its mathematical consistency and its technical and empiric quality. This is indeed a skilled task, for which a study customer may not possess the required competence.

A solution would be to hire a second consultant to evaluate the work of the first one: this is internal validation, performed externally with respect to the study customer.

Another, yet complementary, solution consists in external validation by a series of common-sensical checks. These may include:

- Detailed presentation of network costs and modal split for one or several important O-D pairs.
- Identification of critical network arcs, for which the predicted flow exceeds the limit capacity. Mapping the ratio of flow to capacity will help detecting severe trouble, eg. when there are several oversaturated arcs in quite the same area.
- Traffic prediction when toll fare is zero, as compared to the prediction under realistic toll fare.

J4 Conclusion

The benefits of an enriched model pertain to its enlarged scope and deeper semantic content. These are balanced by formal, technical and empiric costs which may be relieved by either structural benefits (eg. Enhanced stability induced by deeper content) or scope economies which enable the analyst to amortize a model over a wide range of application.