Mathematical description of TRANUS

The activity and land use model

The purpose of the activity location model is to simulate a spatial economic system. Given a region or city divided into zones, the model estimates the activities that locate in each zone and the interactions that they generate for a specific time-period.

Basic concepts

The central element in the activity model is a spatial input-output procedure defined by economic sectors and their production and consumption relationships.

The starting point is the classical structure of an input-output model. The main elements are final demand, intermediate demand and primary inputs. The vector of final demands represents the final destination of production. In input-output models final demand usually includes private consumption, government consumption, exports and investments. The economic system must produce the quantities demanded in each sector; to achieve this, intermediate inputs are required, generating a production-consumption chain. In addition to intermediate inputs there are primary inputs; they include salaries, imports, profits and taxes. The sum of final demand plus all intermediate demands is equal to total production in the system. Similarly, the sum of all intermediate production plus primary inputs is equal to total production.

In TRANUS, the basic concepts of the input-output model have been generalized and a spatial dimension has been added. The term sector is much more general than in the traditional concept. It may include the classical sectors in which the economy is divided (agriculture, manufacturing, mining, government, etc.), factors of production (capital, land and labor), population groups, employment, floorspace, land, energy, or any other that is thought relevant to the spatial system being represented. The number and types of sectors must be defined according to the requirements of each application. The units in which each sector is represented (money, production, jobs, people, hectares, etc.) can also be defined to suit each case. This makes it possible to apply the model to urban or regional areas alike.

A first distinction can be made between *transportable* and *non-transportable* sectors. The main difference is that transportable sectors may be consumed in places different to those in which they were produced. For example, the demand for coal by a steel industry located in a particular place may be satisfied by mining industries located in other regions. Similarly, the demand for labor in a central area may be satisfied by population living in the outskirts. A typical example of a non-transportable sector is land or buildings; these must be consumed in the same place where they are produced. Consequently, transportable sectors generate trades or, in general, economic flows of goods, money or people. Such flows are later turned into demand for trips in the transport model. The transport system, in turn, must make such flows possible, and imposes transportation costs on them. By contrast, non-transportable sectors do not require transport and do not generate flows.

Another distinction is made between *internal* and *external zones*, and receive a different treatment in the model. All economic relations occur between internal zones. External zones are only used to represent imports and exports. However, it is possible to define external zones only for the transport model to represent external or through trips.

In turn, internal zones can be of two types: *first or second hierarchical level*. A first level zone may consist of two or more second level zones, thus affecting the spatial and sectoral distributions. A sector may be transportable at both first and second level zones, may be non-transportable, or may be only transportable at a second hierarchical level. The sum of all activities in second level zones will always equal the activities in the first level zone to which they belong.

Each sector located in specific zones is characterized by a number of specific associated variables. These are defined in the following paragraphs.

Exogenous production

It is the production not generated or demanded by other internal sectors. It is equivalent to final demand in input-output models. The location of exogenous production does not depend on the internal logic implicit in the model. Instead depends on elements not modeled or external to the system. Exogenous production is not subject to the spatial and sectoral distribution procedures of the model. It is a given input, to be added to endogenous, induced, production. The growth of exogenous production from one period to the next by sector and zone may be given, or alternatively study-wide values may be given together with distribution functions. Both approaches may be combined.

Induced Production

It is production generated by internal or external demands. It is allocated to zones with the spatial and sectoral distribution model. The growth of induced production depends on the growth of those sectors that demand it.

Exogenous demand

It is an additional demand to that generated internally. If this additional demand takes place in external zones, it is termed exports. The term exogenous demand, then, will always refer to that taking place in internal zones only. Exogenous demand is distributed together with induced demand. The growth of exogenous demand from one period to the next is dealt with by the incremental model.

Induced demand

It is determined by the consumption requirements of final demand sectors or by the intermediate activities.

Exports

Exports are defined as internal production of the study area consumed in external zones. It is represented as exogenous demand in external zones. The model allocates the required production among internal zones only.

Imports

Imports are defined as a demand in an internal zone satisfied by production in external zones. It is distributed together with the rest of induced demand. Imports compete with internal production, but may be restricted to fixed amounts given exogenously or estimated by the incremental model.

Capacity of production

Total production (exogenous+induced) in a sector and zone may be subject to restrictions. Maximum, minimum or both maximum and minimum restrictions may be imposed for particular sectors and zones.

Consumption cost

It is the average cost of a unit of input at the consumption zone (CIF). The consumption cost is, hence, equal to the production cost plus the transport cost from the production zone to the consumption zone.

Because consumption in a particular zone may be satisfied by production from several zones with different production and transportation costs, consumption cost is calculated as a weighted average.

Production cost

It is the unit cost of production at the production zone (FOB). It is calculated as the sum of the consumption costs of all required inputs, plus value added.

Equilibrium price

It is the value of a sector in a particular zone when production is constrained. It represents a scarce commodity with supply limited to a production restriction. If there are no restrictions, then the price is equal to the production cost. If demand exceeds production capacity, the price increases, generating an excess profit to the producer or *rent*. The opposite case is that of demand being less than a minimum constraint; in this case, the equilibrium price drops below production costs, generating a loss to producers.

Value added

It is the value of capital and labor that is added to all other input commodities in order to obtain a unit of production. Typically value added includes payments to capital (rent), to labor (salaries), taxes or subsidies, capital payments on equipment, and so on.

Consumption utility

It is the logarithmic average of the utility values used in the probabilistic distribution of demand to production zones. Uses transport utility instead of transport monetary cost.

Transport cost

It is the monetary expenditure needed to transport one unit of production from the production zone to the consumption zone. It is calculated by the transport model. In the case of commodities, transport cost is a unit cost. For other sectors, such as residents, this value represents transport expenditure, and depends on the number of trips in a given time period, that is, the time period being represented in the activity model (month, year, etc.). It has an influence on the cost of production.

Transport disutility

Also calculated by the transport model, includes the monetary cost but other elements as well, such as the value of time, subjective elements, etc. Transport disutility is always entered to the activity model as a unit cost, i.e. per trip. It influences the spatial-sectoral distribution of production.

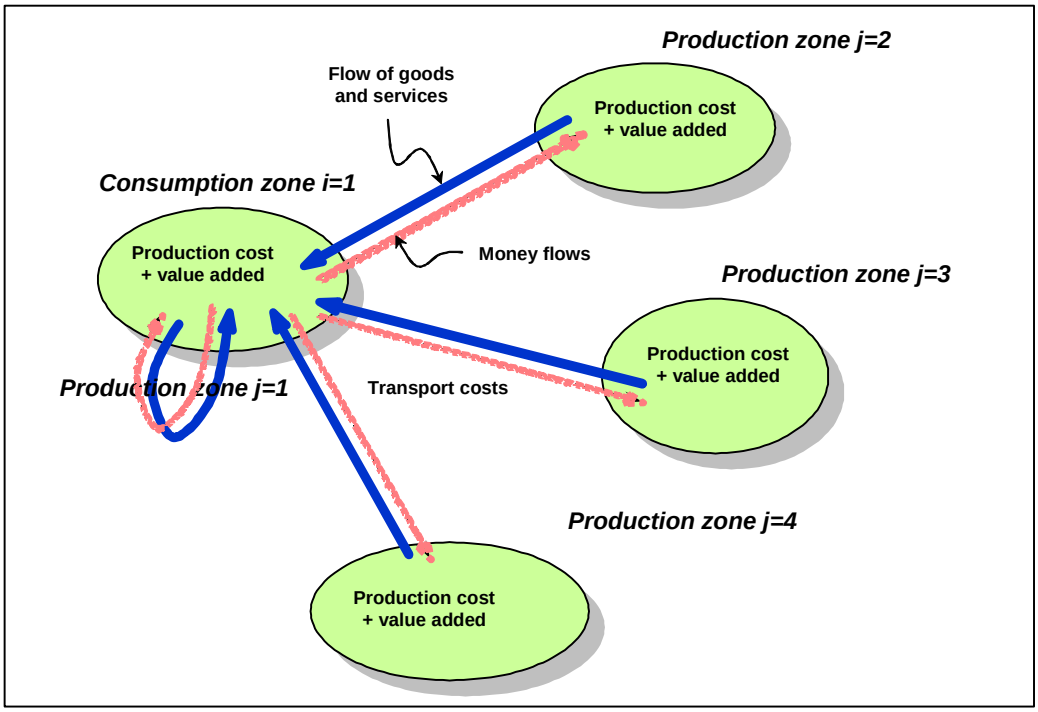
Demand and distribution of production

In principle, every sector requires inputs from other sectors. Hence, part of total production goes to intermediate consumption, and the rest may go directly to final consumption, whether internal or external (exports). Given a certain amount of final demand in one or more sectors and zones, the model estimates induced production through demand functions. The model then allocates this induced production to zones through spatial distribution functions. In turn, induced production requires further inputs, thus generating a *production chain* and the corresponding location of activities.

From the above relations, economic transactions are derived, giving rise to transactions or functional flows if production and consumption take place in different zones. These are defined as *transportable flows* involving people, goods, services or money. From these flows, transport demand is derived at a later stage. In some transactions, non-transportable commodities may be involved, such as land or floorspace. In this case, economic transactions are involved, but no flows are generated. Each sector may generate different transactions and different types of flows. A manufacturing industry, for example, may require non-transportable land and buildings as well as transportable raw materials, other manufactured goods and labor, thus generating a demand for transport.

As shown in Figure 1, a transaction involves a consumption zone and one or more production zones. Usually the consumption zone is, at the same time, a production zone. The model distributes the goods and services purchased among production zones in a probabilistic way. In the diagram, the blue arrows point in the direction of the flows of goods, while money for such purchases will flow in the opposite direction as shown with red arrows.

Figure 1: Relationships between production and consumption



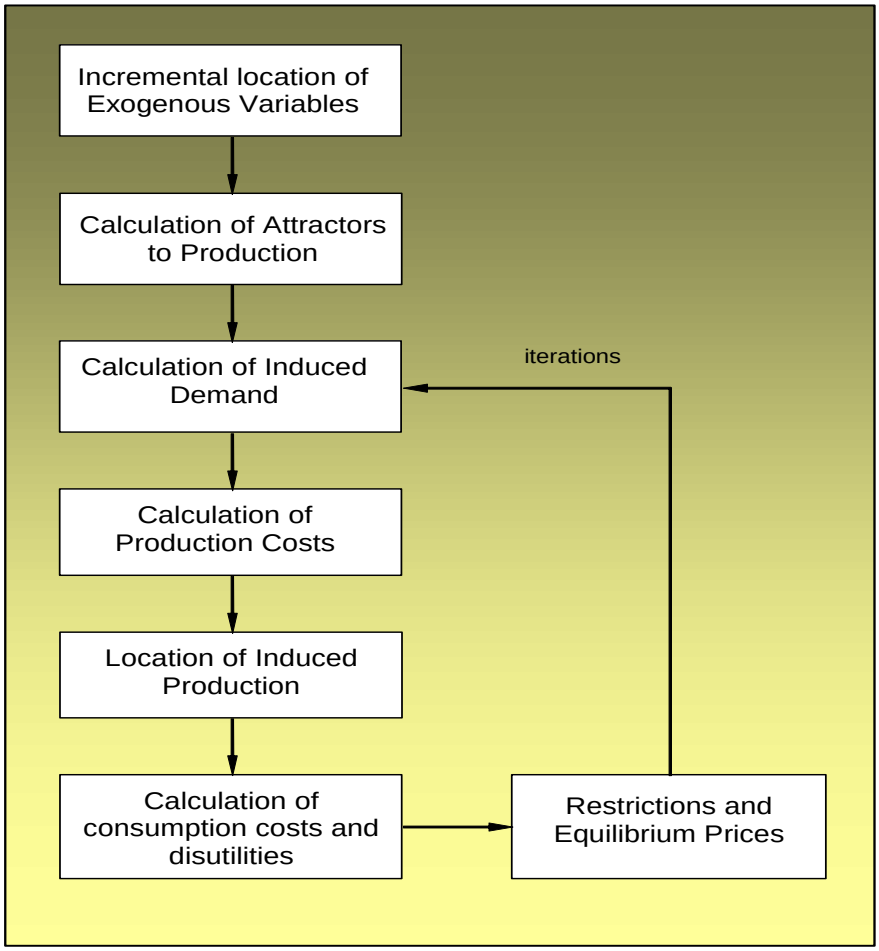
Structure of the activity model

The activity location model performs the calculation steps described in Figure 2:

- incremental location of exogenous variables
- calculation of attractors for induced production
- estimation of induced demand
- estimation of production costs

- location of induced production
- calculation of consumption costs and disutilities
- restriction checks and adjustment of equilibrium prices

Figure 2: Structure of the activity location model



Increments and location of exogenous variables

The first stage in the sequence of calculations of the activity location model is the estimation of the growth of exogenous variables in each sector and zone. By definition, exogenous variables depend on elements not simulated in the model: they are given inputs. Consequently, any increment of these variables in the future must be given to the model in the corresponding time-period.

The exogenous variables that may be modified for each time-period with the incremental model are:

exogenous production;
exogenous consumption;
capacity of production (restrictions);
exports;
imports;
initial attractors.

If *H* denotes any of the above exogenous variables, the increment for sector *n* between period *t-1* and *t* (positive or negative) in a specific zone *i* may be added exogenously as:

$$H_i^{n,t} = H_i^{n,t-1} + \Delta H_i^{n,t} \quad , \qquad (1)$$

However, for the first three exogenous variables, that is, production, consumption and restrictions, it is possible to specify a study-wide global increment. In this case, an attractor function must also be specified. The model will allocate the global increments to zones in the proportions that result from the attraction functions. Here the distribution of increments of total production is described. Exogenous consumption and restrictions use a similar formulation.

If $X_i^{*n,t}$ is the exogenous production of sector *n* in zone *i*, time period *t*, then:

$$X_i^{*n,t} = X_i^{*n,t-1} + \Delta X^{*n,t} \rho_i^{n,t} + \Delta X_i^{*n,t} \quad , \qquad (2)$$

where:

$X_i^{*n,t-1}$ exogenous production of sector n in zone i for time period t-1;

$\Delta X^{*n,t}$ global increment of exogenous production of sector n between t-1 and t;

$\Delta X_i^{*n,t}$ given increment of exogenous production of n in zone i for time period t (user defined);

ρ_i^{nt} proportion of the increment of n allocated to zone i for period t.

The proportion of the global increment assigned to each zone is a function of the attraction function:

$$\rho_i^{n,t} = \frac{A_i^{n,t}}{\sum_i A_i^{n,t}} \quad , \qquad (3)$$

where A_i^{nt} is the attractor of sector *n* in zone *i* for period *t*. There are two options to define the attractor function. The first option is a lineal function predefined in the model with three terms. The modeler may use one or more terms when defining the function.

$$A_i^{n,t} = \sum_k b^{nk} \left(\alpha^{nk} \tilde{X}_i^{k,t-1} + \beta^{nk} p_i^{k,t-1} + \chi^{nk} Q_i^{k,t-1} \right) , \qquad (4)$$

where:

b^{nk}	relative weight of sector *k* in the attraction function of sector *n*;
$\tilde{X}_i^{k,t-1}$	total production (exogenous + induced) of *k* in *i* at period *t-1*
$p_i^{k,t-1}$	price of *k* in *i* at period *t-1*;
$Q_i^{k,t-1}$	excess capacity (maximum constraint – total production) of *k* in *i* at period *t-1*
$\alpha^{nk}, \beta^{nk}, \chi^{nk}$	parameters regulating the relative importance of each element.

As a result, several sectors *k* may be combined into the attractor of a specific sector *n*, and this is regulated by the set of weights b^{nk}. Specific characteristics of the attracting sectors may be specified. For example, if only the price of the attracting sector is of interest, then $\alpha^{nk}=0$, $\beta^{nk}>0$, and $\chi^{nk}=0$.

As a second option, the modeler may freely define an attractor function for the distribution of global increments in exogenous variables. He or she may select the attracting sectors, variables and parameters of three types of equations: linear, power and logit, and also a constant term. Up to four equations may be used to define an attractor function.

Linear: $$A_i^{n,t} = C + \sum_k \left(a1^{nk} X1_i^{k,t-1} + a2^{nk} X2_i^{k,t-1} + \ldots \right) \quad (5)$$

Potential: $$A_i^{n,t} = C + \sum_k \left(\left(X1_i^{k,t-1} \right)^{a1^{nk}} \times \left(X2_i^{k,t-1} \right)^{a2^{nk}} \times \ldots \right) \quad (6)$$

Logit: $$A_i^{n,t} = \exp\left(C + \sum_k \left(a1^{nk} X1_i^{k,t-1} + a2^{nk} X2_i^{k,t-1} + \ldots \right) \right) \quad (7)$$

For each sector *n*, several attractor sectors *k* may be specified, *X1*, *X2*, …, that attract exogenous production of *n* and the appropriate parameters (*a1*, *a2*, ...). The following variables may be specified: Exogenous Production, Total Production (exogenous + induced), Exogenous Demand, Minimum Restriction, Maximum Restriction, Price and Capacity (Maximum Restriction - Total Production). The model takes the amount of the previous period. If some increments are given to specific zones in the current period, they are added to those of the previous period before the distribution of the global increment.

Calculation of attractors for induced production

In the case of induced production, attractors are calculated before the iterative sequence starts. The attraction functions are defined as follows:

$$A_i^{n,t} = \left(\sum_k b_k^n \left(\tilde{X}_i^{k,t-1} \right) \right) W_i^{n,t} \quad (8)$$

where:

$\tilde{X}_i^{k,t-1}$	total production (exogenous + induced) of a sector *k* attracting *n* in zone *i;*

b_k^n relative weight of sector *k* as an attractor to sector *n;*

$W_i^{n,t}$ initial attractor of zone *i* that takes into account non-modeled elements that attract the location of *n*.

The set of relative weights of each sector *k* in the attraction function of sector *n* may be different for first and second level zones.

Generation of induced demand

The amount of inputs that a unit of production of a sector requires from another sector is determined by a demand function. The model includes as options a fixed demand (equivalent to technical coefficients in an input-output model), variable (elastic) demand and the possibility of specifying substitutes. Land is a typical example of a substitute when different types of land are present in the system, such as low density, high density, industrial, commercial, etc.

The general form of the demand function is as follows:

$$a_i^{mn} = \min^{mn} + \left(\max^{mn} - \min^{mn}\right) \cdot \exp\left(-\delta^{mn} U_i^n\right), \qquad (9)$$

where:

a_i^{mn} amount of production of sector *n* demanded by a unit of sector *m* in zone *i*

$\min^{mn}$ minimum amount of *n* required by a unit production of *m*

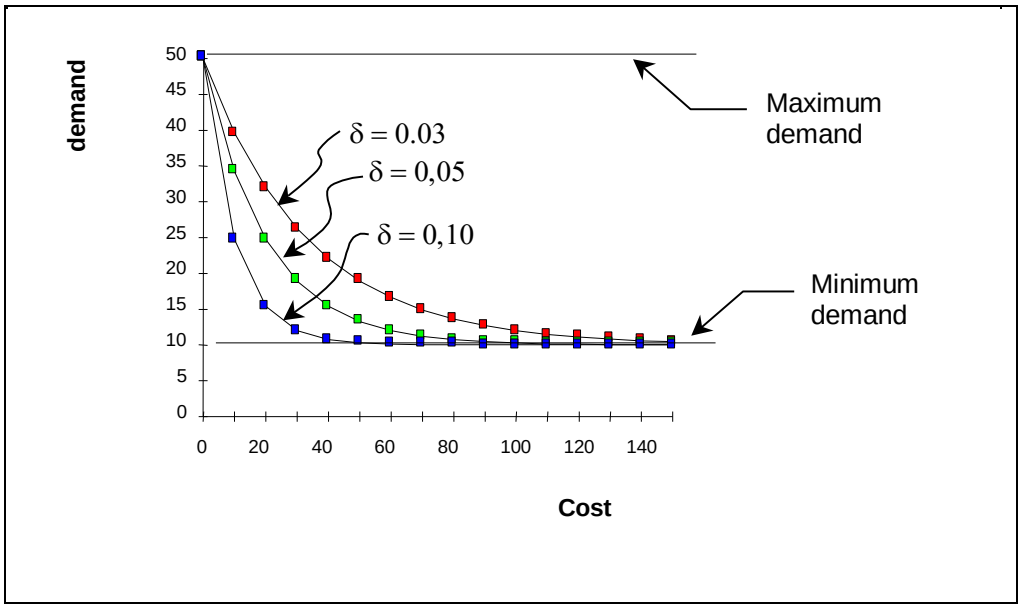
$\max^{mn}$ maximum amount of *n* required by a unit production of *m*

δ^{mn} elasticity parameter of *m* with respect to the cost of input *n*

U_i^n consumption disutility of *n* in *i*.

The resulting form of the demand function is shown in Figure 3. In this example a maximum consumption of *50* and a minimum of *10* is applied for different values of δ^{mn}.

Figure 3: Examples of demand functions



Next, the proportion applied to the demand function to take into account the presence of substitutes is estimated with a multinomial logit model (may be powit as well) of the form:

$$S_i^{mn} = \frac{\exp\left(-\delta \tilde{U}_i^{mn}\right)}{\sum_k \exp\left(-\delta \tilde{U}_i^{mk}\right)} \quad , \forall k, n \in K^n \qquad (10)$$

where K^n is a set of substitutes for *n*, and $\tilde{U}_i^{mn}$ is the scaled utility term of the substitutions model, defined as follows:

$$\tilde{U}_i^{mn} = \frac{a_i^{mn} \tilde{c}_i^{n} \varpi^{mn}}{\left[\min_k (a_i^{mk} \tilde{c}_i^{k} \varpi^{mk})\right]^{\theta^m}}, \qquad (11)$$

where the term $a_i^{mn}\tilde{c}_i^{n}$ is the amount of *n* that sector *m* is willing to consume in zone *i* multiplied by the consumption cost of *n* in *i*, thus representing *expenditure*. This expenditure is, in turn, multiplied by ϖ^{mn} which acts as a penalizing factor. The denominator in this equation scales the utility by dividing it by the utility of the best option, that is, the option with the least penalized expenditure. Finally, θ^m sets the degree of scaling; if $\theta^m=1$, the utility function is fully scaled; if $\theta^m=0$, the utility function is completely unscaled, with possible values in between.

The amount of inputs *n* demanded by sector *m* in zone *i* is, then:

$$D_i^{mn} = \left(X_i^{*m} + X_i^m \right) a_i^{mn} S_i^{mn} \qquad (12)$$

The total demand for inputs *n* in a particular zone *i* is the sum of the consumption of *n* by all sectors *m*, plus possible exogenous demands:

$$D_i^n = \sum_m D_i^{mn} + D*_i^n \,, \qquad (13)$$

where:

D_i^n total demand for *n* in zone *i*,

$D*_i^n$ exogenous demand for *n* in zone *i*.

In the first iteration the system will only have exogenous production and the induced production directly derived from it. In successive iterations, induced demand from the production of all sectors in the previous iteration are added.

Calculation of production costs

The production cost is calculated as the consumption cost of all necessary inputs to produce one unit of *m* in zone *i*, plus the value added to production:

$$c_i^m = \left(\sum_n D_i^{mn} \tilde{c}_i^n \right) + VA_i^m \,, \qquad (14)$$

where VA_i^m is the value added to the production of *m*, and $\tilde{c}_i^n$ is the consumption cost of input *n* in zone *i*.

Location of induced production

Once the amount of production demanded in each zone has been estimated, it must be distributed to production zones. If a sector is non-transportable, all production is assigned to the zone in which it is demanded. If the sector is transportable, demand is distributed to production zones with a multinomial logit model, in which the utility function of each zone is determined by:

$$U_{ij}^n = \lambda^n \left(p_j^n + h_j^n \right) + t_{ij}^n \,, \qquad (15)$$

where:

p_j^n is the price of sector *n* in the production zone *j*,

h_j^n is the shadow price of sector *n* in the production zone *j*,

t_{ij}^n is the transport disutility for sector *n* from the production zone *j* to the consumption zone *i*, and

λ^n is a parameter that regulates the relative importance of prices versus transport disutility in the utility function

The shadow price of production is estimated at calibration stage.

The results of the above calculation are divided by the utility of the best option to obtain the *scaled* utility:

$$\tilde{U}_{ij}^n = \frac{U_{ij}^n}{(\min_j U_{ij}^n)^{\theta^n}} \ , \qquad (16)$$

where θ^n sets the level of scaling of the utility term, $\theta^n = 0 \rightarrow 1$. If $\theta^n = 1$, then the model is said to be *fully scaled*; if $\theta^n = 0$, then the model is *unscaled*. These scaled utilities are entered into the multinomial logit model to estimate the probability that the production of sector *n* demanded in zone *i* is located in zone *j*:

¡Error! No se pueden crear objetos modificando códigos de campo., ***(17)***

$$X_{ij}^n = D_i^n \ \mathrm{Pr}_{ij}^n \ , \qquad (18)$$

where:

X_{ij}^n production of *n* located in the production zone *j* induced by activities in the consumption zone *i*,

A_j^n attractor term for the production of *n* in *j*,

α^n is a parameter that regulates the relative importance of the attractor versus the utility function in the location of sector *n*,

$\tilde{U}_{ij}^n$ scaled utility of location of *n* in zone *j* to satisfy the demand in zone *i*,

β^n dispersion parameter of the multinomial logit model.

If the consumption zone *i* is an internal zone, the distribution is applied to all zones, both internal and external. If the consumption zone is external (exports), the distribution is applied to internal zones only. In other words, the model does not allow the demand for exports to be satisfied by imports.

Finally, total induced production allocated to a zone is obtained by adding over all demand zones:

$$X_j^n = \sum_i X_{ij}^n \ . \qquad (19)$$

Consumption costs

Once demand has been assigned to production zones, consumption costs are calculated, that is, the amount that a sector *m* located in *i* has to pay for the consumption of one unit of input *n*. Because purchases are spatially distributed, an average is calculated, weighted by the price paid in each production zone plus the corresponding transport costs:

$$\tilde{c}_i^n = \frac{\sum_j X_{ij}^n \left(p_j^n + tm_{ij}^n \right)}{\sum_j X_{ij}^n} \quad , \qquad (20)$$

where:

X_{ij}^n amount of production of sector *n* demanded in *i* and produced in *j*,

p_j^n unit price of *n* in the production zone *j*,

tm_{ij}^n monetary cost of transporting a unit of sector *n* from the production zone *j* to the consumption zone *i* (different from transport disutility).

Consumption disutility

The disutility of consuming *n* in *i* is the logarithmic average of the disutilities used in the distribution to the production zones:

$$U_i^n = -\frac{\ln Pg^n}{\beta^n} (\min_j U_{ij}^n)^{\theta^n} \quad , \qquad (21)$$

Note that the expression is multiplied by the minimum disutility, because in the distribution the scaled disutility was used. This returns the original scaling to the composite utility. *Pg* is defined as a series of the following form:

$$Pg^n = \sum_{j=1}^{z} G_j \prod_{h=1}^{j-1} (1 - G_h), \qquad (22)$$

where the term G_j is the exponential element in the numerator of the probability that the production demanded in *i* locates in zone *j*, and z is the total number of internal zones. From equation (17):

$$G_j = \exp\left(-\beta \tilde{U}_{ij}^n\right), \qquad (23)$$

Checking for restrictions and adjustment of Equilibrium prices

The production of a sector in a zone may be limited to the minimum and/or maximum capacity of production. If the production assigned to a zone after the distribution lies within the established limits, the price is equal to the production costs plus value added. If, however, production is above the maximum or below the minimum, then the price is determined by demand-supply equilibrium. At the end of each iteration, the model checks for restrictions and adjusts the prices accordingly; the price is increased if the maximum restriction is violated, and is reduced if the minimum restriction is violated. These variations in price affect the distribution of production in subsequent iterations, until an equilibrium is reached. Prices are adjusted as follows:

$$p_j^{n,t} \begin{cases} < p_j^{n,t-1}, \left(X_j^{*n} + X_j^n\right) < R\min_j^n \\ > p_j^{n,t-1}, \left(X_j^{*n} + X_j^n\right) > R\max_j^n \\ = c_j^{n,t}, R\min_j^n = 0, R\max_j^n = \infty \end{cases}, \qquad (24)$$

where:

$p_j^{n,t-1}$ unit price of sector *n* in zone *j* in the previous iteration *t-1*

$p_j^{n,\tau}$ unit price of sector *n* in zone *j* in the current iteration τ,

$R\min_j^n$ and $R\max_j^n$ minimum and maximum restriction to the production of sector *n* in zone *j*

$C_j^{n,t}$ production cost of sector *n* in zone *j* in the current iteration

$X_j^{*n} + X_j^n$ total production: exogenous production + induced production of sector *n*

The rate at which prices are changed from one iteration to the next is also affected by a *smoothing parameter*.

Convergence

In each iteration the convergence in prices and production are evaluated. Both are calculated for each zone and sector as the percentage variation with respect to the previous iteration. These indicators are calculated separately for each sector, and adopt the value of the worst zone, that is, the zone that varied the most:

$$Cp_j^{n,\tau} = \max_j \left| \frac{p_j^{n,\tau} - p_j^{n,\tau-1}}{p_j^{n,\tau-1}} \right|,$$

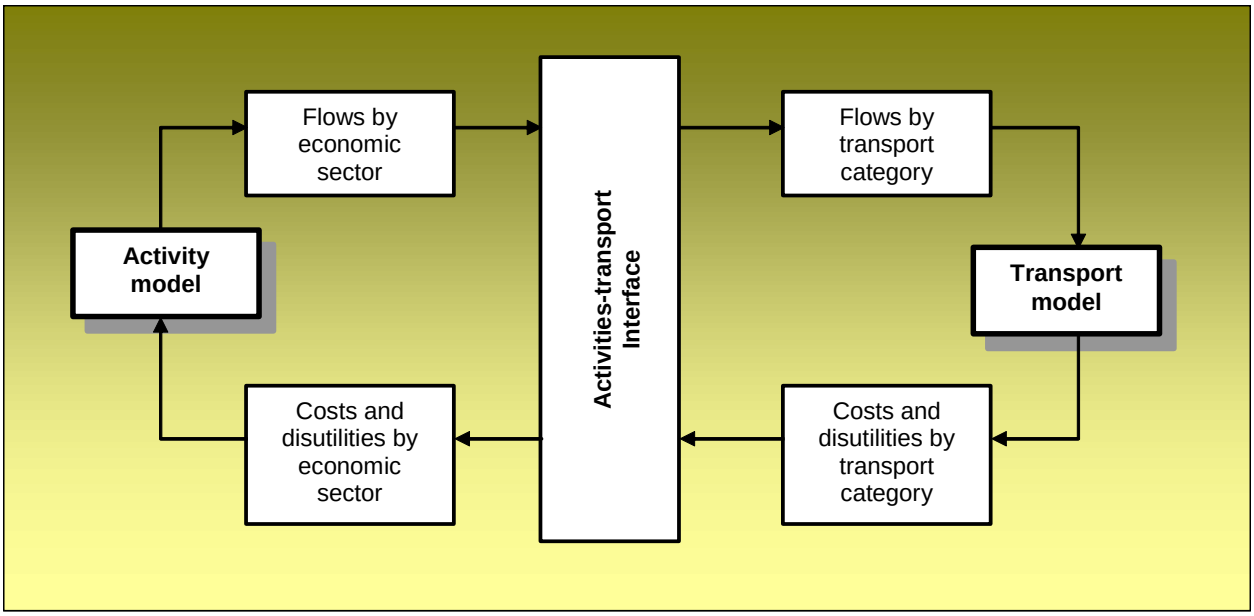
$$CX_j^{n,\tau} = \max_j \left| \frac{X_j^{n,\tau} - X_j^{n,\tau-1}}{X_j^{n,\tau-1}} \right|, \qquad (25)$$

where $Cp_j^{n,\tau}$ is the price convergence indicator, and $CX_j^{n,\tau}$ is the production convergence indicator. The model ends the iterative process when both convergence indicators are smaller than a pre-specified convergence criteria, or when a maximum number of iterations is reached.

Activities-transport interface

The activity location model produces as output, among others, a set of matrices of economic flows by (transportable) economic sector. Transport demand derivates from these flows in several transport categories. In turn, the transport model calculates transport costs and disutilities by transport categories, which must be transferred to the sectors that originated the trips. In other words, a two-way set of transformations must be performed: from activities-to-transport and from transport-to-activities. These are shown in Figure 4 below.

Figure 4: Activities-transport interface



The interface module performs other transformations to make compatible the units of time, magnitude and direction of the flows between the land use and transport model. The following transformations are possible:

- formation of transport categories from economic flows,
- time factor,
- volume/value factor
- direction of flows.

Each one of these possibilities is described in the following sections.

From economic flows to transport categories

The interaction between transportable economic activities generates flows of goods or people. In a regional application, for instance, the interaction between industries such as agriculture, manufacturing and so on, generate movements of commodities. Similarly, the interaction between employment and residents generate movements of commuters. Each transportable sector generates a corresponding matrix of flows. It may be that

some categories coincide with those in the transport model, but this may not be the case. For example, economic transactions between productive sectors (measured in money units) may generate movements in several transport categories such as heavy bulk, general cargo, containers, and so on.

The definition of transport categories occurs in the activities→transport direction. Once the transport model has finished its calculations, it will generate matrices of transport costs and disutilities by transport categories. The interface module transforms these back, to assign the corresponding quotes to the economic sectors that generated them. The transformation rules will be the same, but in reverse.

Consider the example of an agricultural sector that generates bulk, general and container transport categories in 60%, 30% and 10% proportions. The interface module will split the agricultural flows in these proportions. Next, the transport model will assign them to transport supply in terms of daily Tons. As a result, the transport model will estimate matrices of costs and disutilities for bulk, general and containers. These are fed back to the agriculture sector in the activities model. The interface weights each one by the same proportions in which they were generated. Thus, each cell in the transport cost matrix for agriculture will be the sum of 60% of the cost transporting bulk, 30% of transporting general cargo and 10% of transporting containerized cargo.

In other words, the number and types of sectors in the activity model does not have to match transport categories. The activity-transport interface allows for such transformations, applying given parameters.

Time factors

It is common that the activity model is set on a different time scale as the transport model. In a regional application, production by sector usually refers to annual amounts, while the transport model will probably work on a daily basis. In an urban application, monthly units are probably the most convenient way to represent salaries rents and expenditure, and a peak hour simulation is required for the transport system.

To make time units compatible, the interface makes a distinction between two types of flows: *normal flows* and *habitual or commuter-type flows*. Commodity movements are typically represented as normal flows. For instance, the annual flow of Tons of agricultural products will generate a certain daily amount, defined with a time factor. Movements of people are typically represented as commuter-type flows. Flows between jobs and residents will generate trips that take place every day, even if the activity model is in monthly terms. Hence, for this type of flows time factors are not applied.

These transformations are made in the activities→transport direction. In the opposite the criteria is reversed. In the case of commodities, the transport model will come out with the cost of transporting, say, one Ton of produce. This cost will be the same, regardless of the time period involved. In the case of commuter flows, the activities model wants to know transport expenditure, that is, how much money was spent on commuting, so that this amount can be compared to other expenditures, such as services, land and floorspace. As a result, transport costs and disutilities are not multiplied by the time factors in the transport→activities direction for normal type flows, while commuter type flows are.

It is important to note that the transport model makes a distinction between transport costs and disutilities. As will be described later, the transport model applies an elastic trip generation model to transform flows by transport category into actual trips. In the case of transport costs, the emphasis is in expenditure. We want to know how much people or firms spend on transport. Consequently, the unit cost of transporting a good or passenger gets multiplied by the number of trips made during the *transport period*. Assume the typical home-work commuter from *a* to *b* spends $ 2 per trip, and makes two such trips, resulting in $ 4 per day. If (s)he makes 22 such trips per month, according to the time factor, then the total monthly expenditure will be 4*22=88.

By contrast, transport disutilities a) cannot be multiplied by trip generation rates, and b) it is irrelevant if they are or are not multiplied by the time factor. These considerations are dealt with in the transport model, but are worth pointing out to improve the understanding of the interface. We shall take each argument in turn.

Consider first argument a), that is, that transport disutilities cannot be multiplied by trip generation rates. Because trip generation is elastic, any improvement in the transport system will reduce transport disutility, and hence, induce more trips. If transport disutilities were multiplied by an increased number of trips made, the effect of the improvement could cancel out, or result in a larger value. Any transport improvement must reduce disutility. Because transport disutilities represent accessibilities and are an important component of the utility function in the allocation/interaction process, they must reflect any possible transport improvement adequately. In other words, if disutilities get multiplied by the number of trips made, then the activities model would get the wrong message.

In the case of b) the time factor is a constant that multiplies all elements of a given disutility matrix. Consequently, any improvement in the transport system resulting in a reduction of disutilities is not lost, and remains proportional. In Tranus, disutilities are multiplied by the time factors, but only to approximate them to the scale of costs.

The following table summarizes the application of time factors.

	Activities→Transport (flows)	Transport→Activities (costs and disutilities)
Normal flows	Divided by time factor	Time factor ignored
Commuter-type flows	Time factor ignored	Multiplied by time factor

Value-to-volume factors

The units in which activities are represented in the activity model do not necessarily have to correspond to those used to represent transport flows. For instance, manufacturing industry might be represented in money units or even employment units (jobs) in the activities model. In the transport model, however, the movements of industrial products are represented in physical units such as Tons. Similarly, population might be represented in terms of households, while the transport model work in terms of trip makers. The set of constants used to transform activities units into transport units are termed in general as *value-to-volume factors*.

In the activities→transport direction, flows by socioeconomic categories are multiplied by value-to-volume factors; in the transport→activities direction, costs and disutilities are divided by the value-to-volume factors.

Direction of flows

The activity model always generates economic flows from consumption zones to production zones, that is, in the direction that purchases take place. This, however, might not be the direction of the transport movements of people or goods. In an urban application, for example, residents are usually generated from employment. Thus, the resulting economic flows will be in the work→home direction. If the transport model is going to derive peak hour trips from these flows, it will be necessary to reverse the direction of the flows. If total day trips are being considered, then both directions will be relevant to represent return trips. These transformations are also dealt with by the interface module of Tranus.

Transformation of flows equation

All the transformations described in the preceding paragraphs may be represented in a single equation as follows:

$$F_{ij}^{s} = \sum_{n} \left(X_{ij}^{n} \frac{vol^{ns} pc^{ns}}{tiem^{ns}} + X_{ji}^{n} \frac{vol^{ns} cp^{ns}}{tiem^{ns}} \right), \qquad (26)$$

where:

F_{ij}^{s} flow of transport demand category *s* from origin *i* to destination *j*, in transport units;

X_{ij}^{n} production of the transportable economic sector *n* located in *j* and consumed in *i*;

vol^{ns} value-to-volume factor for the economic flow *n* that forms part of the transport category *s*;

$tiem^{ns}$ time factor for the economic flow *n* that forms part of the transport category *s*;

cp^{ns} proportion of the economic flow that moves in the direction consumption→production;

pc^{ns} proportion of the economic flow that moves in the direction production→consumption.

Summation is made over all economic flows *n* that form part of the transport category *s*.

Note that the cp^{ns} and pc^{ns} terms represent both the proportion in which each economic sector contributes to a transport category, as well as the direction of the flows. Consider an example in which three economic sectors: agriculture, mining and industry, give rise to three transport categories: general cargo, bulk and containers. In this case, we want all flows to be in the production→consumption direction, so that all $cp^{ns}=0$, and all pc^{ns} will have significant values. The following table shows possible values for the pc^{ns} terms:

Economic categories	Transport categories		
	General cargo	Bulk	Containers
Agriculture	0.3	0.6	0.1
Mining	0.2	0.8	0.0
Industry	0.4	0.3	0.3

In this example, 30% of agricultural produce travels in the form of general cargo, 60% as bulk and 10% in containers. There are no restrictions on the proportions, although it is common that $\sum_{s} pc^{ns} = 1.$

It is essential that every transportable economic sector forms part, at least, of one transport category, however small the proportion. This is because all transportable sectors need matrices of transport costs and disutilities. If a transportable sector has not been assigned to any transport category, the interface will not know how to build such matrices.

The are no restrictions on the proportions that determine the direction of flows. For example, in an urban application, residents might generate service employment. If the transport model is being applied to the morning peak hour, then the direction of the economic flows (residents→services) is probably the same as the corresponding transport flows (home→services). In this case:

$$cp^{ns} = 1 \text{ and } pc^{ns} = 0.$$

In the same application, jobs might generate residents, but for the morning peak we are interested in the home→work direction, that is, in the production→consumption direction. In this case:

$$cp^{ns} = 0 \text{ and } pc^{ns} = 1.$$

Assume the transport model is representing a two-hour morning peak period, and that we know that 90% of workers are going from home to work, and 10% go in the opposite direction. In this case:

$$cp^{ns} = 0.1 \text{ and } pc^{ns} = 0.9$$

If a two-way directionality is needed, as for total day trips, then:

$$cp^{ns} = 1 \text{ and } pc^{ns} = 1.$$

The transport model

Basic concepts

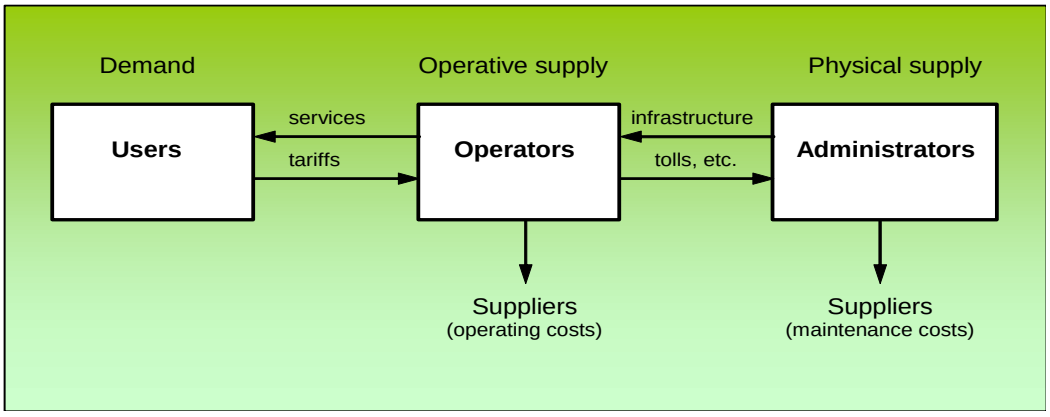
The main purpose of the transport model is to estimate travel demand and assign it to transport supply, such that an equilibrium is reached. As described in the previous sections, the activities model estimates a set matrices of flows by economic sector. The interface transforms such matrices into flows by transport categories. The transport model takes these flows as inputs and estimates the number of actual trips made in the given transport period. Trips are then assigned to supply, represented by the various modes and infrastructure. Transport costs and disutilities resulting from demand/supply equilibrium are used by the activities model to simulate a subsequent time period.

The two main elements of the transport model are, then, supply and demand. Users represent demand, that is people or goods that require a transport service for freight or passengers. On the supply side, it is possible to distinguish between *operative supply* and *physical supply*. The operative supply is the set of private or public organizations that operate vehicles of several types. Physical supply is the transport infrastructure required by the operators to perform their functions. An *administrator* is usually in charge of the physical supply. Users, operators and administrators are the main entities in the transport system.

Figure 5 shows these entities of the transport system and their main economic relationships. Users demand a transport service from the operators, and for this they pay tariffs. Operators charge users, and in turn pay for operating costs and for the use of the infrastructure. Administrators, in turn, charge operators and pay their suppliers for maintenance costs.

This is the general scheme, but in practice it may adopt specific forms. Private transport is a special case in which users are their own operators. It is also common that railways companies both operate the service and administrate their own infrastructure. The transport model, however, will always consider them as separate entities for accounting purposes. Also payments may be direct or indirect. Passengers, for instance, may pay directly to a bus operator in the form of fares, but car users indirectly pay themselves to cover their costs. Operators may have to pay administrators directly in the form of tolls, port tariffs, parking charges, and so on, but there might be indirect payments, such as fuel taxes, road taxes, and the like. There are also subsidies to consider, such as reduced fares for children and elders, or free roads and parking.

Figure 5: Elements of the transport system



Transport demand

Users are classified by transport categories. This allows for a separate treatment of passengers and freight. Passengers, in turn, may be classified by income group, trip purpose, or combinations of both. Each person category may have an associated car availability that limits the selection between public and private modes of transport. There may be several types of freight demand categories, such as bulk or containers.

From each travel option, users perceive a disutility, that includes the monetary cost of travel, the value of travel time and waiting time, and subjective elements, such as comfort, reliability, safety, and so on.

Operative supply

Transport supply is organized hierarchically in three levels: modes, operators and routes. Modes represent a set of operators that provide a service for a particular kind of user. In TRANUS, modes represent broad categories such as public, private, light or heavy commodities. Each trip category may choose among specific modes, such that commodities can only choose from freight modes and people can only choose from passenger modes. For each transport category a choice set must be defined, consisting of a list of modes available to that particular category.

Each mode, in turn, may consist of several operators. An operator is characterized by a service of some general characteristics, such as vehicle type, tariffs, operating costs, transfer costs, energy consumption, and so on. Many operators may belong to a common mode. A user can freely transfer from one operator to another, provided such operators belong to the same mode. A typical application may consider three modes: freight, public and private. The freight mode may contain light and heavy trucks, railways, barges, and so on, as operators, such that a consignment may combine them to reach its destination. The private mode usually considers a single operator: cars, but parking or HOV cars may be included as well. Public transport probably provides the most complex structures, and a large number of different operators may be specified, such as buses, minibuses, light and heavy subways, feeder buses, jitneys, walk, byke, etc. Certain rules may be imposed on transfers; integrated fares is one example, and some transfers may be prohibited altogether, even if the source and destination operator belong to the same mode. For instance, a transfer between a normal car and a HOV car may be prohibited.

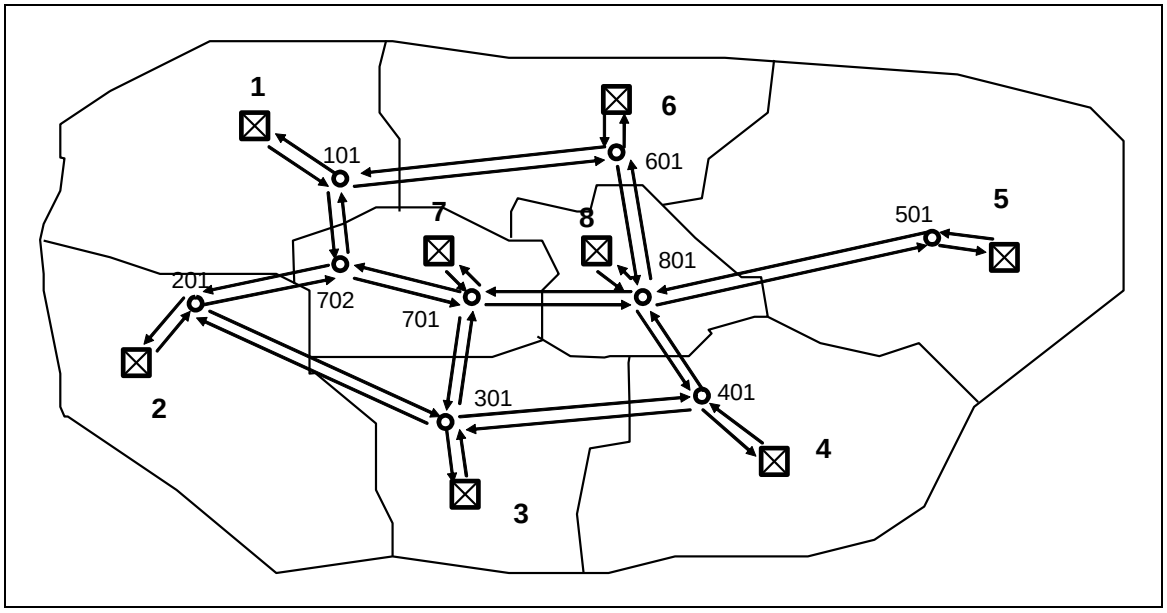
At a more detailed level, a public transport operator may be organized as a set of routes. The definition of a route is a service that must follow a specific sequence of links in the network. If a user wants to transfer from one route to another, he or she will have to pay a transfer cost and waiting time, except for possible integrated fares.

Physical supply

A *transport network* represents physical supply in the model. The network is defined as a directed graph, or a set of one-way links and nodes, like the example of Figure 6. Nodes may represent road junctions, points at which the characteristics of a road may change, stations, bust stops, ports, and the like. A subset of nodes are called *centroids*. Centroids represent zones, and for the transport model all trips either start or finish at centroids. Each centroid must be connected to one or more nodes of the network. Links, on the other hand, may represent street sections, highways, rural roads, railways, airways, waterways or any other kind of relevant infrastructure. Links have specific characteristics, such as distance, capacity, speed, and so on. Some of these characteristics are link-specific (distance, capacity), but others are defined generically, in the form of *link types*. All links of the same type share common characteristics (speeds, toll charges, maintenance costs, etc.). Link types also define which operators can use them; for instance, a highway link type may be used by trucks, cars and buses, but not by trains or vessels.

In the TRANUS graphical interface, a link may contain a number of intermediate nodes to form a *polyline*. See the TUS Manual for more details. These intermediate *polynodes* only affect the calculation of length from geographical coordinates, and do not enter in the numerical calculations. Links in the analytical network may only have one origin and one destination node.

Figure 6: Example of a transport network



The following characteristics define a link:

origin node
destination node
link type
link length
physical capacity
transit routes
prohibited turns or turn delays

The link type attribute specifies the following generic characteristics:

free flow speed by operator
car-equivalent units by operator
distance-related operating cost of vehicles per operator
charges or tolls by operator
administrator in charge
fixed and marginal operating costs
capacity restriction function

Usually, the physical capacity of a link is measured in car-equivalent units per hour or daily. In special cases other units may be used for convenience, such as trains, coaches, Tons, or any other. In very dense parts of a city, the capacity of a link may be determined by the capacity of intersections.

Transit routes are coded directly in each link. Each route has a specific frequency (vehicles per time unit). The model assigns the corresponding vehicles to the network. It is also possible to specify transit routes with an undefined frequency, in which case the model will adjust the frequency to demand.

Prohibited turns may be coded for each link, to indicate nodes towards which vehicles cannot turn. This a simple way of coding turn prohibitions that minimizes the possibility of errors and avoids the need for fictitious nodes

and links. Turn delays may also be coded, to represent signals, left turn conflicts, and other delays such as load and unload, access to ports, tolls stations and so on.

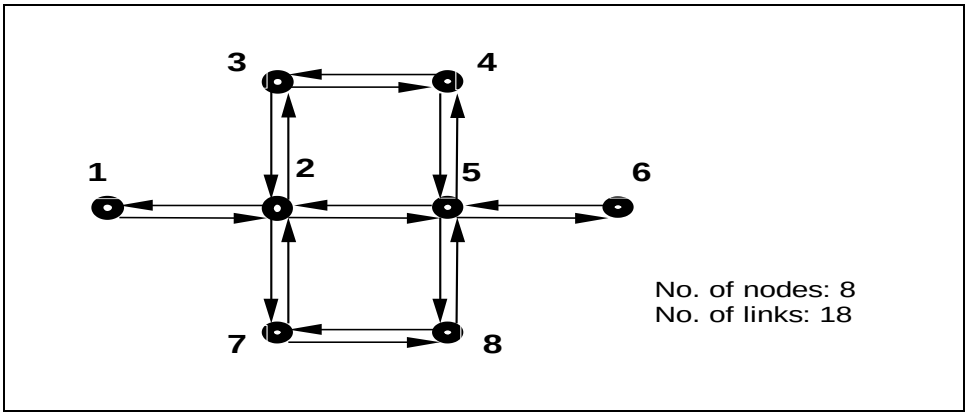
Representation of the transport network

The transport model represents the transport network internally as a *dual graph*. The network is coded in the traditional way, with links representing road sections and nodes representing intersections. Then the program automatically turns the external links into nodes of an internal dual graph, and creates internal links to represent possible connections. This method is completely transparent to the model user, because once all calculations have been performed, the model translates back the results in the original form. In other words, the modeler does not have to do anything, but it is important to bare in mind the concept behind the dual graph technique and its consequences for coding the network and interpreting the results. This section presents a brief conceptual description of the dual graph technique. A full description and mathematical specification is given in 'Dual Graph Representation of Transport Networks', by J Añez T de la Barra and B Pérez, *Transportation Research B,* Vol 30, pp 209-216, 1996. This paper is also available from Modelistica's home page.

Figure 7 shows an example of a simple network, consisting of eight nodes and 18 links. As in most transport models, this network is coded in the form of a simple list of links as follows:

Origin node	Destination node
1	2
2	1
2	3
3	2
2	5
…etc.	

Figure 7: Example of a network with the original direct representation



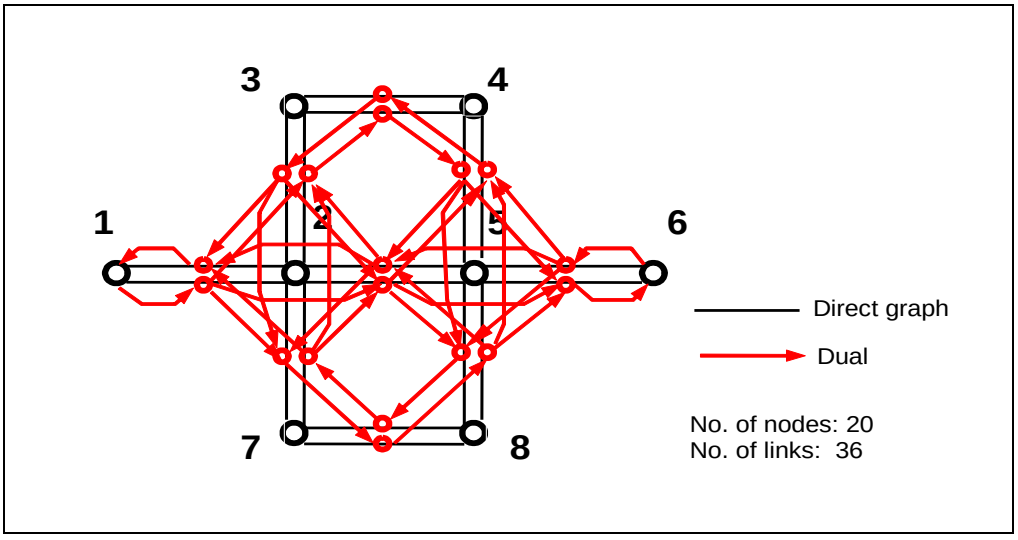
From this list, the dual graph is generated by the model. In the dual, each original link becomes a node or vertex of the graph, and links represent possible connections. Figure 8 shows the resulting internal dual graph created by

the model. For instance, the original link 1-2 is connected to links 2-3, 2-5 and 2-7 in the original network. In the dual network, link 1-2 becomes a node, and the resulting list of links would look something like:

Origin node	Destination node
1-2	2-3
1-2	2-5
1-2	2-7
2-3	3-4
…etc.	

Because this dual list is easy to automate, the model builds it internally. The total list of nodes will now be made of 20 dual nodes and 36 dual links.

Figure 8: The dual graph created by the model



The main advantage of this scheme is that it is very easy to define prohibited turns and turns delays. The difficulty of defining them in the direct graph is well known. This is usually done by replacing each node by several nodes, and then the modeler must code a number of fictitious links to represent turning movements. This process is tedious and error prone, and results in a large number of unwanted fictitious nodes and links. Some models automate this process, but the computational burden can be significant, and the results are not reliable.

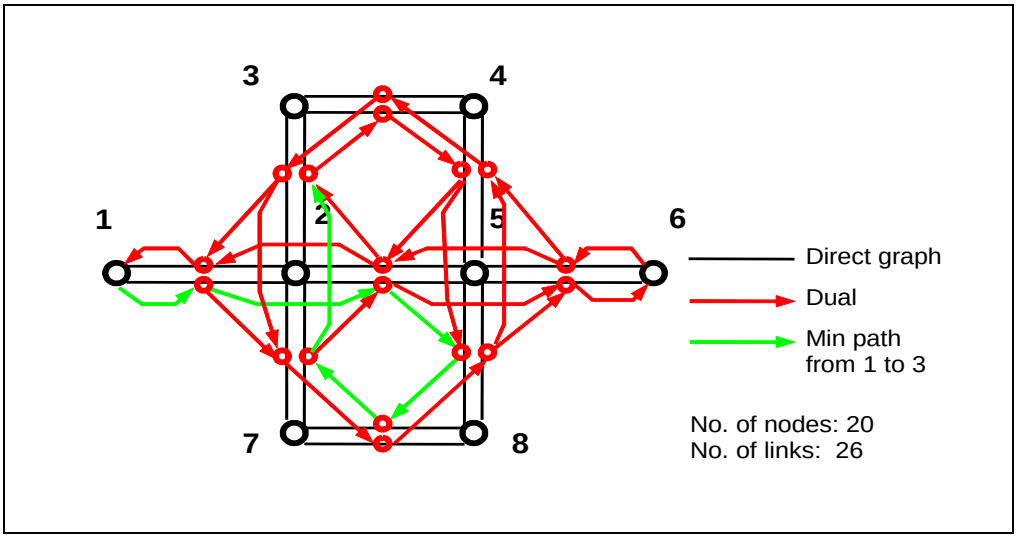
The dual graph representation complete avoids fictitious links. All the modeler has to do is to specify, for each direct link, the nodes towards which vehicles cannot turn, if any. Assume that all left-turns are prohibited in the simple network above. The network code thus becomes:

Origin node	Destination node	Prohibited turns
1	2	3
2	1	
…etc.	…	
2	5	4
5	2	7
…etc.		

When the model builds the dual graph, it checks for possible prohibited turns. A dual link is created only if the destination node of the link to which it is connected is **not** in the list of prohibited turns. The more prohibited turns are defined, the smaller the generated network; contrary of most traditional models.

Figure 9 shows the resulting dual graph when all left turns are banned. It may be seen that the dual now contains fewer links, 26 instead of 36. In fact, as more turn prohibitions are indicated, the size of the network becomes smaller. In traditional models the opposite happens. The figure also shows the resulting minimum path from 1 to 3 in the dual.

Figure 9: The dual graph created by the model with left turn prohibitions



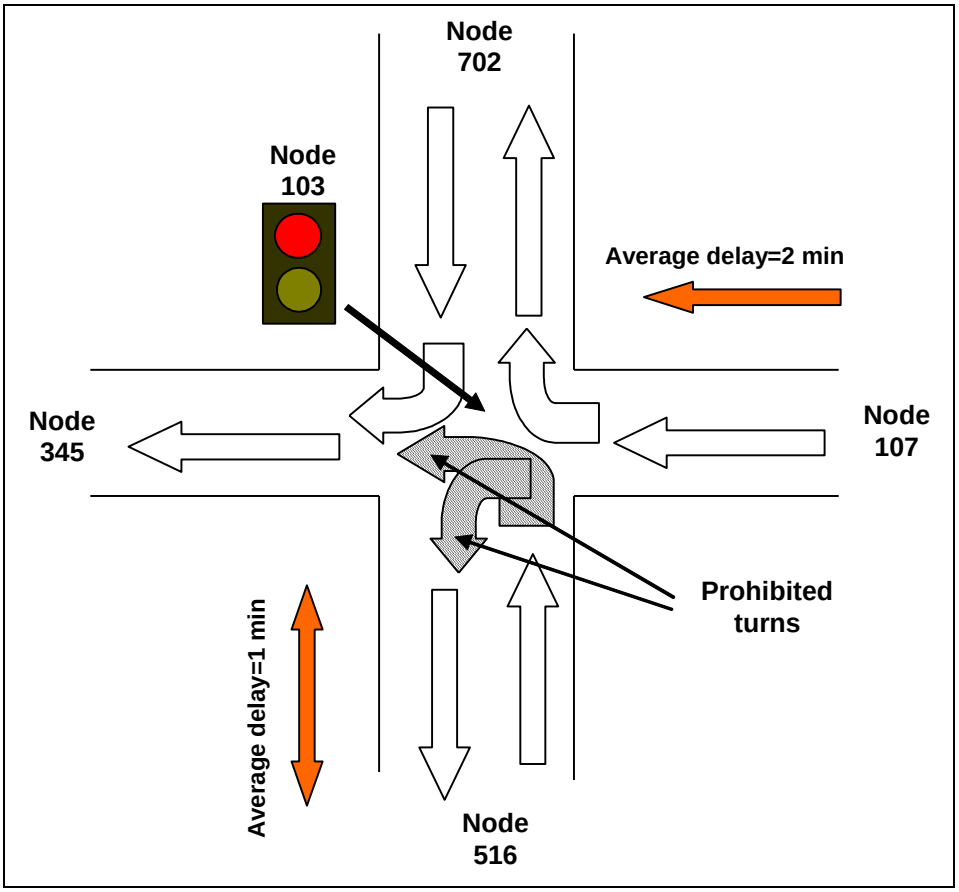
Turn Delays

Optionally, it is possible to specify turn delays, which is useful for representing traffic signals and many others, such as loading-unloading for freight trips, tolls stations, access to ports and so on. The input data is the average delay, defined in a similar way that turn prohibitions. Figure 10 shows a signalized intersection 103. The average

delay for the East-West road is two minutes (0.033 Hr) and one minute (0.017 Hr) for the North-South road. Left turns are prohibited.

The table under the figure shows permitted turns at intersection 103 and the corresponding delays. For instance, node 516 can not be accessed from link 107-103 (infinite delay=turn prohibition). Turning to 702 imply a delay of two minutes. Delays are added to travel time and have effect on all vehicles (cars, buses, trucks) assigned to the multimodal network of TRANUS, which is described in the next section.

Figure 10: Turn Delays



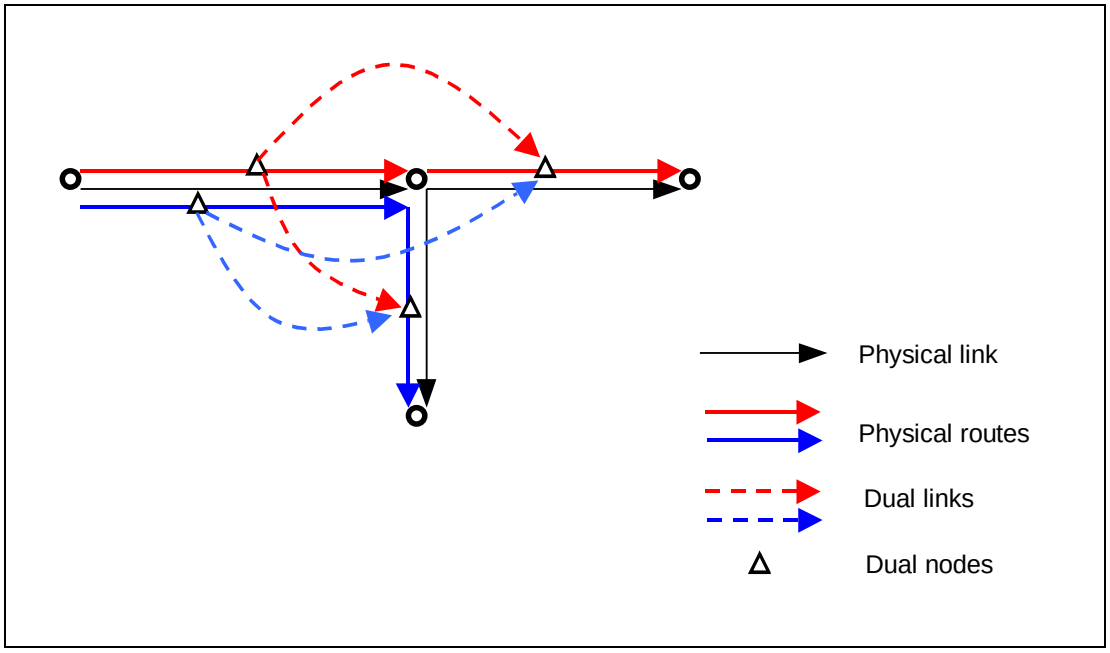
Origin Node	Destination Node	Turn to node	Delay (hours)
107	103	702	0.033
107	103	345	0.033
107	103	516	Infinita
516	103	702	0.016
516	103	345	Infinita
702	103	516	0.016
702	103	345	0.016

Multimodal Network

The dual network concept is extended to the *multimodal* network. Depending on the link type, several modes, operators or routes may use a common link. For example, a typical road may allow for cars, several bus routes, minibus routes, trucks, bikes and pedestrians. Each physical link is expanded by the model, such that each link/operator-route combination becomes a link. Again this is an internal process, transparent to the modeler.

Figure 11 shows a simple example of a multimodal network. There are three physical links and two bus routes. The model generates one multimodal link for each combination of physical link and route. In the dual network, each multimodal link becomes a node, and dual links are generated for every possible combination. If the origin and destination routes are **not** the same, such as a transfer from the red route to the blue route, transfer costs are added. Transfer costs may include monetary costs (boarding tariff) and the value of waiting time.

Figure 11: Generation of dual links in a multimodal network



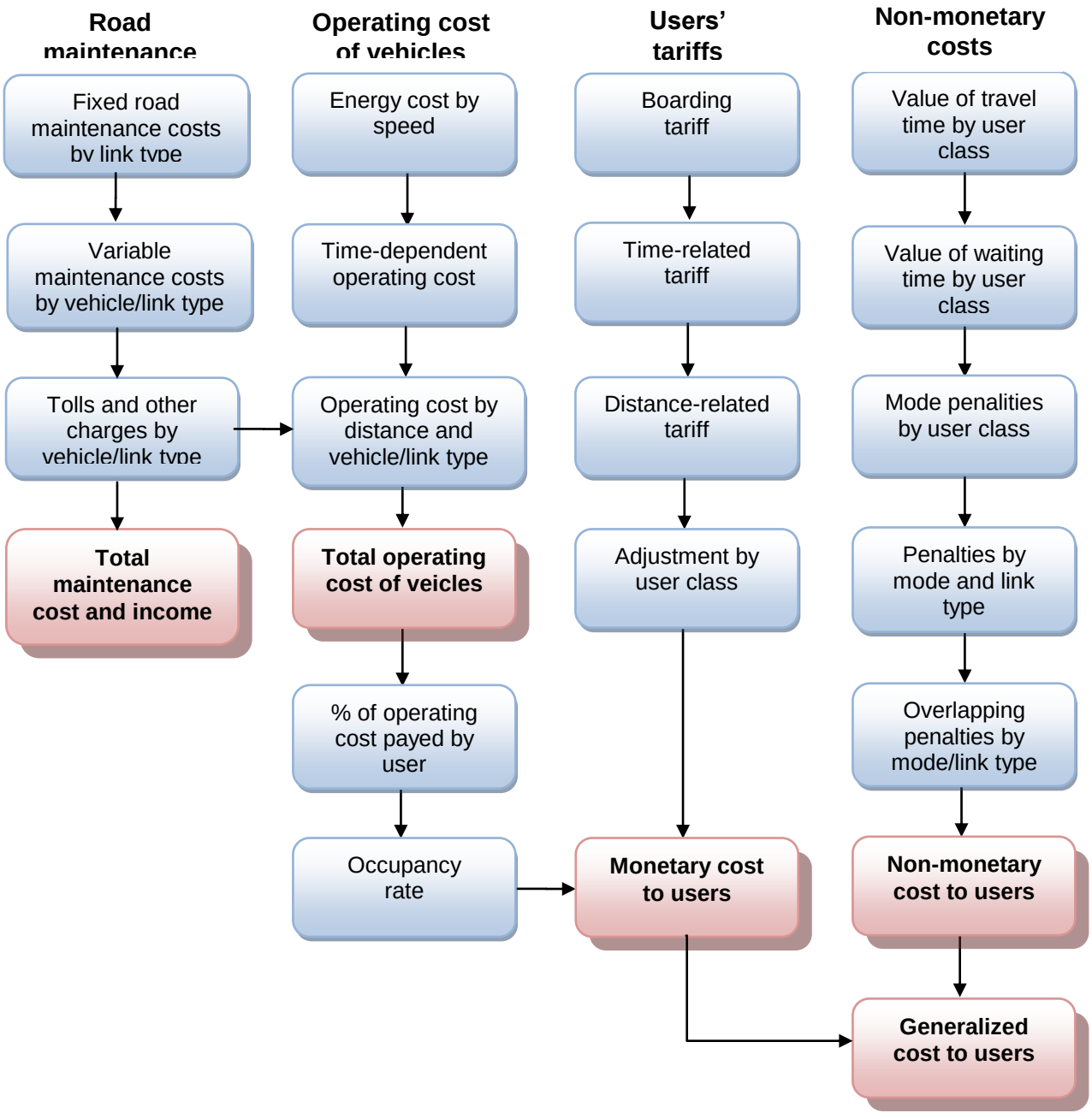
Structure of transport costs

As described in Figure 5, the transport model makes a distinction between three types of monetary costs, corresponding to the three main entities in the transport system: users, operators and administrators, as follows:

- Users' costs, including monetary and non-monetary costs, adding up to what is called *generalized costs*. They are measured by demand unit being transported (passengers, Tons, etc.)
- Operating costs, strictly monetary, in terms of vehicles
- Infrastructure maintenance costs, also monetary, in terms of distance

The first two items may affect demand and the assignment of demand to supply, while maintenance costs will not affect demand but are important to economic accounting. Figure 12 shows in detail all cost elements that are calculated in the transport model. The following paragraphs will describe each component.

Figure 12: Detailed calculation of transport costs in TRANUS



Monetary costs to users

The monetary component of user costs is termed *tariff*. It is the amount of money that users pay to the operators. Two broad types of tariffs may be specified, depending on whether they are related or independent of operating costs. It is common that public transport services operate with tariffs that are quite independent of their operating costs. Freight operators, by contrast, tend to link tariffs with their actual operating costs, (including a profit margin). In the case of private cars, the user and operator are the same, so that tariffs should strictly correspond to operating costs; however, there is considerable evidence that car users do not perceive their full costs, but only a small proportion. For convenience, the term tariff refers to monetary costs to users, even if in the case of cars it might sound inadequate.

Furthermore, not all demand categories pay the same tariffs. It is common that students and senior citizens pay preferential fares. It could also be that some types of freight operators may charge different fares depending on the commodity being transported. All these elements are taken into account by the model when calculating tariffs.

In general, the tariff charged by operators to users is calculated as:

$$t_o^s = tp_o^s\left(tf_o + tt_o + td_o + \frac{tc_o c_o}{to_o}\right), \qquad (27)$$

where:

t_o^s is the tariff paid by trip makers of type *s* to operator *o*;

tp_o^s is the proportion of the *full tariff* that trip makers of type *s* pay to operator *o*; the rest of the elements of the equation above represents full tariff;

tf_o fixed tariff when operator *o* is boarded; if there are integrated tariffs, then tf_o will depend on the previous operator;

tt_o time-related tariff of operator *o*, calculated as the time that the operator takes to perform the trip, multiplied by a tariff per unit of time;

td_o distance-related tariff of operator *o*, calculated as the distance travelled, multiplied by a tariff per unit of distance;

$\frac{tc_o c_o}{to_o}$ the operating cost of operator *o*, c_o, multiplied by a factor tc_o representing the proportion of the operating cost to be added to the tariff, and divided by the occupancy rate of operator o, to_o.

The operating cost of vehicles is represented in terms of vehicles, such as per car or per truck. Dividing by the occupancy rate implies that the operating cost is being shared by all users in the vehicle. In the case of private cars, for instance, if the occupancy rate is 1.4, then each passenger will pay 1/1.4 of the cost. Similarly, if a truck carries 8 Tons on average, then each Ton will pay 1/8th of the operating cost.

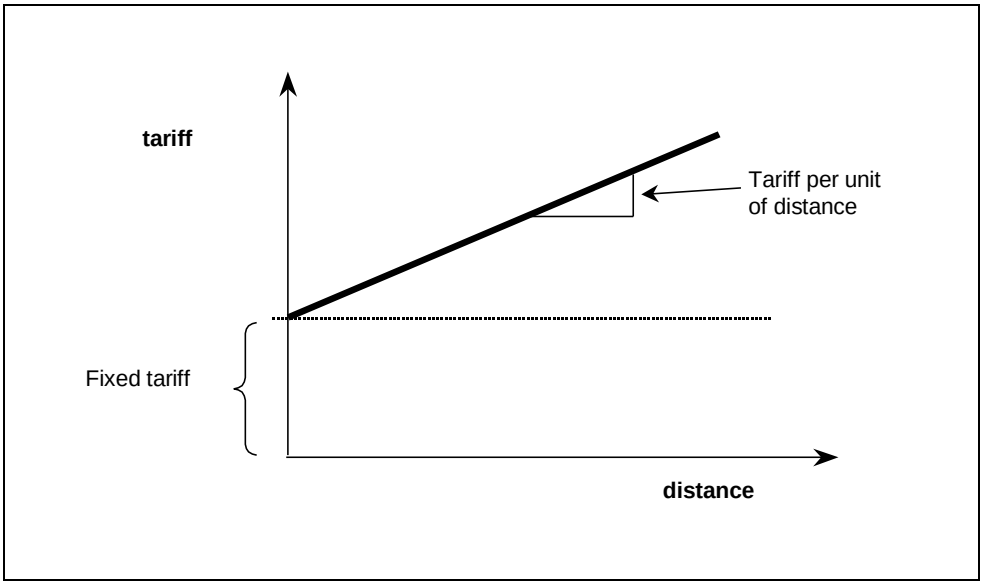
Different tariff functions may be specified for each operator. In the case of a freight service, it is common to specify tariffs as a function of operating costs; hence only tc_o will have a significant value (e.g. 1.2), leaving the

rest of the elements as zero. In the case of private cars, it is common that all elements are set to zero, except tc_o that is set to 0.5 or 0.6 or whatever value is considered as representative of the proportion of operating costs perceived by users. In the case of transit, it is common to have tariffs as a function of distance and/or the fixed component, setting tc_o to zero. In general, tariffs either have a specific function related to distance and time, or are related to operating costs. A combination of both is very rare.

A typical function for transit operators is presented in Figure 13. In this example there is a fixed component and a distance-related tariff. In the case of taxis, it is common to have a fixed component, together with both distance and a time-related components.

Note that congestion will affect tariffs only if there are time or operating cost-related components. For operators of this kind, congestion will increase the tariffs paid by users. This may occur if a specific time-related tariff has been defined, or if there is an energy function and/or a time-related operating cost, and the operator transfers the cost to the user.

Figure 13: Typical fixed+distance-related tariff



Operating costs

Operating costs per vehicle, c_o of a particular operator o includes a fixed element, a distance-related element, a time-related element and possible charges. The latter represents the amount that operators pay to administrators, representing elements such as tolls, parking charges, port tariffs and others. These elements are represented in the following expression:

$$c_o = cf_o + ct_o + cd_o + ch_o + ce_o, \quad (28)$$

where:

cf_o	fixed operating cost of a vehicle of operator *o* to be applied only when the vehicle is boarded, that is, once for every trip made; usually refers to administrative costs and loading/unloading in the case of goods vehicles;
ct_o	operating cost per unit of time; usually includes drivers' salaries and capital payments;
cd_o	operating cost per unit distance of a vehicle of operator *o*, usually including tires, spares, maintenance, lubricants, and others; this cost varies by link type;
ch_o	charges paid by operator *o* to administrators, usually representing tolls, parking, duties, etc.;
ce_o	energy cost of operator o, a function of distance and speed.

The cost of energy is calculated as:

$$ce_o = \left[ed_o^{\min} + \left(ed_o^{\max} - ed_o^{\min}\right) * \exp\left(-\delta^o V_o\right)\right] pe_o, \quad (29)$$

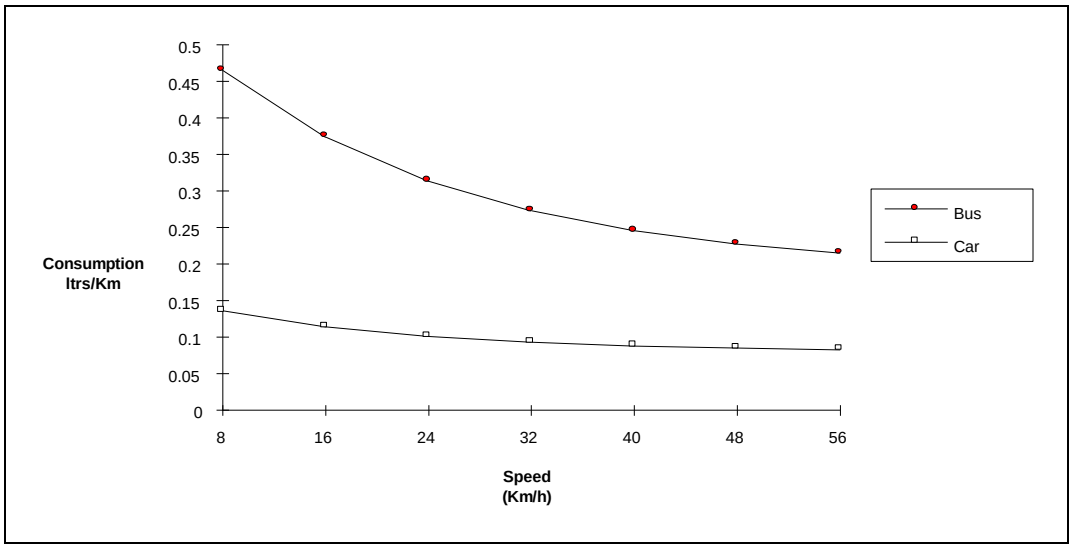
where:

ce_o	energy-related cost per unit distance of a vehicle of operator *o;*
$ed_o^{\min}$	minimum consumption of energy per unit distance when a vehicle of operator *o* travels at free flow speed;
$ed_o^{\max}$	maximum consumption of energy per unit distance when a vehicle of operator *o* travels at a speed close to zero;
V_o	speed of vehicle of operator *o*, after capacity restriction;
δ^o	parameter regulating the steepness of the energy consumption curve;
pe_o	price of a unit of energy.

This calculation is made on a link-by-link basis, because the speed is link-specific. The function is negative, because as speed increases, energy consumption is reduced. Figure 14 shows examples of typical energy consumption curves for two different operators.

The cost of energy may be calculated in different units for each operator. For example, cars may be calculated in terms of liters of gasoline, truck in liters of diesel, and rail in KWh. If the energy function is not known, the corresponding cost may be specified as a distance-related cost.

Figure 14: Examples of energy consumption curves



Maintenance costs

Administrators pay for fixed and marginal maintenance costs per unit distance. Fixed costs include routine maintenance and any other, assuming that no vehicles use the infrastructure. Marginal costs represent the maintenance costs attributable to each additional vehicle traveling along the link per unit distance. Each administrator *a* is in charge of a particular set of link types T^a . If *lt* is a link of the set L^τ of links of type τ, the cost of maintenance of administrator *a* is:

$$cm^a = \sum_{\tau \in T^a} \sum_{lt \in L^t} \left[mf_\tau * d_l^\tau + \sum_o ma_\tau^o * Ve_l^o \right], \qquad (30)$$

where:

mf_τ	fixed maintenance cost per unit distance of links type τ
d_l^τ	distance of link *l* of type τ
ma_τ^o	marginal cost of maintenance of links type τ per vehicle of operator *o*
Ve_l^o	number of vehicles of operator *o* traveling along link *l*

Structure of the transport model

The transport model follows a calculating sequence as described in Figure 15. Two distinct procedures may be distinguished: path building and the transport model itself. Path building may be seen as the process of generating

travel options for each origin and destination, by mode. Based on the description of the network and parameters and functions describing costs and disutilities, the path building module generates a set of such options. Instead of generating a single minimum path, this procedure will generate the first *n* paths.

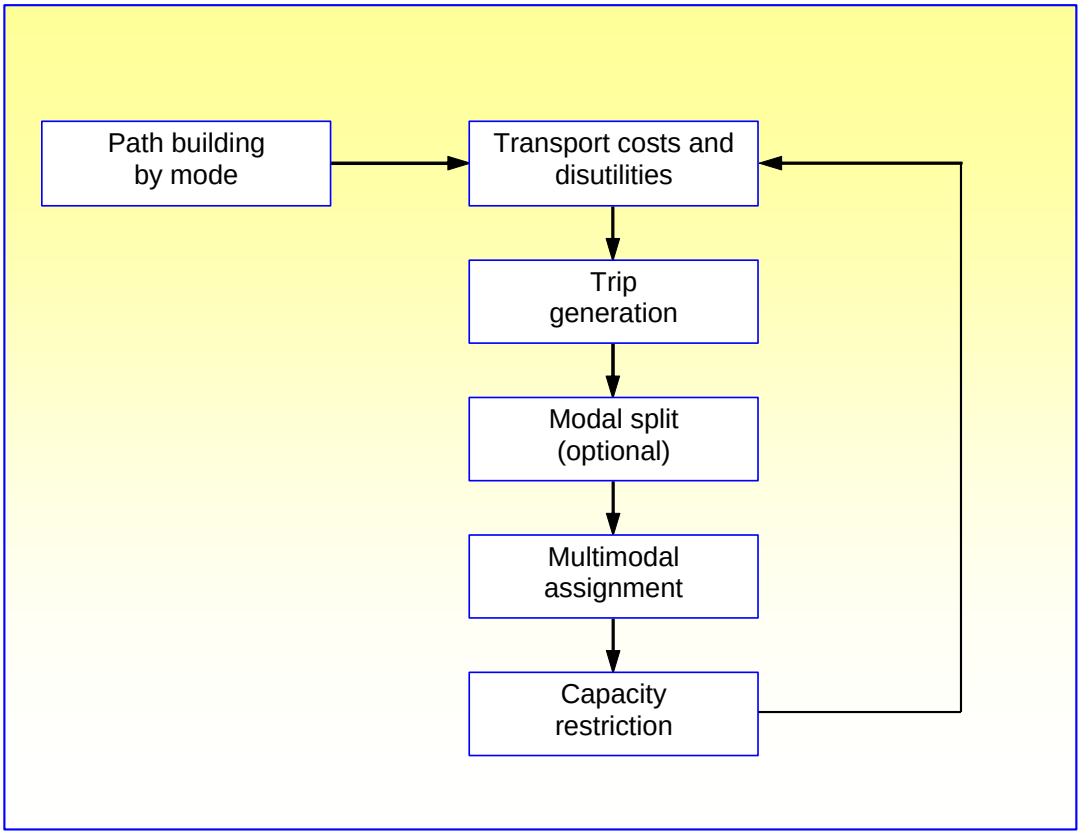
The transport model takes the path descriptions and performs a number of calculations. It begins by calculating the costs and disutilities involved in each path for each transport category. These costs are aggregated over all paths to generate costs and disutilities by mode, and aggregated again to obtain costs and disutilities by transport category.

The next stage is to estimate the number of trips generated by O-D pair and category. This is an elastic function of the flows and disutilities. Trips are then separated by mode, and are then assigned to paths and links of the network. The modal split stage is optional, since it may be combined entirely with the assignment stage. This is a powerful feature that will be explained in detail further below. Finally a capacity restriction procedure is applied to adjust speeds and waiting times according to the demand/capacity ratios in each link and route. Because speeds and waiting times affect costs and disutilities, an iterative procedure is performed. The process is repeated several times, until a demand/supply equilibrium is reached.

In the first iteration, costs and disutilities are calculated on the basis of free-flow speeds and minimum waiting times. At the end of the iterative process, speeds and waiting times may have changed considerably, as the network is loaded with trips. These are termed *loaded* costs and disutilities.

The following sections describe each component.

Figure 15: Structure of the transport model



Path building

As was mentioned, the purpose of the path building procedure is to derive a set of travel options from an origin to a destination by a particular mode. A path is not just a sequence of links, but a sequence of link/operators (or routes) combinations. There might be two paths with identical physical links, but with different operators or transit routes.

Formally, a path may be described as:

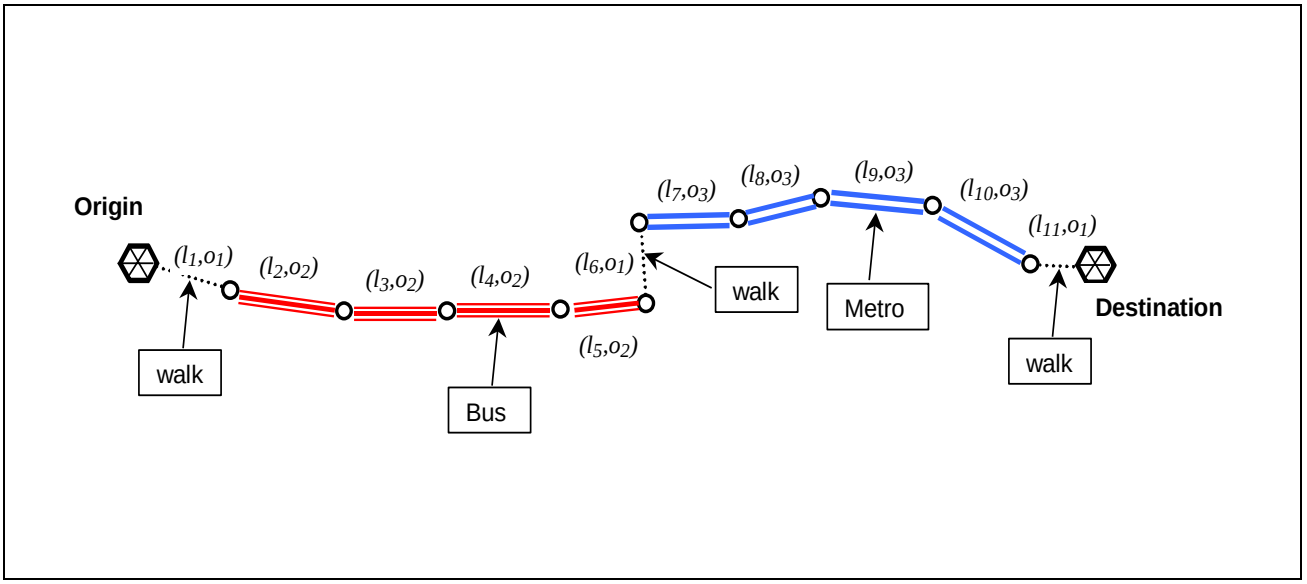
$$m_1, m_2, m_3, \ldots m_z$$

$$mi = (li, oi)$$

where m_i denotes a particular combination of a physical link l_i and an operator (or route) o_i along the path sequence. The origin node of l_1 must be the centroid of the origin zone, where operator o_1 is boarded. Along the path sequence, a change from one operator to another may occur, and this introduces the possibility of transfers. Finally, the destination node of l_z corresponds to the centroid of the destination zone of the trip.

Consider the example of a multimodal path in Figure 16. this travel option involves 11 physical links and three operators (walk, bus, and metro). Trip makers walk to the bus stop, then take a bus, then walk to a metro station, then take a metro and finally walk to the destination.

Figure 16: Example of a multimodal path



During path search, the model calculates the *generalized cost* for each path, accumulating the following elements for each link/operator combination *m* that forms part of the sequence:

$$c_{ijp}^{ks} = \sum_{m=1}^{z} RT_m^s + RD_m^s + TR_{m-1,m}^s , \qquad (31)$$

where:

c_{ijp}^{ks}	generalized cost of path *p* from *i* to *j* by mode *k* for the category *s;*
RT_m^s	time-related costs in link combination *m*(*l*,*o)* for demand category *s*;
RD_m^s	distance-related costs in link combination *m(l,o)* for demand category *s*;
$TR_{m-1,m}^s$	transfer costs, that is, the cost of boarding a new operator or route; this can take place either at the beginning of a trip or when there is a transfer somewhere along the path, i.e., when *o(m-1)≠o(m).*

In the following paragraphs each one of these cost elements are described. Most of the elements used to calculate these costs have already been defined when describing operating costs and tariffs.

Time-related costs

Includes two main components: monetary and perceived, as follows:

$$RT_m^s = tv_m\left(tt_o + \frac{ct_o tc_o}{to_o}\right)pc_o^s + tv_m\left(vv^s\, pt_m\, pg_o\, pp_o^s\right), \qquad o \in m \quad (32)$$

where:

RT_m^s	time-related cost for the link/operator combination *m* as perceived by trip-makers *s*.
tv_m	travel time of operator *o* in link *l*, a function of the length of the link and the speed of the operator;
tt_o	time-related fare charged by operator *o;*
ct_o	time-related operating cost of operator *o;*
tc_o	proportion of the operating cost that operator *o* transfers to users;
to_o	occupancy rate of operator *o*;
pc_o^s	proportion of fare paid by users of category *s* to operator *o*;
vv^s	value of travel time of the trip category *s*;
pt_m	penalizing factor associated to the link type of the combination *m(l,o);*
pg_o	penalizing factor associated to operator *o*;
pp_o^s	penalizing factor associated with the operator *o* and travel demand category *s* combination.

The first part of the equation represents the monetary components: a possible time-related tariff (tt_o), and time-related operating costs being transferred to users (ct_o tc_o/ to_o). The monetary component is, in turn, multiplied by the proportion that each category pays to the operator; this is because some categories, such as school trips or the elderly, may pay reduced fares. The second part of the equation represents the non-monetary, subjective or perceived part of the generalized cost. Includes the value of time, multiplied by three penalizing factors associated with the link type, the operator and the travel demand category.

It is important to note that the value of time is only related to the trip-making category. All other aspects related to the subjective perception of time are taken care of by the penalizing factors (by operator/category, or by link type).

The factor associated with the link type is used to represent non-modeled factors related to the quality of the infrastructure, such as convenience, safety, road-side services, and so on. Consider the example of an origin connected to a destination by two alternative roads: a small curvy road and a large highway. Assume that travel times and costs are the same in both cases, but the highway has better signals, the possibility of accidents is smaller, has lighting, emergency facilities and good road-side services. In such conditions users will prefer the highway, and this is represented in the model with a smaller penalizing factor.

Similarly, the penalizing factor associated with operators is used to represent non-modeled elements such as reliability, safety, comfort, and so. In practice, a value of 1.0 is usually assigned to the penalizing factor of the best option, and bigger values for the less attractive options, such as 1.1, 1.3, etc.

Finally, the operator-travel demand category penalizing factor is used to represent preferences. It may be that a specific category prefers a specific mode, resulting in a reduced penalty for such combination. A typical example of this is a bulk commodity preferring rail over trucks. In the case of passengers, this factor may be used to represent preferences by income group. High income travelers may show a strong preference for cars, while low income travelers might show a preference for transit modes. This does not mean that low-income travelers dislike the car option in principle; however, the choice of cars could mean that the consumption of other goods may have to be sacrificed if the corresponding household must incur in the high cost of buying a car. This is a very important feature, because in this way the modal split stage in the model may be skipped altogether. In this case, modal split is entirely dealt with at the assignment stage, providing great flexibility and realism in the simulations. Car availability, instead of being an input parameter to the model, becomes an output and an endogenous variable.

It is recommended that the modal split stage is avoided in the model design. Instead, the use of operator-category penalty factors is encouraged. Modal split has been maintained in the model structure for compatibility with previous versions that did not include this powerful feature.

Distance-related costs

Distance-related costs to users include distance-related tariffs, and distance-related operating costs transferred to users:

$$RD_m^s = d_l\left(td_o + \frac{cd_m tc_o}{to_o}\right)pc_o^s,\ l,o \in m \quad , \qquad (33)$$

where:

RD_m^s	Distance-related cost for the link/operator combination *m* and category *s*;
d_l	length of link *l;*
td_o	distance-related fare charged by operator *o;*
cd_m	distance-related operating cost of operator *o* in link *l;*
tc_o	proportion of the operating cost that operator *o* tranfers to users;
to_o	average occupancy rate of vehicles of operator *o*;
pc_o^s	proportion of fare paid by users of category *s* to operator *o*.

Transfer costs

Transfer costs are paid by the user when there is a transfer, that is, a change of operator or route somewhere along the path. These may include a monetary boarding fare, possible fixed operating costs transferred to the user, and a perceived value of waiting time. These costs are only computed if the current operator-route of the link/operator combination *m* is different to the operator-route of the previous combination *m-1* along the path, that is, if $o(m-1) \neq o(m)$.

$$TR_m^s = \left(tf_o + \frac{cf_o tc_o}{to_o} \right) pc_o^s + te_m ve^s \quad , o \in m \ , \qquad (34)$$

where:

TR_m^s	transfer cost for category *s* when boarding operator *o* at link *l*;
tf_o	boarding fare of operator *o*;
cf_o	fixed operating cost of operator *o;*
tc_o	proportion of the operating cost that operator *o* transfers to users;
to_o	average occupancy rate of vehicles of operator *o*;
pc_o^s	proportion of fare paid by users of category *s* to operator *o*;
te_m	waiting time for a vehicle of operator *o* in link *l* (applies only to transit);
ve^s	value of waiting time for trip category *s*

Note that the boarding fare tf_o is a matrix of the form *(from operator, to operator)*; this is to allow for integrated fares, in which case $tf_o=0$, reduced fares or to prohibit a specific combination with $tf_o=\infty$. An example of an integrated fare is the case of a metro system combined with a bus-feeder system. An example of a reduced fare is a port, in which both railways and trucks come to unload; if the railway has a highly automated and mechanized unloading system, the transfer cost from railway-to-ship could be less than the truck-to-ship cost. An example of

a prohibited combination is a case in which there are buses, walking and bicycles; the model specification may allow for people to walk to buses, but not bike to buses, if there are no special facilities to do so.

Prohibited combinations are also useful to represent park-and-ride situations. In this case a P&R operator may be defined and assigned to specific links in the network. Next, car options may be allowed to combine with P&R but not with buses, and in turn P&R may be combined with buses. This means that car-travelers may only combine with buses at specific P&R facilities.

Waiting time depends on two main factors: the frequency of the service involved, and the demand/capacity ratio. The frequency of the service defines a minimum waiting time. As the demand/capacity ratio increases, waiting time also increases in addition to the minimum. The way in which waiting time varies as a function of the demand/capacity ratio is dealt with at the capacity restriction stage, described further below. If the network is empty, as in the initial path search or the first iteration of the transport model, then waiting time is only a function of the frequency. Assuming a random arrival of passengers, the average waiting time is:

$$te_m = te\min_o + \frac{1}{2 f_o} \quad , o \in m \ , \qquad (35)$$

where:

te_m	waiting time when boarding a vehicle of the route or operator o in link l, assuming the demand/capacity ratio is very low at that point;
$temin_o$	minimum waiting time for vehicles of the operator o in addition to the frequency-related time;
f_o	is the frequency of the route or operator o (vehicles per unit of time).

Two types of transit operators are recognized in the model: scheduled and unscheduled. If the transit operator is scheduled, only the first part of the equation above applies, that is $te_o = temin_o$. This form is particularly useful when representing rural or inter-urban services with very low frequencies. For example, if a bus route only has one bus per day, it is obvious that passengers will not wait on average half a day, since they will probably know the time-table of the service. Tranus allows simulating transit services without specifying routes; in this case frequencies are calculated automatically in function of demand.

Additionally, Tranus allows defining a range for frequencies by operator. The model estimates the frequency that maximize the operator income. The resulting frequency is f_o

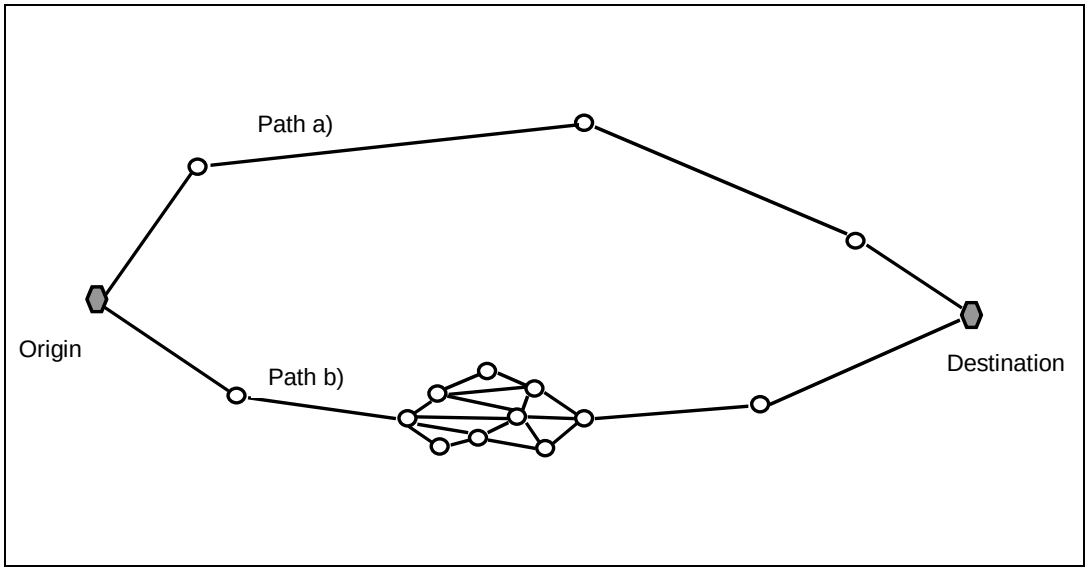
The path search algorithm

The costing elements described in the previous sections are the basis of path search. For each origin and destination, the path building procedure analyzes all possible link/operator or route combinations that may form a reasonable travel option. This may lead to a very large set of options, particularly if the network and the operators and routes make a densely connected system. In order to reduce the number of options, the path search procedure selects a smaller subset, that is, a set of n-paths. Two main criteria are used to select the set of n-paths:

- those of least generalized cost;
- those that make clear, distinct options.

The first criteria is quite obvious, but if it was the only one to be applied, some of the resulting paths could be very similar; some paths could share a large number of link/operator combinations. This would represent very close options, that travelers would consider as a single option with very small variations. There is a risk that small variations around a successful path may rule out some very reasonable options. Consider the example of Figure 17. At first sight there are two main paths a) and b); path b) offers a considerable number of close variations. Assume that path b) and all its variations have a smaller generalized cost compared to path a). If the generalized cost was the only criterion for selection, only path b) and six of its variations would get selected, completely neglecting path a).

Figure 17: Simple network to show overlapping effects



The path search algorithm includes a procedure to control for the independence of options, to avoid highly correlated paths with small variations as separate distinct options, called *overlapping control*. The method keeps track of the degree of coincidence among competing paths in terms of link/operators combinations. As a result, paths are selected as those with the least generalized cost and least overlapping.

Overlapping control is achieved through a penalizing factor called the *Oz* factor, a positive number greater than one. Bigger values of *Oz* increase the effect of overlapping in the selection of paths. With this factor, the path search model proceeds in the following steps:

a) search for the minimum path from *i* to *j* by mode *k* and store it;

b) penalize by *Oz* the link-related costs of all link/operators *m* that form part of the path, excluding transfer costs;

c) go back to step a) and iterate until:

d) the minimum path found in a) is identical to any of the paths that got stored.

If *Oz=1* the minimum path found in the first search will immediately re-emerge in the second search; in this case the model yields a single path for each O-D pair. As the value of *Oz* increases, the number of paths that succeed in getting stored also increases. Note that a same operator/link combination may emerge several times in different paths and is penalized over and over again. Also note that no paths will be stored with a generalized cost grater than the cost of the minimum path multiplied by *Oz*, so that this value also acts as a maximum dispersion factor.

Ideally, the procedure to select paths should be applied to each category and mode. Each category of travelers will value time differently, and so will consider different choice sets. However, to avoid the computational burden, paths are only selected by mode. Hence, the modeler must make sure that the path set is wide enough to accommodate all categories of transport users. When the path choice model is applied, as described further below, each category will be treated with its own values to simulate path choice.

Disutilities and probabilities

Transport disutilities are measures of accessibility that influence both location decisions and transport choices. Transport monetary costs form part of disutilities, and are used directly to calculate production costs in the activities model. The transport model keeps track of both disutilities and monetary costs separately.

As shown in Figure 18, three main levels of decisions are considered in the transport model: path choice, mode choice and trip making choice, that is the decision as to how many trips to make. For each one of these levels, the model estimates disutilities, and these in turn, form the basis for the calculation of probabilities. Probabilities are calculated with scaled multinomial logit models.

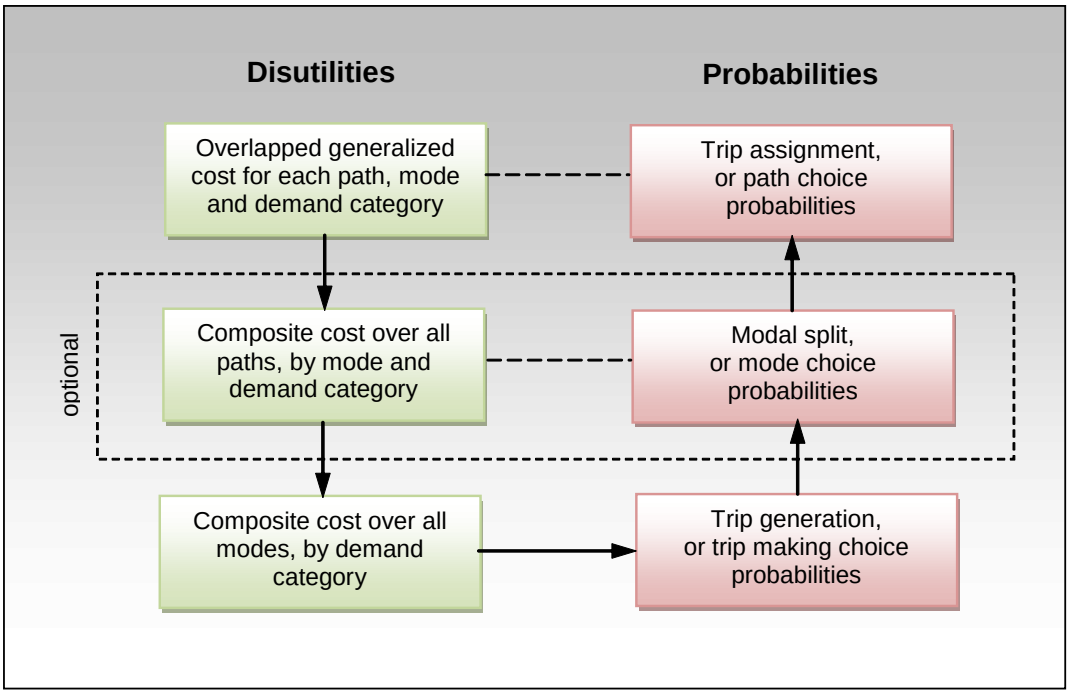
The arrows in the figure show the sequence in which these calculations are made. The process begins by calculating the disutility at a path choice level. Because this is the lowest level in the inverted decision tree, the utility function for path choice corresponds to the generalized cost of each path. The way in which generalized costs are calculated were described in the previous sections, and include monetary costs and perceived travel and waiting times. Generalized costs are then aggregated over all competing paths to form a composite cost by category and mode, which is the utility term at a mode choice level. Finally composite costs are aggregated over all competing modes to form a general composite cost for each transport demand category, which is the utility term at the trip making level.

Trip making choice is estimated by the trip generation model, mode choice is estimated by the modal split model, and path choice by the assignment model. These three models are linked to each other by the composite costs, forming a large hierarchical or nested multinomial logit model.

It is important to point out that the modal split stage in TRANUS is optional, and may be skipped altogether in certain circumstances. This is because of the intermodal nature of the assignment algorithm, that combines physical links and operators and routes. For a given transport category, such as people of a certain economic level, a single mode may be specified: "passenger". This single model may include all operators available for passengers, such as cars, buses, trains, walk, and so on. In this case, the model assumes that all such *supply options* may be combined and linked together to form a single travel option, subject to certain rules. For instance, a direct transfer from car to metro may not be allowed, except at park-and-ride facilities. The single-mode configuration is highly recommended in situations where there is a relatively high proportion of trip makers with car availability.

The detailed formulation for disutilities and probabilities at each stage in the calculation process is described in the following paragraphs.

Figure 18: Disutilities and probabilities at each decision level



Path choice level

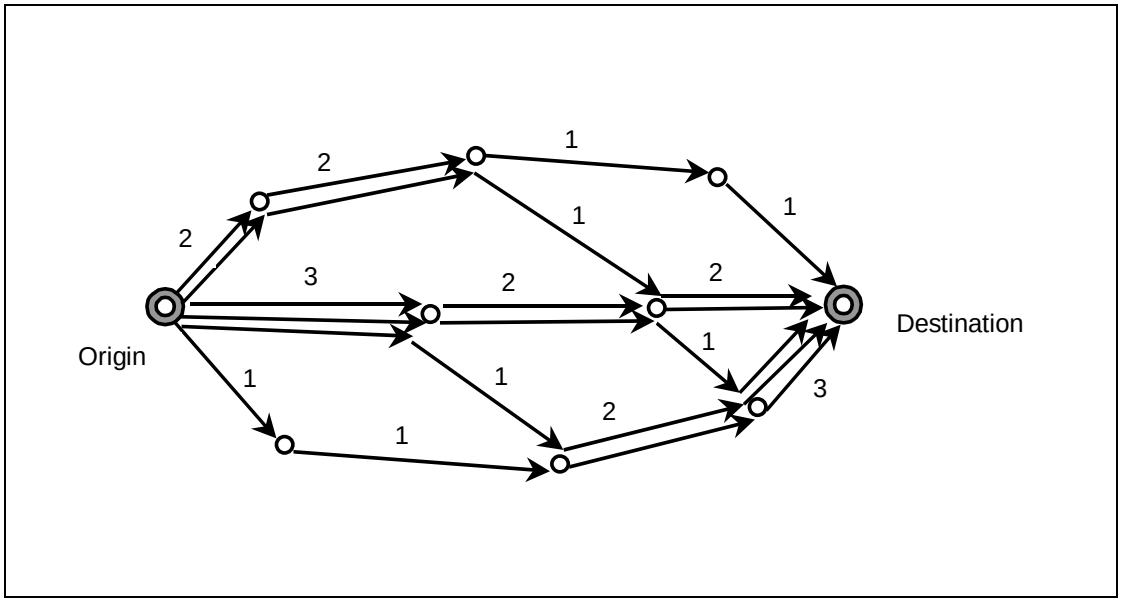
As was mentioned, the utility term at the path choice level is based on the generalized cost of each path. On the basis of the generalized cost, an *overlapped cost* term is built. Overlapping is different to the overlapping control used to generate paths. The latter was used to make the resulting path options as distinct as possible. Overlapping in the assignment model is used to compensate for possible remaining attribute correlation or overlapping among competing paths. The probability that trip makers of category *s* choose path *p* when travelling from *i* to *j* by mode *k* is given by the following scaled multinomial logit model:

$$P_{ijp}^{ks} = \frac{\exp\left(-\gamma^{s}\tilde{c}_{ijp}^{ks}\right)}{\sum_{p}\exp\left(-\gamma^{s}\tilde{c}_{ijp}^{ks}\right)}, \qquad (36)$$

where $\tilde{c}_{ijp}^{ks}$ is the scaled **and** overlapped generalized cost of travel, and γ^{s} is the dispersion parameter in the logit path choice model.

Overlapping in the assignment model is represented by multiplying the link-related elements of the generalized cost of each link/operator combination by an *overlapping factor*, defined as the number of paths that share the combination. Figure 19 shows a simple example of overlapping factors, with values of 1, 2 and 3, for a total number of paths = 6.

Figure 19: Example of overlapping factors in the assignment model



If $\dot{c}_{ijp}^{ks}$ represents the resulting overlapped generalized cost of each path, as described above, then the scaled disutility of a path becomes:

$$\tilde{c}_{ijp}^{ks} = \frac{\dot{c}_{ijp}^{ks}}{\left(\min_{p}(\dot{c}_{ijp}^{ks})\right)^{\theta^s}}, \qquad (37)$$

where θ^s sets the degree of scaling in the utility function.

The composite travel disutility $\tilde{c}_{ij}^{ks}$ from *i* to *j* by mode *k* to trip makers of category *s* is estimated by aggregating over all paths in the following way:

$$\tilde{c}_{ij}^{ks} = -\frac{\ln Pg^{ks}}{\gamma^k}\left(\min_{p}(\dot{c}_{ijp}^{ks})\right)^{\theta^s}, \qquad (38)$$

where:

θ^s is a parameter regulating the degree of scaling for category *s*.

This equation is multiplied by the overlapped cost of the minimum option because scaled utilities have been used. Pg^{ks} is defined as a series in the following way:

$$Pg^{ks} = \sum_p G_p \prod_{h=1}^{p-1}(1-G_h), \qquad (39)$$

where function G_p is the numerator of the logit model of equation (36):

$$G_p = \exp\left(-\gamma^s \tilde{c}_{ijp}^{ks}\right), \qquad (40)$$

Mode choice level

As mentioned, the mode choice stage may be skipped and replaced by operator-category penalizing factors in the assignment stage. The model user may still want to specify a mode choice level. What follows is a description of the modal split stage, if this option is used.

The probability that trip makers *s* choose a specific mode *k* to travel from *i* to *j* is calculated from the composite disutilities of each mode $\tilde{c}_{ij}^{ks}$ with the following scaled multinomial logit model:

$$P_{ij}^{ks} = \frac{\exp\left(-\lambda^s\left(\tilde{c}_{ij}^{ks}/(\min_k(\tilde{c}_{ij}^{ks}))^{\theta^s}\right)\right)}{\sum_k \exp\left(-\lambda^s\left(\tilde{c}_{ij}^{ks}/(\min_k(\tilde{c}_{ij}^{ks}))^{\theta^s}\right)\right)}, \quad k \in K^s \qquad (41)$$

where:

λ^s	dispersion parameter of the logit mode choice model;
K^s	is the set of modes *k* available to category *s*; e.g. goods only choose from freight modes, people from passenger modes, etc.;
$\tilde{c}_{ij}^{ks}$	is the composite disutility of mode *k;*
$\min_k(\tilde{c}_{ij}^{ks})$	is the composite disutility of the best mode in the choice set;
θ^s	is a parameter regulating the degree of scaling for category *s*.

Finally, the composite disutility for all trip makers *s* traveling from *i* to *j* is estimated aggregating over all modes:

$$\tilde{c}_{ij}^{s} = -\frac{\ln Pg^s}{\lambda^s}\left(\min_k(\tilde{c}_{ij}^{ks})\right)^{\theta^s}, \quad k \in K^s \qquad (42)$$

This equation is multiplied by the utility of the best mode, to unscale it back.

Pg^s is defined as a series in the following way:

$$Pg^s = \sum_k G_k \prod_{h=1}^{k-1}(1-G_h) \,, \qquad (43)$$

where function G_k is the numerator of the logit model of equation (41):

$$G_k = \exp\left(-\lambda^s \tilde{c}_{ij}^{ks}\right) \,, \qquad (44)$$

Trip generation

The purpose of the trip generation model is to calculate the number of trips derived from a functional flow estimated by the activity location model and transformed into flows by transport category by the land use/transport interface. The number of trips generated by a category *s* for an O-D pair for a particular time period is a function of the composite disutility calculated in equation (42). Strictly speaking, trip generation should be represented as a discrete choice model with a standard multinomial logit form, similar to path or mode choice. In this way trip makers would be viewed as choosing between making one trip per flow, two, trips,… or no trips at all. However, it is difficult to associate a disutility to each one of these options. A much easier way of doing this is to represent trip generation as an elastic demand curve:

$$T_{ij}^s = F_{ij}^s \left[v_{\min}^s + \left(v_{\max}^s - v_{\max-\min}^s \right) \exp\left(-\eta^s \tilde{c}_{ij}^s\right)\right] \,, \qquad (45)$$

where:

F_{ij}^s	flow by transport category *s* from *i* to *j*;
$v_{\min}^s$	minimum number of trips per unit of flow made by category *s*, whatever the value of the composite disutility;
$v_{\max}^s$	maximum number of trips per unit of flow made by category *s*, when the composite disutility tends to zero;
η^s	elasticity of category *s* with respect to the composite disutility.

In each iteration of the transport model, disutility increases because of congestion. As a result, the number of trips decreases, depending on the elasticity of the trip category. When the system converges to an equilibrium, the difference between the number of trips estimated in the first and last iterations is called *repressed demand*, that is, the number of trips that were not made because of congestion.

Modal split

If included in the model design, the modal split model estimates the number of trips of category *s* that choose mode *k*, from the modal probabilities of equation (46) and the number of trips by category calculated in (45):

$$T_{ij}^{ks} = T_{ij}^{s} P_{ij}^{ks} \left[\varphi^{s} + \left(1 - \varphi^{s}\right) B^{k} \right] , \qquad B^{k} = \begin{Bmatrix} 1 \text{ if } k \text{ is public} \\ 0 \text{ if } k \text{ is not public} \end{Bmatrix} \qquad (46)$$

Where φ^{s} is the car availability rate for transport category *s*.

Note that the probability applies only to trip makers that have a car available, while transit captive population only choose between public modes. Also note that the term *car availability* is used, and not car ownership; a trip maker may not own a car but still have a company car or share the trip with others.

Trip assignment

Trips by category and mode are assigned to paths with a scaled multinomial logit model, with the path choice probabilities calculated in equation (36) applied to the trips by mode calculated in (46):

$$T_{ijp}^{ks} = T_{ij}^{ks} * P_{ijp}^{ks} , \qquad (47)$$

Once the assignment process has finished for all O-D pairs, categories and modes, the model calculates and displays the following results:

T_m demand in the link/operator combination *m(l,o)*, in proper units (e.g. Tons, passengers)

V_m number of vehicles traveling along link/operator *m*, applying occupancy rates, except for transit routes with given fixed frequencies

VE_l number of vehicles in equivalent units on link *l* estimated as:

$$VE_l = \sum_m V_m \, eq_m , \qquad (48)$$

where eq_m are car equivalent rates by operator *o* and link *l*.

Transit operators get a special treatment; if a frequency has been defined for a particular route, the number of corresponding vehicles is a given data that the model assigns directly to the links involved. However, if a frequency has been left undefined, the model calculates the required frequency from demand figures, applies given average occupancy rates and assigns the resulting vehicles to links.:

$$f_m = \max_l \frac{T_m}{to^{o}} \quad \textit{, if the frequency is undefined,} \qquad (49)$$

where to^{o} is the occupancy rate for operator *o*. The capacity *q* of an operator or route in a link is:

$$q_m = f_m * to^o \ , \qquad (50)$$

The demand/capacity ratio for each operator in a link is:

$$dc_m = \frac{T_m}{q_m} \qquad (51)$$

The overall demand/capacity ratio for the link is calculated dividing the equivalent vehicles that share it by the given physical capacity of the link:

$$DC_l = \frac{VE_l}{Q_l} \ , \qquad (52)$$

Speeds and waiting times for each operator or route are also presented as a result of the assignment process after capacity restriction. This is described in the following section.

Capacity restriction

The general purpose of the capacity restriction procedure is to adjust travel times and waiting times as a function of the demand/capacity ratios. Such adjustments are made at the end of each iteration, once all demand has been assigned to supply. This procedure involves two distinct set of adjustments:

- the speed of all vehicles
- waiting times for transit .

In the first case, speed of vehicles in each link are reduced in function of congestion levels, respect to a fixed capacity. The model estimates queuing vehicles and the upstream queue in the precedent links. In the second case, the model use queue theory to increase waiting times as the spare capacity of a transit service is reduced.

The following sections describe the way in which these adjustments are calculated.

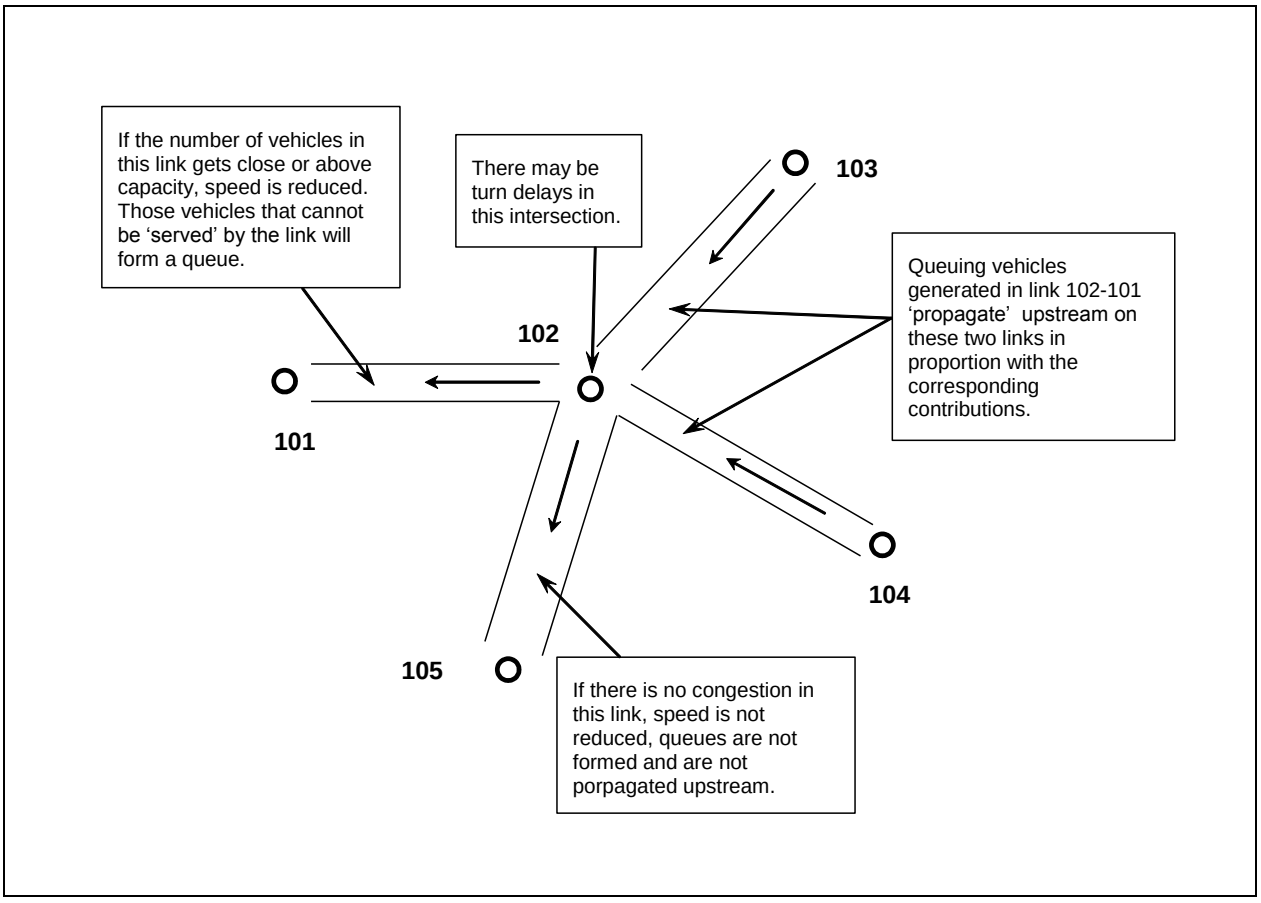
Vehicle speed adjustments

Figure 20 shows the main components of the link-based capacity restriction with a simple example. Two links converge to intersection 102 connected to other two links. When the volume assigned to link 102-101 approximates the capacity of the link, the following actions occurs:

- In the link 102-101 the speed of all vehicles are reduced
- A number of vehicles in the link 102-102 are queuing
- The queue extends to the upstream links 103-102 and 104-102 in proportion to the volume accessing the link 102-101

The traveling speed of each operator in each link is adjusted at the end of each. For those links with an undefined capacity, speeds remain constant, equal to the initial free flow speeds.

Figure 20: Link-based capacity restriction



The model applies the capacity restriction using the increased volume of the link, defined as the assigned volume plus vehicles in queue due to downstream congestion:

$$IV(l) = V(l) + Qv(l), \qquad (53)$$

where:

IV: the increased volume in the link *l*
V: assigned volume in the link *l*
Qv: queuing vehicles in the link *l* due to downstream congestion

The queuing delay function in a link approximates a Poisson curve using the following parameters:

Service Time	S = 1/free flow speed
Rate of arrival	F = Volume(*l*)
Server	C = Capacity(*l*)

Queuing delay is:

QD(*l*) = Poissons(S,F,C)

The queuing delays are added to the costs of the path, multiplying it by the value of time of the transport category. For capacity restriction only, queuing vehicles in a link *l* are added to the upstream links in the proportion of the volume incoming link *l*.

The model adjusts the speed of all operators using a link *l* with the following group of equations, which define an hyperbolic secant. The subscript *m* express a link-operator combination (*lo*).

$$V_m^\tau = V_m^0 - \text{sech}\left[\rho\left(DC_l\right)^\beta\right], \qquad (54)$$

$$\rho = \text{sech}^{-1}\left(1-\alpha\right), \qquad (55)$$

$$\beta = \frac{\ln\left[\text{sech}^{-1}\left(\upsilon\right)/\rho\right]}{\ln\gamma}, \qquad (56)$$

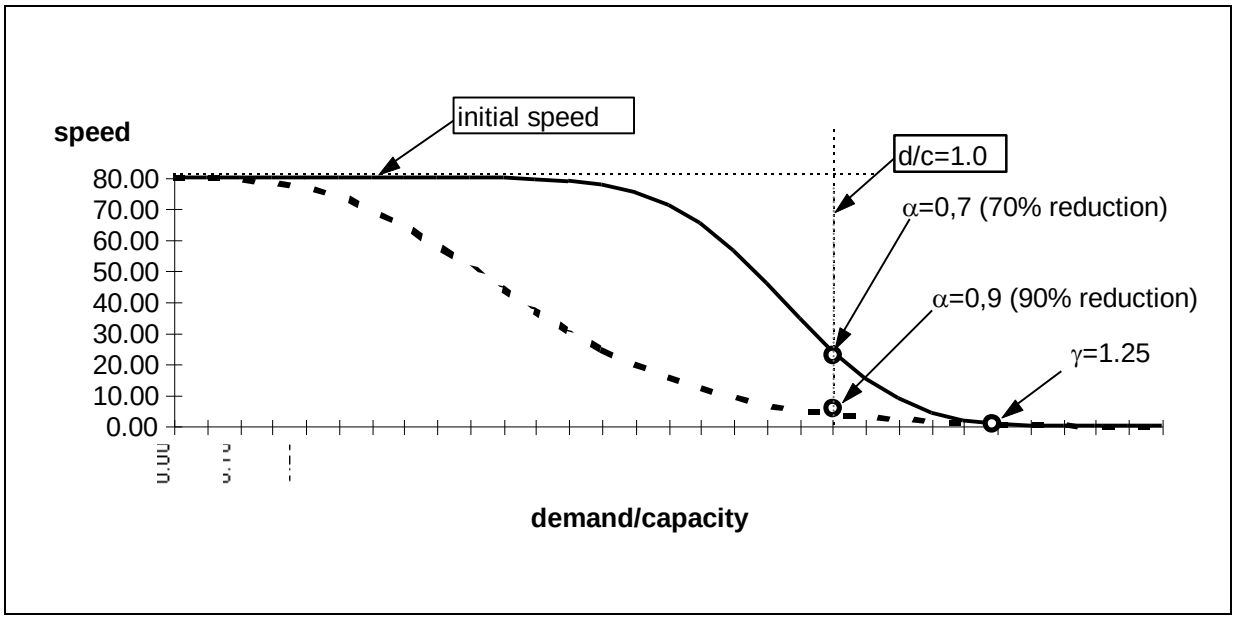
where:

V_m^τ	Speed of operator *o* in link *l*, for the iteration τ;
V_m^0	Initial (free flow) speed of operator *o* (iteration 0) in link *l;*
DC_l	Demand/capacity ratio for link *l;* queuing vehicles added to Demand
α	given proportion in which the initial speed is reduced when the demand/capacity ratio is =1
υ	given proportion in which the initial speed is reduced when the demand/capacity ratio = γ
γ	demand/capacity ratio at which the initial speed is reduced to the minimum value υ ($\gamma > 1$)

The value of DC, that is, the demand/capacity ratio is calculated by dividing the number of equivalent vehicles assigned to the link (plus queuing vehicles), over the capacity of the link, also measured in terms of equivalent vehicles. Parameters α υ γ are given inputs to the model.

Figure 21 shows an example of a speed reduction function for an operator with an initial free-flow speed of 80 Km/hr, with a value of $\gamma = 1.25$. The full line corresponds to a value of α=0.7 and the dotted line to a value of α=0.95. In both cases υ was set to 0.01. As can be seen, when *DC=0* there is no reduction to the initial speed; as *DC* increases, the speed is reduced, until it reaches a value of $(1-\alpha)V_m^0$ when *DC=1*. Beyond *DC=1*, reduction continues until DC=γ where speed is υV_m^0. From thereafter speed is asymptotic to zero.

Figure 21: Speed reduction functions for different values of α



There are two main reasons for choosing the above formulation for the capacity restriction function in Tranus:

a) Produced realistic results according to the abundant empirical evidence. The curve may be adjusted to closely match the well-known BPR curves developed by the Transport Research Laboratory of Great Britain.

b) In contrast to most curves used in transport modeling, including BPR curves, the formulae described here uses a single equation, and consequently has no discontinuities or inflexion points, which facilitates convergence.

To make convergence even easier, the speed to be used in the following iteration $\tau+1$ is the weighted average between the current speeds in iteration τ and those of the previous iteration $\tau-1$:

$$V_m^{\tau+1} = V_m^{\tau-1} + (V_m^{\tau} - V_m^{\tau-1})/(1+w)\,, \qquad (57)$$

Where w is the weight given to the previous iteration $\tau-1$ with respect to the current iteration, $w>=0$. It may be seen that if $w=0$ speed is equal to the current speed; if $w=1$ both speeds have the same weight resulting in a simple average; as w increases >1 speed converges to that of the previous iteration. If the model finds it difficult to converge, especially in the early stages of model development, the use of high values of w is recommended, such as 3.

Adjustment of waiting time

The capacity restriction procedure also adjusts waiting times for transit passengers as the number of passengers boarding units get close to capacity. This is a very important procedure that allows the model to equilibrate demand and supply in public transport services. According to queuing theory, waiting times will increase sharply as the demand/spare capacity ratio gets very close to 1. According to queuing theory, when this ratio becomes

equal to 1, waiting time is infinite. In this formulation, however, a 'softened' version is used to facilitate convergence.

The calculation of waiting time begins by specifying a minimum waiting time, that is, the time that passengers will have to wait even if the demand/spare capacity is very close to zero. This minimum waiting time, in turn, is determined by two factors: a constant term and the frequency of the route. The constant term represents the time needed for buying tickets, looking the time-tables, etc., and maybe small. The frequency of the route affects minimum waiting times if we assume that boarding passengers arrive randomly to the transit stop or station. In some cases, however, passengers may not arrive randomly if the transit service is scheduled. In such cases passengers know when the next service will come, and will arrive to the stop just before that. Hence, the minimum waiting time will be:

$$MW_m = CW_m, \qquad (58)$$

if the service is scheduled, where MW_m and CW_m denote the minimum and waiting times of route o in link l, and:

$$MW_m = CW_m + \frac{1}{2 f_o}, \qquad (59)$$

if the service is unscheduled, where f_o is the frequency of route o. This assumes that passengers arrive randomly to the station, so that on average, passengers wait for the inverse of half the frequency.

In addition to the minimum waiting time, total waiting time increases as the demand/spare capacity ratio of the route also increases. For example, if there are a few passengers arriving randomly to a bus stop and buses arrive relatively empty, then passengers will wait on average half the inverse of the frequency. However, if buses arrive with only a few spaces left, the probability that some passengers will not be able to board immediately will increase. Also, as the number of waiting passengers increases at the stop, the probability of having to wait for subsequent units increases. Furthermore, the additional waiting time will be larger if the frequency of the service is low. Thus, the average waiting time for passengers increases as the demand/spare capacity gets close to 1, and is affected by the frequency of the service.

This phenomenon may be represented as a queuing model. According to queuing theory, the case of waiting passengers represents a typical example of customers arriving randomly with the service arriving at a constant rate, with the capacity of servicing many clients at the same time. The latter is known as *bulk service*. If we assume that the rate at which passengers arrive at a station with an exponential distribution, then demand may be assumed to have a Poisson distribution. If ρ_m denotes the demand/spare capacity for route o in link l, total waiting time is calculated as:

$$TW_m = MW_m + \frac{\rho_m \frac{1}{f_o}}{1 - \rho_m}. \qquad (60)$$

Note that as the frequency f_o increases, total waiting time gets smaller. Also note that, as the demand/spare capacity ρ_m approaches to 1, the denominator $1-\rho_m$ tends to zero, so that waiting time tends to infinity. In other words, this function is characterized by its vertical asymptote at $\rho_m = 1$, and by being undefined for larger values of ρ_m. This may be seen in Figure 22.

Such asymptotic behavior is not numerically treatable in an iterative model such as TRANUS, in which ρ_m may well have values larger than one in intermediate iterations, and even in the final iteration of a model that is not yet calibrated. For practical terms, then, the model calculates the equation above as a series and then truncates it.

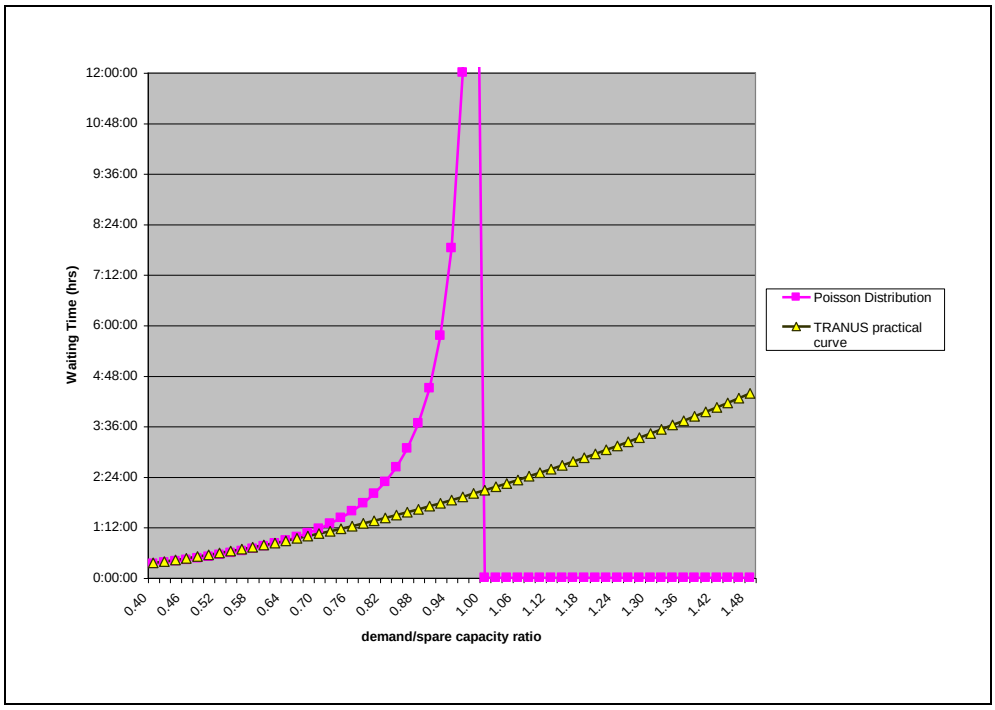
Figure 22 compares the curves obtained from the theoretical curve and that of the approximation used in TRANUS, both assuming that spare capacity = *1*. The curve used in TRANUS is very close to the Poisson distribution up to about ρ_m=*0.7*. From this point the 'practical' curve departs from the theoretical one and assumes a non-asymptotic, monotonically increasing shape.

Like speeds, waiting times are also averaged at the end of each iteration:

$$TW_m^{\tau+1} = \frac{TW_m^{\tau} + TW_m^{\tau-1}}{2} \quad , \qquad (61)$$

These last calculations end the current iteration. A fresh iteration starts with new estimates of operating costs, generalized costs and disutilities as a result of the adjusted travel and waiting times; these in turn affect trip generation, modal split and assignment, causing a new set of adjustments in capacity restriction. Once all necessary iterations have been performed, the resulting costs and disutilities will affect the location and interaction of activities in a new time period.

Figure 22: Increase in waiting time as demand/spare capacity increases



Convergence

In the transport model, convergence is checked for all links as the percentage difference between the current iteration and the previous one, considering two variables: operating speeds and traffic flows. The iterative process ends when such differences are both below a pre-defined convergence criterion. The model reports the worst case links in terms of speeds and traffic flows.

4/9/2012