SAPIENZA
UNIVERSITY OF ROME
FACULTY of ENGINEERING

Ph.D. Thesis in
Infrastructures and Transports
XXIV cycle

Models for dynamic network loading and algorithms for traffic signal synchronization in congested networks

Tutor:

Prof. Ing. Guido Gentile

Candidate:

Daniele Tiddi

My most sincere gratitude to those who really spent themselves to give a contribution to my career.
Prof. Guido Gentile and Ing. Lorenzo Meschini, for years of patient teaching and my professional growth.
My parents for anything they could.
My colleagues for the time spent together and the worthwhile collaboration.
Luciano Comelli and Luca D'Offizi for their willingness.
Donatella for her kind final review.

Index

Notation

B	lane breadth
C	signal cycle time
I	stage intergreen time
J	jam density
K	critical density
L	length
N	number of lanes
Q	capacity
S	legal speed limit
T	last time instant of the simulation
V	freeflow speed
W	jam wave speed
A	additional red time
BS	backward star set
E	cumulative outflow
EF	emission factor
F	cumulative inflow
FS	forward star set
G	cumulative space slots
$\hat{G}$	temporary cumulative space slots
H	cumulative arrives
$\hat{H}$	temporary cumulative arrives
N	cumulative vehicles
PE	pollutant emissions
Q	queue length
R	vertical storage
S	vertical queue
U	number of vehicles
VK	traffic production (vehicle-per-km)
Y	residual green
Z	new vehicles in the queue
A	set of links
C	set of conflict areas
E	set of externalities

$\boldsymbol{H}$	set of legs
$\boldsymbol{I}$	set of user classes
$\boldsymbol{J}$	set of junctions
$\boldsymbol{L}$	set of lanes
$\boldsymbol{M}$	set of vehicle types
$\boldsymbol{N}$	set of nodes
$\boldsymbol{S}$	set of stages
$\boldsymbol{T}$	set of lane turns
$\boldsymbol{X}$	set of signalized junctions
$\boldsymbol{Z}$	set of network zones
k	density function
q	flow function
t	travel time function
v	flow speed function
w	wave speed function
a	index of upstream link
b	index of downstream link
c	index of conflict area
d	turn demand flow
$e()$	outflow
$\hat{e}$	index of externality
$\tilde{e}$	modal-equivalent outflow
$f()$	inflow
$\tilde{f}$	modal-equivalent inflow
g	index of green split
h	index of leg
i	index of user class
j	index of junction
k	index of density
l	index of lane
m	index of vehicle type
$m()$	main stage
n	turn receiving flow
p	turn probability
q	index of flow
$r()$	receiving flow

$s()$	sending flow
t	index of lane turn
$t()$	travel time
$u()$	forward wave arrival time instant
v	index of speed
w	index of wave speed
x	index of node / space point
y	turn flow
z	index of zone
$z()$	backward wave arrival time instant
Γ	alternative value of density in Gentile polynomial
Θ	set of queue buckets
Λ	stage matrix
Ξ	link model
Π	node model
Φ	maximum flow
Ψ	alternative value of speed in Gentile polynomial
Ω	genetic objective function
α	multiclass flow coefficient
β	multiclass density coefficient
γ	curvature of fundamental diagram
δ	constant delay
ζ	elitism size
η	exit bottleneck share
θ	given queue bucket
ι	stage green share
κ	bucket capacity
λ	stage matrix element
μ	entry bottleneck share
ν	objective function multiplier
ξ	receiving share
π	turn priority
ρ	minimum ratio
ς	spillback condition
σ	multiclass queue share
τ	given time instant

υ	sneaking factor
ϋ	environmental sneaking factor
φ	turn capacity
χ	share of vehicle type in the fleet
ψ	squeezing factor
ω	del Castillo exponent
Б	genetic algorithm population set
Ж	genetic algorithm crossover set
З	genetic algorithm gene set
Ш	genetic algorithm mutation set
б	index of individual
ж	index of crossed over individual
з	index of gene
ш	index of mutated individual

1 Introduction

In the last decades transport demand grew up so fast that it became considerably unbalanced with respect to the available supply. The subsequent congestion levels yield incredibly high social costs, in terms of delays, fuel consumption, air and acoustic emissions and individual stress for travellers. Congestion is particularly sensible in urban areas, due to the high density of both residential and commercial activities. Additional costs generated by the congestion have a significant impact on the quality of life of its inhabitants and they hinder the social and economical development of the territory. Thus, the need to intervene on the transport system to decrease travel times and to improve the overall accessibility of town areas.

In urban networks travel times are highly affected by the presence of at-grade junctions, where the trajectories of drivers cross each other. According to the definition given in [5], junctions are the physical spaces of the transportation network where several flow streams merge into each other, diverge or crisscross to move from one road to another. Vehicle interactions yield negative effects both in terms of traffic speed and safety: these can be countered or mitigated through interventions which modify driver behaviour by removing or reducing conflicts. Such interventions can be generally grouped in four categories:

- altimetric separation (ramps, tunnels);
- planimetric shifting (roundabouts);
- time shifting (traffic signals);
- priority regulation (stop and yield signs).

The main objectives usually are safety, environmental impact and costs. On one hand, building ramps and tunnels is definitely the most effective solution, but it involves high investment costs. On the other hand, intersections with traffic lights or priority signs are cheap, but less effective. In urban areas the number of junctions and paths and the limited space rarely allow to choose the first solution, so the most common regulation strategy on junctions where traffic streams are significant is the introduction of a traffic signal. Subsequently, due to the incredibly high number of traffic signals in densely populated and busy areas, the determination of the best settings for each of them plays a very relevant role in local traffic optimization. The problem becomes even more significant, considering that close traffic signals have mutual influence on their level of service.

Generally, authority intervention at any level requires adequate decision support systems. This is due to their relevance, both in terms of the mere operational costs and of the subsequent costs sustained by the network users, which are proportional to the traffic flows involved by the intervention. In the last years, support decision systems improved their effectiveness and the authorities started to consider them for middle and long-term planning. Currently, the hardest issue is improving them to satisfy not just a *what if* methodology but to answer a *what to* question. To do this, it is necessary to gather the know-how developed by research teams and to put it into usable tools.

Moreover, the real world phenomena should be deeply inspected in order to produce realistic quantitative evaluations.
Aim of the present work is to focus on the signal control strategy, proposing a suitable formulation of the problem and an effective evaluation methodology. In the present Chapter 1 we have given an overview of the regulation strategies for junction with conflicting manoeuvres among vehicles. In Chapter 2 we introduce the signal synchronization problem, its variables and an overview of the state of the art in the methodologies about it. In Chapter 3 we offer an overview of the mathematical models which can be used to simulate the traffic phenomena on traffic networks; so as to get some useful key indicators for the optimization. In Chapter 4 the proposed formulation of the signal synchronization problem and the selected optimization technique are presented. In Chapter 5 we illustrate TOSCA, the software package integrating the proposed traffic model and optimization algorithm. In Chapter 6 the results obtained by the proposed methodology are shown, including both the traffic simulation and the evaluation of the optimization solutions. In Chapter 7 we draw our conclusions and some possible future developments.

2 The signal synchronization problem

A traffic signal is mainly described by the variables which will be briefly presented here. Each manoeuvre on the junction is allowed if the corresponding green light is on, else the stop is imposed. The first significant variable is the duration of lights (red, amber, green) for each given manoeuvre. In general, only green and red times are relevant for optimization aims while amber is dependent on the junction configuration. So it is usually worth considering an *effective green*, which is the time period during which vehicles actually cross the intersection, and an *effective red*, during which vehicles wait at their stop line. For each manoeuvre the sum of all lights is the same: this duration is called *cycle* of the signal. Thus, a cycle is the period of time after which the same sequence of lights is repeated again. Having two different green sequences during the cycle is very rare, as this usually yields drawbacks both in service and safety performances. Given the cycle time, each light duration is often expressed through its *split*, i.e. the ratio between it and the cycle. There are two ways for handling the timings of manoeuvres of one signal: *stages* and *signal groups*. A *stage* is the time interval during which no variation in lights of all manoeuvres occurs. A stage enables a specific set of possibly low-conflicting manoeuvres amongst all possible turns occurring at the junction. A *signal group* is a set of manoeuvres whose lights always coincide. Thus, manoeuvres of the same signal group are always enabled and disabled simultaneously. Two different signal groups can be contemporaneously allowed but there will be at least one instant during which one will be enabled and the other one will not. Stage and signal group representation are alternative ways to represent the same information. Given a group-stage matrix Λ, whose element λ_{ij} is positive if signal group *i* is allowed during stage *j*, the former method gives timings by row, while the other considers these by column. Sometimes the specific configuration of the intersection or some manoeuvres require the introduction of constraints to the variables. Box constraints are one type of these, i.e. minimum and/or maximum duration, both of cycle and green/red times. Another type is due to intersection clearance: there can be cross constraints between the end of one signal group and the start of another one (*intergreen time*) or of any other one (*all-red*). Finally, when considering a network of signalized junctions we need to introduce the *offset*, i.e. the time interval between a reference instant (the same for all traffic lights) and the starting instant of each signal cycle.

Difference between signal coordination and synchronization has often been a discussed subject, because of the absence of a formal definition. Here we refer to coordination as the determination of the optimal offset among several junctions with the same cycle time. Timings of these junctions are supposed to be predetermined, stemming from a previous optimization and then considering each of them as an isolated junction. We refer to synchronization as the simultaneous optimization of cycle, timings and offset of all the junctions in the network. Often, in real networks not all junctions are synchronized together. Instead, several groups of mutually synchronized junctions can be found, even close to each other. These sets are called *coordination groups*. Thus, all

the junctions belonging to the same coordination group have the same cycle, or a submultiple of it.
The signal synchronization problem is to determine the optimal values for signal parameters with respect to a given objective function, complying with the given constraints. Several data can be useful to this aim: flows on each manoeuvre, saturation flow, saturation rates, loss times up to traffic mix, road slope, presence of parking slots, transit stops and pedestrian crossings close to the junction, distance or travel time among junctions in the network.
Further data which could be subject of optimization is the sequence of the stages: this makes the problem considerably harder and it is often neglected, assuming the stage sequence as given. An example of the optimization of the stage sequence for left turn movements is given in [51].
Signal control strategies based on flow counts can be divided as follows:

- predefined plans;
- plan selection;
- actuated.

Off-line approach by predefined plans aims at determining the best signal settings for a given day period, based on given demand flows, e.g. by historical surveys; suitable optimization algorithms are generally used to this aim. Splitting the day in several periods is the only solution to make the control strategy time-varying in a certain way. Real-time regulation requires real-time traffic data, e.g. by loops or cameras; it typically implements some fast and simple reaction rules based on the detected flows. For example, it selects in real-time the best signal settings among a given set of off-line solutions built in advance (plan selection). Otherwise, the rule aims to bring the system at some desired status using an optimal control logic instead of explicitly minimizing some overall objective functions. In that case there is no actual plan as signal settings vary with a predefined frequency. In each of the previous cases, switching from one (optimal) plan to another is not trivial and optimality can be lost if this is not properly done. This aspect is particularly relevant especially in the actuated control logic, where plans are continuously determined and set. The switching problem is considered e.g. in [16], but we will not deal with it further, as this is beyond the scope of the present work.
Finding an optimal solution to the signal setting problem, also in the off-line scheme, is a task which is hard to accomplish in an analytical way, that is why heuristic methods have been often applied. Depending on the objective function and on the design variables, the algorithm to solve the signal timing problem varies substantially. Traffic lights improve safety, but they usually increase travel times, so that minimizing the total delay, or equivalently the time spent by all vehicles, may be an intuitive objective.

2.1 State of the art

The automated traffic light as we know it today, for vehicle traffic regulation, has been introduced in the ‘20s. Before that time, pedestrian and train models were already available. The problem of signal setting optimization has been studied from a theoretical point of view approximately from the half of the last century. One of the first

contributions in this field was the minimum cycle method conceived by Webster in 1958 ([68]): he proposed a formula to calculate the minimum cycle length to minimize total delay suffered by vehicles on an isolated junction. The proposed methodology had several limitations but today it is still used in academic courses as a significant example to introduce the traffic signal setting problem.
Another approach for isolated intersections was the maximization of the reserve capacity, i.e. a multiplier of demand flows, constrained by link saturation flows. This approach is complementary to the previous one, as it focuses on the maximization of the intersection throughput, disregarding delays. In the case of isolated intersections the latter can be assimilated to the Webster problem, while in junction networks the two objectives can differ significantly. E.g. in [69] a mixed heuristic is presented, minimizing the cycle time of isolated intersections and then adopting the maximum of the resulting cycles.
When considering junction networks the previous methods are inadequate, as new variables arise (signal offsets) and the interactions among different junctions turn more and more relevant the closer they are. In particular, further than the offsets, green times mutually interact too, because flows leaving an upstream junction are strictly related to flows reaching the downstream junction. One of the first significant contributions in this direction was given by Allsop in 1968 ([2]). More recently, mixed-integer formulations were presented by Gartner *et al.* in [28] and in the following [29]. Unfortunately mixed-integer optimization problems have neither efficient nor effective solution techniques and they are usually hard to solve.
However, the offset consideration led instead to the issue of the synchronization of an arterial road. Synchronizing signals along a predefined path is a particular case of the more general junction network. The specific assumptions defined for this problem yield a significantly different problem, for which suitable formulations with particular mathematical properties, convexity above all, have been proposed in the years. The most common is the bandwidth maximization. Bandwidth maximization is a quasi-concave problem, thus analytical optimization algorithms can be used to find a global solution. Bandwidth is defined as the share of cycle time during which a vehicle is able to leave the first junction and drive along the whole path up to the last junction with no stops along his run. The greater the bandwidth is, the more the number of vehicle trajectories which potentially can do this. The first and most famous solution to this problem was presented in 1966 by Little (see [45]), who introduced the MAXBAND algorithm. From 1966 the method has been further extended, e.g. to consider also the bandwidth of the opposite direction of the arterial, as in [26] and [27]. The bandwidth maximization problem has been the most common approach about synchronization until recent times, because of its properties and the possibility of an analytical solution. Its objective is intuitive, nevertheless it may lead to solutions that are far away from the minimum delay ([3]). The first important remark about maximum bandwidth approach is that the vehicle arrivals at the first junction are assumed to be uniform in time; this is the condition to make its effectiveness actually proportional to the bandwidth. Moreover, it only assesses trajectories along the given path, disregarding the vehicles

leaving or entering the arterial in middle junctions and their delay before joining the coordinated path. Actually, it does not take into account travel demand or estimated traffic flows at all. Besides, the methodology requires the travel times from junction to junction, which rarely consider congested conditions and the presence of queues. Final drawback, it is not suitable to coordinate signals on a network. Hence, it is commonly used to coordinate major urban arterials, where flows on access roads can be neglected with respect to the main stream and possibly one priority direction can be defined (e.g. inbound in the morning, outbound in the afternoon).

The strongest limitation of previous methods consists in the fact that all of them assume any intersection to work in uncongested conditions. This is a very remarkable limit when considering urban areas with current congestion levels, as they proved to be ineffective to the aims introduced in Chapter 1. Dealing with oversaturation is an issue which recently assumed greater importance: it requires both the problem modelling and its solution approach to be significantly revised. In this sense, an interesting comparison among delays in undersaturated and oversaturated conditions is performed in [19]. [8] gives an example of optimization in oversaturated conditions.

The delay minimization in signal synchronization is typically a non-convex problem, so a global optimum cannot easily be found through analytical optimization. That is why alternative formulations are often proposed, whenever no effective methods to estimate the delay are available. As soon as advances in optimization allowed it, the complex problem of delay minimization has been tackled, with several approaches and different results. Hybrid methodologies including delay calculation in bandwidth maximization have been proposed, such as in [50] and in the already considered [51]. Local optimization was instead proposed in [70].

The first method explicitly conceived to minimize the delay was proposed in 1969 by Robertson ([64]); it was called TRANSYT and from it one of the most popular commercial software packages for signal optimization was written. Now TRANSYT latest release is n. 14 ([66]), whereas several extensions or hybrid models have been proposed by scientific researchers from its birth. TRANSYT core methodology performs a hill-climbing search about all problem variables. It optimizes cycle time, green splits and offsets of every intersection node. According to this, it can optimize a junction network; delays are estimated through a statistical traffic model considering node interactions based on the distance among them. Hill-climbing is an optimization technique finding local minima, thus it is strongly affected by the starting point of the algorithm. Several studies were presented, aiming to overcome TRANSYT limits and to increase its performance. In [9] Cohen used the maximum bandwidth solutions given by MAXBAND as TRANSYT algorithm starting points. Later he extended the research in [10], using bandwidth as a constraint for TRANSYT minimum delay problem. SYNCHRO ([40]) is another commercial software for signal optimization which deserves to be mentioned thanks to its popularity, especially in the UK market. SYNCHRO was developed later than TRANSYT but it essentially targets the same topics. SYNCHRO optimization is similar to the one in TRANSYT and it is based on

statistical models. PASSER ([65]) is another commercial software developed in the US, employed for control strategies on signalized networks.
In latest years all the commercial software illustrated above have introduced genetic algorithms among their optimization tools. Genetic algorithms are stochastic optimization methods, particularly suitable for the optimization of complex problems, where finding an analytical solution seems to be unfeasible. They find sub-optimal solutions following "*survival of the fittest*" criteria and other algorithm steps inspired to the evolution sciences. Full details will be given in Chapter 4 sotto. Genetic algorithm optimization for signal setting problem was first applied in 1993 by Hadi and Wallace ([37]) for bandwidth maximization, using the TRANSYT traffic model. Six years later in 1999 Park *et al.* showed its effectiveness on oversaturated intersections ([58]), using a mesoscopic simulator for traffic indicators. Genetic algorithms are still very used to tackle the traffic signal setting problem, especially under congested conditions ([15]), coupled with traffic simulation models: macroscopic ([22], [24]), mesoscopic ([11]) or microscopic ([59]). The approach revealed to be promisingly effective, compared with other commercial tools (see [46] for a comparison with TRANSYT-7F). Finally, genetic algorithms are particularly suitable because the problem formulation and solution does not depend on the specific objective, so different objectives can be specified, for example the number of stops, queue lengths, vehicle externalities, etc. Sometimes these objectives are integrated into multi-criteria approaches.
An exhaustive review of signal control strategies must include the automated adaptive signal control systems currently used all over the world. Among them, SCOOT is today the most popular system, with hundreds of installations in the world. SCOOT ([39]) is a direct derivation of the TRANSYT strategy, determining the optimal green and offset for a network of signalized junctions, based on traffic flows detected through traffic sensors. UTOPIA ([20]) is a regulation system which performs in real-time a bilevel optimization: at a lower level it manages the single intersection greens and offsets, considering interactions with adjacent intersections, while at the upper level it regulates driver travel paths and speeds based on expected demand data and generates the constraints for the lower level solutions, to make them reciprocally compatible. Applications of UTOPIA can be found in the Italian cities of Rome, Turin and Bologna. OPAC was developed after UTOPIA by Gartner ([25]). It is a fully demand-responsive system, mostly performing each time step a new plan selection and then adopting a rolling-horizon strategy. OPAC's main installations are overspread in the US. BALANCE ([23]) is a product of the German academy. BALANCE focuses on the system modularity, thus it is immediately scalable. It includes a microscopic traffic model for performance evaluations and it explicitly allows to include public transport systems and to apply specific strategic policies (e.g. transit priority). The academic research proposed several other non-commercial responsive management strategies, such as [1] and [18]. As an example, the latter is suitable for the transit priority, like BALANCE.

3 The traffic model

Years of applications showed that the optimization quality is essentially related to the conformity of the traffic model to reality. In fact, as said before, signal optimizers require an index assessed to each solution which somehow represents its expected goodness. The index can be returned by a statistical model, by a traffic model or - more effectively – by a traffic simulator. The latter is assumed to return the realest answer; at the same time it allows a greater number of indicators but it requires greater computational efforts. Nevertheless, the computational power reached by modern processors makes the approach feasible and the results generally obtained make of it the most desirable.

Congestion is a traffic phenomenon which is extremely dynamic: it usually raises temporarily and then disappears. All traffic phenomena related to the signal setting problem require a dynamic representation: demand flows, traffic lights, stops, queues, spillback are just some of the most relevant of them. Therefore, a dynamic traffic simulator is the natural choice to fulfil our aims. Refer to [14] for a detailed review on traffic dynamics.

Finally, dynamic simulators can be divided into three main classes: microscopic, mesoscopic and macroscopic. Microscopic models explicitly represent each single vehicle in the network, with its respective attributes (e.g., speed, acceleration, destination, wished manoeuvre according to other vehicles, priority and traffic lights). The subtending models seek to represent the driver behaviour with respect to the available conditions which affect it. Microscopic models represent the most detailed behavioural models but they require very high computational efforts, due to the high number of data involved. Note that their complexity is proportional to the number of vehicles in the network, i.e. the greater the considered flows, the slower the simulation runs. VISSIM and AIMSUN are the most popular commercial software packages for microsimulation. Mesoscopic models aggregate flow data with respect to some aggregation unit. This is usually the platoon, i.e. a group of vehicles which are assumed to have a homogeneous behaviour. Platoon aggregation and dispersion models are included. Mesoscopic models are faster than microscopic models and they allow the reconstruction of disaggregate data, such as vehicle trajectories. Mesoscopic models are less popular than the others. An example of a mesoscopic model for signal optimization is illustrated in [42]. Finally, macroscopic models represent aggregate traffic variables, such as inflow, outflow and average speed by link. Due to the smaller decomposition, they are the fastest simulators, at the cost of a minor level of detail. An overview of a macroscopic model behaviour is given in [31] and [32]. In [49] a comparison between the mesoscopic and the macroscopic traffic models available in TRANSYT is performed. In particular, the macroscopic model available in TRANSYT implements the Cell Transmission Model (CTM), presented by Daganzo in [12] and [13].

The CTM is a first-order implementation of the kinematic wave theory (KWT), which was developed independently in about 1955 by Lighthill and Whitham in [44] and Richards in [63]. The CTM discretizes the space in cells of equal dimension and

homogeneous characteristics and in these cells flow propagates according to a given state equation. A Link Transmission Model (LTM), which does not require space discretization along links, is presented in [71] and [72]. In Chapter 3.1 sotto we introduce some KWT fundamentals and the General Link Transmission Model (GLTM), which is the traffic model proposed for our optimization scopes.

3.1 The General Link Transmission Model

The GLTM extends the idea of the CTM to any concave fundamental diagram, without discretizing the space into cells and considering the link as a whole. Through the GLTM it is possible to reproduce with an appropriate level of detail the main congestion phenomena that affect vehicle travel times in urban contexts, i.e. the temporal evolution of queues along links and their spillback at junctions. Here we introduce the main concepts and equations of the model, referring to [30] for further details about the formulation and the solution algorithm.

3.1.1 The fundamental diagram

GLTM is one of the models which refer to the kinematic wave theory (KWT) in its first-order implementation based on cumulative flows, according to Newell studies ([52], [53] and [54]). An academic overview of the approach is given in [33]. In such models the traffic along a link a is represented in the space x and time τ as a macroscopic mono-dimensional fluid of partially compressible particles. We introduce five main variables (we will omit index a for simplicity of notation):

$N(x,\tau)$	the number of vehicles that traversed section x until time τ
$q(x,\tau) = \partial N(x,\tau) / \partial\tau$	the flow through section x at time τ
$k(x,\tau) = -\partial N(x,\tau) / \partial x$	the density at section x and time τ
$v(x,\tau) = q(x,\tau) / k(x,\tau)$	the flow speed at section x and time τ, where $dN(x,\tau) = 0$
$w(x,\tau) = dq(x,\tau) / dk(x,\tau)$	the wave speed at section x and time τ, where $dq(x,\tau) = 0$

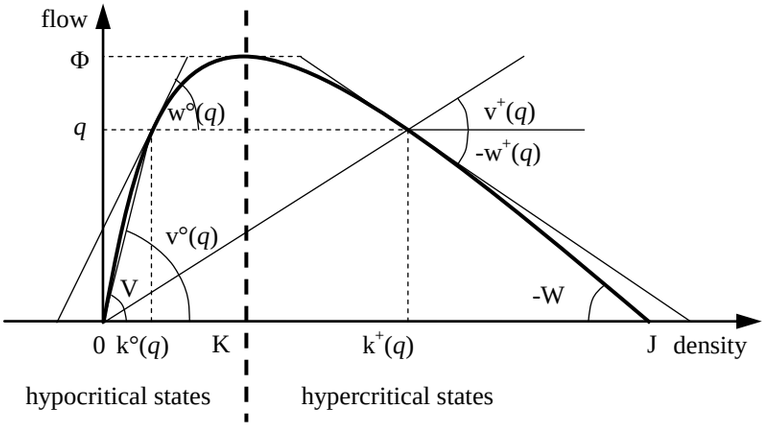


Figure 3.1: The fundamental diagram

The fundamental diagram $q = q(k)$, depicted in Figure 3.1, is the experimental relation between flow and density. It represents the two main phenomena of driver behaviour when driving along a road channel with homogeneous characteristics and no bottlenecks, that is:

- the variance of the desired speed in free flow conditions among different vehicles assuming no overtaking, so that the frequency of delays due to a slower car ahead increases with the density of the flow (hypocritical spacing);
- the need for a vehicle of keeping a safety distance from the car ahead, which depends on the speed of the flow (hypercritical spacing).

The fundamental diagram holds for stationary conditions; KWT assumes its validity also for non-stationary traffic flows. This implies that vehicles react instantaneously to changing flow states with no reaction time (although their spacing takes it into account) by adapting their speed without instabilities.

Referring separately to the two branches (hypocritical and hypercritical) of the fundamental diagram (see Figure 3.1), we can then introduce the following functions of the flow:

$k°(q) = q^{-1}(q)$	the hypocritical density (trivially $k°(0) = 0$)	(1)
$v°(q) = q / k°(q)$	the hypocritical vehicle speed	(2)
$w°(q) = 1 / [dk°(q) / dq]$	the hypocritical wave speed	(3)
$k^+(q) = q^{-1}(q)$	the hypercritical density	(4)
$v^+(q) = q / k^+(q)$	the hypercritical vehicle speed	(5)
$w^+(q) = 1 / [dk^+(q) / dq]$	the hypercritical wave speed	(6)

All the parameters defining the shape of one branch of the fundamental diagram are assumed to be constant in space and time. We denote in particular:

$V = v°(0) = w°(0)$	free flow speed	(7)
$W = -w^+(0)$	jam wave speed	(8)
$J = k^+(0)$	jam density	(9)
$\Phi = max\{q(k): k \in [0, J]\}$	physical capacity	(10)
$Q = UB\{q(k)\}$	nominal capacity	(11)
$K = k°(\Phi) = k^+(\Phi)$	critical density	(12)

Classical forms are the triangular shape and the parabolic shape. The parameters (7)-(12) are rather "physical" quantities which can be robustly estimated by direct measurements, although we often recur to model transposition and previous experience. We introduce both a physical capacity Φ and a nominal capacity Q because sometimes the input capacity of the model (Q) is neglected, if unrealistic (usually too high).

Despite there are some degrees of correlation among these parameters, they can yet be considered independent of each other; indeed in practice we can find road types with all sorts of value combinations. However, the most popular models are rather simple and they are often not capable of accommodating for more than few independent parameters, deriving others. In the following we report some well known fundamental

diagrams, from the works of Greenshield ([36]), Greenberg ([35]), Underwood ([67]) and the first studies by Newell. Among parentheses the independent parameters are given:

$q(k) =$

$k \cdot V \cdot (1 - k / J)$	Greenshield quadratic (V, J)	(13)
$k \cdot W \cdot \ln(J / k)$	Greenberg logarithmic (W, J)	(14)
$k \cdot V \cdot \exp(- k / K)$	Underwood exponential (V, K)	(15)
$min\{k \cdot V, (J - k) \cdot W\}$	Newell triangular (V, W, J)	(16)
$min\{k \cdot V, Q, (J - k) \cdot W\}$	Newell trapezoidal (V, W, J, Q)	(17)

More complex models have been recently proposed by some authors. Among these, we cite the negative power model proposed by del Castillo in [17]:

$q(k) =$

$J \cdot V \cdot [(W / V \cdot (1 - k/J))^{-\omega} + (k/J)^{-\omega}]^{-1/\omega}$ del Castillo neg. pow. (V, W, J, K) (18)

where:

$$\omega = 1 / [\ln(V/W) / \ln(J/K - 1) - 1] \quad (19)$$

For $\omega \to \infty$, i.e. for $K = J \cdot (V/W - 1)$, we have the triangular model.

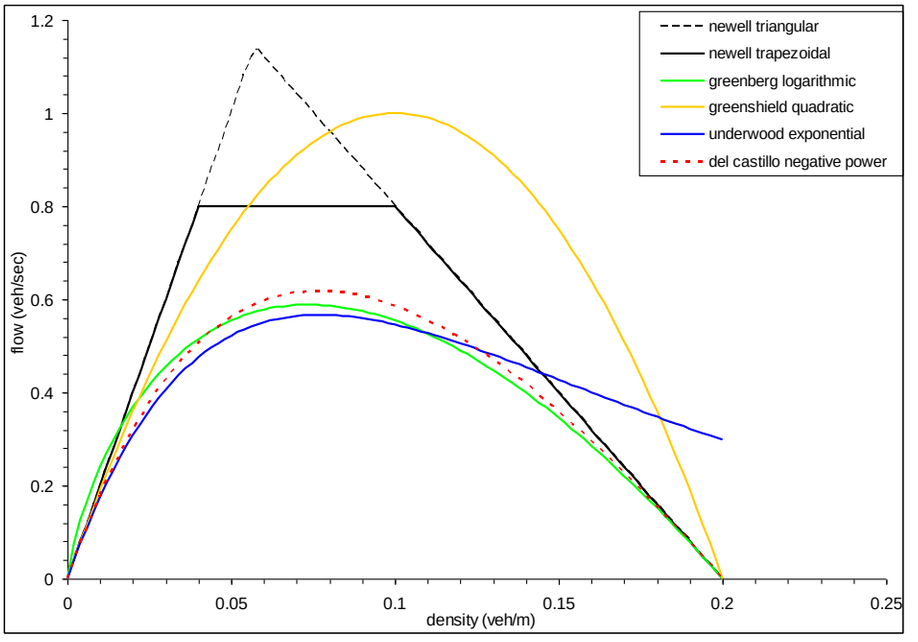


Figure 3.2: *q-k* relation for different theoretical fundamental diagrams.

3.1.2 The flow propagation

Given a generic link of length $L > 0$, let $f(\tau) = q(0,\tau)$ be its inflow ($x = 0$ is the initial section) and $e(\tau) = q(L,\tau)$ its outflow ($x = L$ is the final section of the link) at time τ. By

definition, the cumulative inflow and outflow, i.e. the number of vehicles that passed respectively the initial point and the final point of the link until that instant, are given by:

$$F(\tau) = N(0,\tau) = {}_0\int^{\tau} f(\zeta) \cdot d\zeta \tag{20}$$

$$E(\tau) = N(L,\tau) = {}_0\int^{\tau} e(\zeta) \cdot d\zeta \tag{21}$$

The instant $u(x,\tau) \geq \tau$ when the forward kinematic wave generated at time τ on the initial point of the link by the hypocritical inflow $f(\tau)$ reaches section x is given by:

$$u(x,\tau) = \tau + x \,/\, w^{\circ}(f(\tau)) \tag{22}$$

In general, $u(x,\tau)$ is not invertible, since more than one kinematic wave generated on the initial point may reach the final point at the same time (for decreasing inflows). If $f(\tau)$ is the prevailing flow state at time $u(x,\tau)$ in the final point, the corresponding cumulative flow $\hat{H}(x,\tau)$ is given by $F(\tau)$ plus the number of vehicles that have passed the forward kinematic wave with slope $w^{\circ}(f(\tau))$ generated at τ in the initial point:

$$\hat{H}(x,\tau) = F(\tau) + f(\tau) \cdot x \cdot [1 \,/\, w^{\circ}(f(\tau)) - 1 \,/\, v^{\circ}(f(\tau))] \tag{23}$$

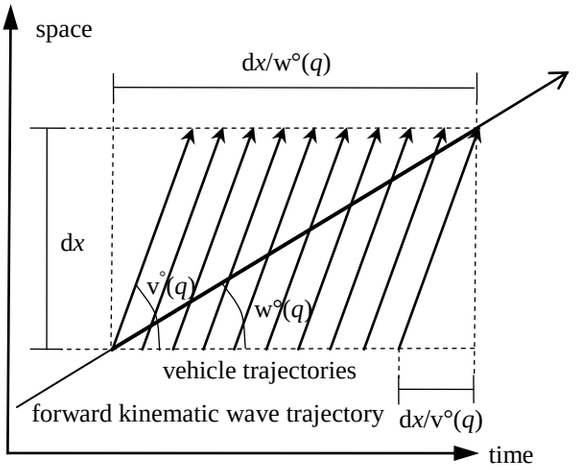


Figure 3.3: Space-time diagram of a forward kinematic wave and traversing vehicle trajectories.

The Newell-Luke Minimum Principle (NLMP) states that, among all forward kinematic waves that reach the final point at time τ, the one yielding the minimum cumulative flow, denoted $H(\tau)$, dominates the others:

$$H(x,\tau) = min\{\hat{H}(x,\sigma): u(x,\sigma) = \tau\} \tag{24}$$

The instant $z(x,\tau) \geq \tau$ when the backward kinematic wave generated at time τ on the final point of the link by the hypercritical outflow $e(\tau)$ reaches the initial point is given by:

$$z(x,\tau) = \tau - L \,/\, w^{+}(e(\tau)) \tag{25}$$

As above, $z(x,\tau)$ is not invertible, since more than one kinematic wave generated on the final point may reach the initial point at the same time (for decreasing outflows). If $e(\tau)$ is the prevailing flow state at time $z(x,\tau)$ in the initial point, the corresponding cumulative flow $\hat{G}(x,\tau)$ is given by $E(\tau)$ plus the number of vehicles that have passed the backward forward kinematic wave with (negative) slope $w^+(e(\tau))$ generated at τ in the final point:

$$\hat{G}(x,\tau) = E(\tau) + e(\tau) \cdot x \cdot [-1 / w^+(e(\tau)) + 1 / v^+(e(\tau))] \quad (26)$$

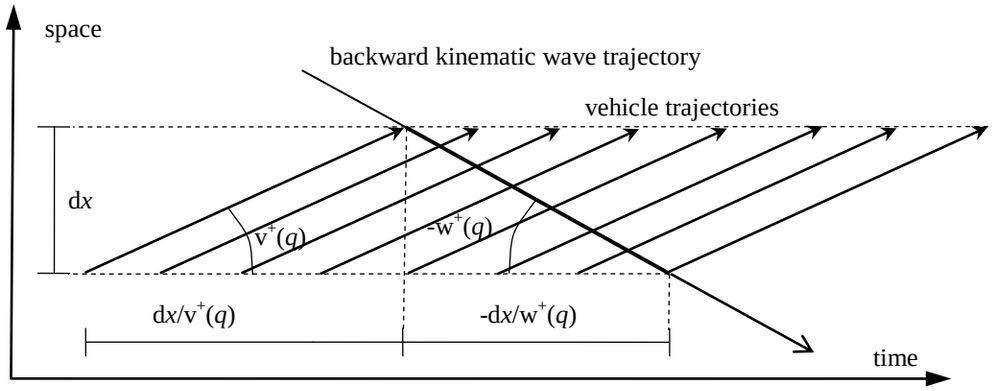


Figure 3.4: Space-time diagram of a backward kinematic wave and traversing vehicle trajectories.

The NLMP states that, among all backward kinematic waves that reach the initial point at time τ, the one yielding the minimum cumulative flow, denoted $G(\tau)$, dominates the others:

$$G(x,\tau) = min\{\hat{G}(\zeta): z(\zeta) = \tau\} \quad (27)$$

The network thus is modelled as a set of links, each one consisting of an homogeneous channel with one bottleneck at its entrance and one at its exit, that connect a set of nodes, each one consisting of an intersection where mergings and diversions take place. Cumulative flows $H(L,\tau)$ and $G(0,\tau)$ are used in the GLTM to determine respectively the sending and receiving flows, which are the input of the node model.

3.1.3 The link model

The link model takes the inflows and outflows of previous instants as an input. It provides as an output the receiving and sending flows of the next instant, associated, respectively, to the entering and exiting bottleneck, as explained in the following.
The vertical queue at time τ, denoted $S(\tau)$, is defined as:

- the vehicles entered at the initial point of the link that propagating forward would reach the final point no later than time τ if no queue was present there, represented by $H(\tau)$;
- minus the vehicles that exited the link no later than time τ, defined by $E(\tau)$;

$$S(\tau) = H(\tau) - E(\tau) \tag{28}$$

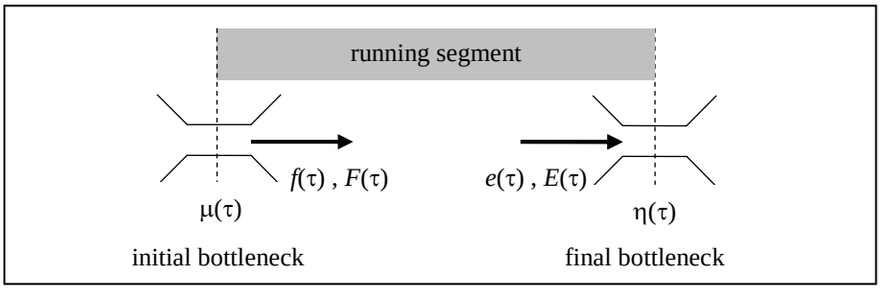


Figure 3.5: Link model and flow notation.

The vertical storage at time τ, denoted $R(\tau)$, is defined as:

- the storage capacity given by $L \cdot J$, plus the free spaces left by the vehicles at the final point of the link that propagating backward would reach the initial point no later than time τ if a queue was present there, both represented by $G(\tau)$;
- minus the vehicles that entered the link no later than time τ, defined by $F(\tau)$;

$$R(\tau) = G(\tau) - F(\tau) \tag{29}$$

The sending flow $s(\tau)$ at time τ results from the minimum between:

- the maximum flow that can exit the link under free flow conditions, which for $d\tau \rightarrow 0$ is given by $dH(\tau) / d\tau$, if the vertical queue $S(\tau)$ is null, and tends to infinity otherwise;
- the exit capacity, given by the reduction $\eta(\tau)$ of the physical capacity Φ, e.g. controlled by a traffic signal;

$$s(\tau) = min\{S(\tau) / d\tau + dH(\tau) / d\tau \, , \, \eta(\tau) \cdot \Phi\} \tag{30}$$

Analogously, the receiving flow $r(\tau)$ at time τ results therefore from the minimum between:

- the maximum flow that can enter the link under spillback conditions, which for $d\tau \rightarrow 0$ is given by $dG(\tau) / d\tau$, if the vertical storage $R(\tau)$ is null, and tends to infinity otherwise;
- the entry capacity, given by the reduction $\mu(\tau)$ of the physical capacity Φ;

$$r(\tau) = min\{R(\tau) / d\tau + dG(\tau) / d\tau \, , \, \mu(\tau) \cdot \Phi\} \tag{31}$$

Travel times can be determined applying the FIFO rule, which is formally expressed as:

$$F(\tau) = E(t(\tau)) \tag{32}$$

where $t(\tau)$ is the exit time of a vehicle entering the link at time τ.

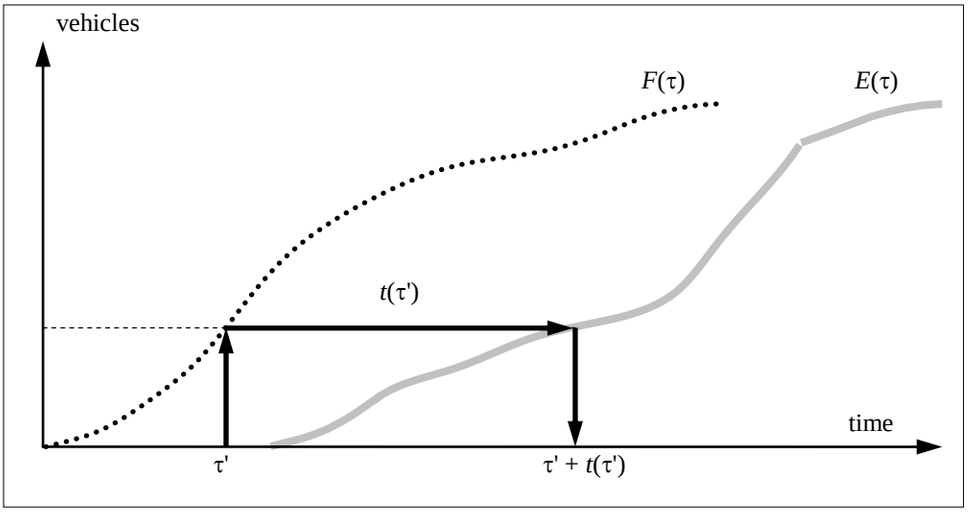


Figure 3.6: Computation of travel time based on cumulative flows.

3.1.4 The node model

The node model, referring to a given instant τ (dropped from the notation for the sake of simplicity), takes as an input the receiving flow of all its forward links and the sending flow of all its backward links, and provides as an output the inflow of all its forward links and the outflow of all its backward links, as explained in the following.

In a merging node $x \in \boldsymbol{N}$, where no routing may occur, the problem is to split the receiving flow r_b of the link $b \in FS(x)$ available at time τ among the links belonging to its backward star, whose outflows compete to get through the intersection. In principle, we assume that the receiving flow is partitioned proportionally to the priority of each link $a \in BS(x)$, defined by $\pi_{ab} \cdot \eta_a \cdot \Phi_a$, where π_{ab} is the priority coefficient of turn ab. In this way, it may happen that for some link c the turn flow y_{cb} is lower than the share of receiving flow assigned to it, so that only a lesser portion of the latter is actually exploited. Let ς_{cb} be 1 for such links and 0 for the others. The rest of the receiving flow $r_b - \sum_{c \in BS(x)} y_{cb} \cdot \varsigma_{cb}$ shall then be partitioned among the links that are in spillback from link b (i.e. all links $c \in BS(x)$: $\varsigma_{cb} = 0$). On this basis, the number of vehicles n_{ab} which can accomplish the turn ab is given by:

$$n_{ab} = \xi_b \cdot (\pi_{ab} \cdot \eta_a \cdot \Phi_a) \tag{33}$$

$$\xi_b = (r_b - \sum_{a \in BS(x)} y_{ab} \cdot \varsigma_{ab}) / (\sum_{a \in BS(x)} \pi_{ab} \cdot \eta_a \cdot \Phi_a \cdot (1 - \varsigma_{ab})) \tag{34}$$

$$\varsigma_{ab} = \begin{cases} 1 & y_{ab} < n_{ab} \\ 0 & \text{otherwise} \end{cases} \tag{35}$$

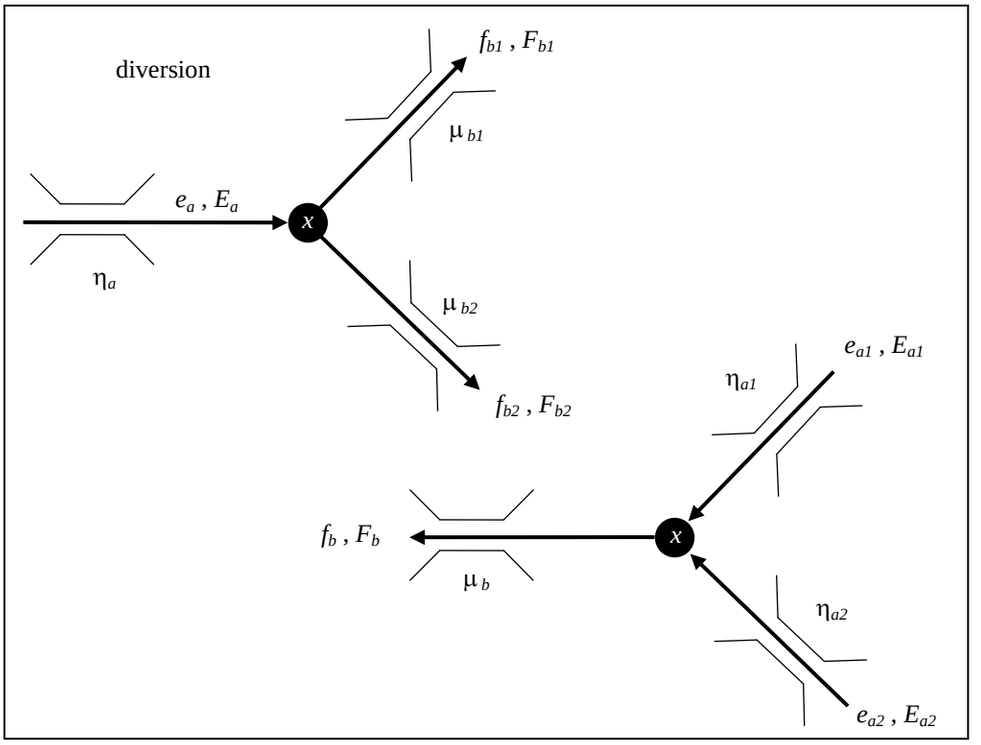


Figure 3.7: Node model and flow notation.

Calculation of definitive receiving flows is accomplished by iterating (34)-(35) at most $|BS(x)|$ times.

In a diversion node $x \in \boldsymbol{N}$, where routing takes place, the node model consists in propagating flows consistently with given path choices and satisfying the FIFO rule (no overtaking allowed), for any time instant. Path choice are represented here by the splitting rate p_{ab}, expressing the probability that the next link of the path is $b \in FS(x)$ for vehicles coming from link $a \in BS(x)$, so that the demand flow d_{ab} of turn ab is given by:

$$d_{ab} = s_a \cdot p_{ab} \tag{36}$$

The problem is to determine at the generic time τ the most severe reduction, if there is any, to the demand flow d_{ab} from link $a \in BS(x)$ among those produced by the receiving flow n_{ab} of each link $b \in FS(x)$ and by the turn capacity φ_{ab}. In order to ensure the FIFO rule applied to the vehicles exiting from link a, the share of sending flow ρ_a that actually gets through is the same for all links $b \in FS(x)$:

$$\rho_a = y_{ab} / d_{ab} = e_a / s_a \tag{37}$$

$$\rho_a = min\{1, \varphi_{ab} / d_{ab} , n_{ab} / d_{ab} : b \in FS(x), d_{ab} > 0\} \tag{38}$$

When considering a generic node $x \in \boldsymbol{N}$ with both mergings and diversions, the above relations shall hold jointly. Finally the resulting inflows and outflows are simply given as follows:

$$f_b = \sum_{a \in BS(x)} y_{ab} \quad (39)$$

$$e_a = \sum_{b \in FS(x)} y_{ab} \quad (40)$$

In the particular case where in node $x \in \boldsymbol{N}$ several separate mergings occur, i.e. flows cross each other without sharing neither origin nor destination link, e.g. when:

$$y_{ab} > 0 \Rightarrow y_{ac} = 0, \forall\ a \in BS(x), b \in FS(x), c \in FS(x)\text{-}\{b\} \quad (41)$$

we introduce the hypothesis that drivers do not occupy the intersection if they cannot cross it due to the presence of a queue on their successive link, but they wait until the necessary space becomes available. Indeed, the proposed node model itself is not capable of addressing the deterioration of performances due to a misusage of the intersection capacity. To this aim a suitable junction model is introduced later below.

3.1.5 Problem formulation

Based on (20)-(31) we can formalize the proposed link model, denoted by function L, as the following functional, for each link $a \in \boldsymbol{A}$ and time τ:

$$(s_a(\tau), r_a(\tau)) = \Xi_{a\tau}(\, f_a(\tau'), e_a(\tau')) : \tau' < \tau \quad (42)$$

Note that the link model Ξ is separable in space but non-separable in time.
Based on (34)-(40) we can formalize the proposed node model as the following functional, for each node $x \in \boldsymbol{N}$ and time τ:

$$(f_b(\tau), e_a(\tau)) = \Pi_{x\tau}(s_a(\tau), r_b(\tau)) : a \in BS(x), b \in FS(x) \quad (43)$$

Differently from the link model, the node model Π is separable in time but non-separable in space.
The link model presented in previous section provides the main input for the node model, that are the sending and receiving flows. On the other side, the output of the node model are the inflow and outflow rates, that constitute the main input for the link model.
Combining the above link and node models, we can then formulate the Dynamic Network Loading as a system of differential equations (42)-(43), which can be easily solved in chronological order.

3.2 The proposed extension models

3.2.1 The polynomial fundamental diagram

A new functional form for the fundamental diagram previously introduced in the GLTM paper is given. Let us introduce the following two parameters, additional to the parameters (7)-(12) given in Chapter 3.1.1 sopra:

S	legal speed limit	(44)
γ	convexity factor	(45)

We can express the flow-density relation as:

$$q(k) =$$

$$Q \cdot (1 - [1 - (k - \Gamma) \cdot \Psi / Q / \gamma]^{\gamma}) \quad \text{Gentile polynomial model (V, W, J, Q)} \quad (46)$$

where:

$$\Psi = \begin{cases} V & k \le K \\ -W & k > K \end{cases} \quad (47)$$

$$\Gamma = \begin{cases} 0 & k \le K \\ J & k > K \end{cases} \quad (48)$$

$$\gamma = J / Q / (1/V + 1/W) \quad (49)$$

$$K = \gamma \cdot Q / V = J - \gamma \cdot Q / W \quad (50)$$

The equation involves four independent parameters, while the other two (K, γ) can be derived from (49), (50).

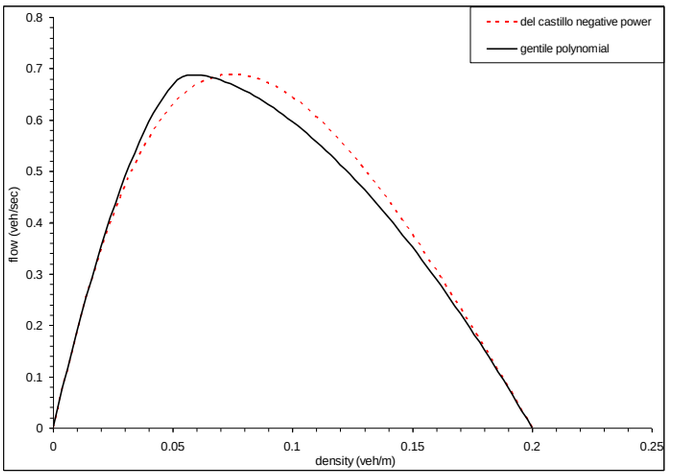


Figure 3.8: Comparison between del Castillo negative power and Gentile polynomial.

The proposed model has some key benefits, also with respect to the negative power model, although their shape do not differ significantly (see Figure 3.8):

- it is able to accommodate directly four of the usual main parameters, namely V, J, W and Q, while only the critical density K is derived;
- its shape parameter γ is related to the other parameters through the capacity;
- it comprehends Newell triangular model (16) in the case:

$$\gamma = 1 \text{ i.e. } Q = J / (1 / V + 1/ W) \quad (51)$$

- it coincides with Greenshield quadratic model (13) in the case:

$$\gamma = 2 \text{ i.e. } W = V,\ Q = J \cdot V / 4 \tag{52}$$

- it is everywhere continuously differentiable, although the second derivative is different for the two branches of the diagram joining at the capacity;
- it is invertible, i.e. we can derive analytically the following functions expressing the density and the wave speed in terms of the flow, for the two branches, separately:

$$k(q) = q^{-1}(q) = \Gamma + \gamma \cdot Q / \Psi \cdot [1 - (1- q / Q)^{1/\gamma}] \tag{53}$$

$$w(q) = 1 / [dk(q)/dq] = \Psi \cdot (1 - q / Q)^{1-1/\gamma} \tag{54}$$

Speed and speed derivative can be obtained as follows:

$$v(q) = q / k(q) \tag{55}$$

$$dv(q)/dq = [1 - q / k(q) / w(q)] / k(q) \tag{56}$$

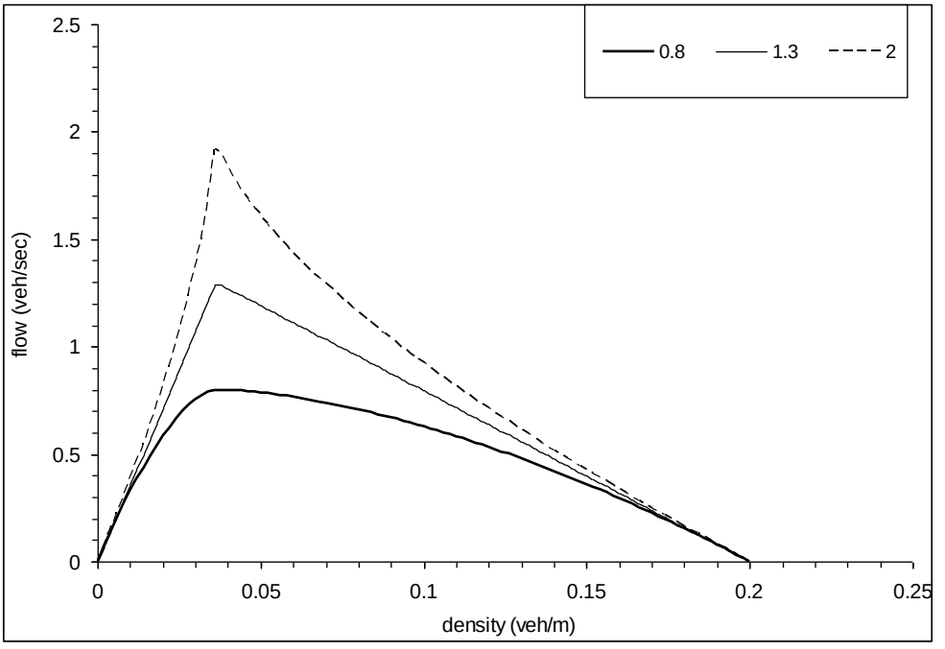


Figure 3.9: Effects of parameter Q on the shape of Gentile polynomial.

As a result of (49), if the input capacity Q is higher than that of the corresponding triangular model expressed in (51) then the resulting model in not anymore concave. In Figure 3.9 three possible values for Q are given, generating respectively a concave, triangular and convex shape. To satisfy the concavity condition, Q is consistently lowered if needed:

$$Q \leq J / (1 / V + 1/ W) \tag{57}$$

i.e. the fundamental diagram is at most triangular and its effective maximum capacity is at most the triangular one given in (51) (in Figure 3.9 this is 1.3).

From (55) we can compute the link travel time for vehicles as:

$$t(q) = L / min\{S, v(q)\} \qquad (58)$$

where L is the link length and S is the legal speed limit introduced above. Let use the index # to denote the flow equations which consider the delay due to the compliance with the legal speed limit. Then, the hypocritical branch of the fundamental diagram becomes the lower envelop between the original function $q°(k)$ and the inverse of function $k^{\#}(q) = q \cdot t(q) / L$. In Figure 3.10 is illustrated the case of a road whose characteristics would allow drivers to maintain a freeflow speed of 130 km/h but the speed limit is set to 90 km/h. As a result, the flow speed is 90 km/h (linear first piece hypocritical branch) until the flow state impose a safe speed below this limit. It is relevant to notice that when the constraint is active, i.e. when $S < v(q)$, this reduces the effective value of $q(k)$.

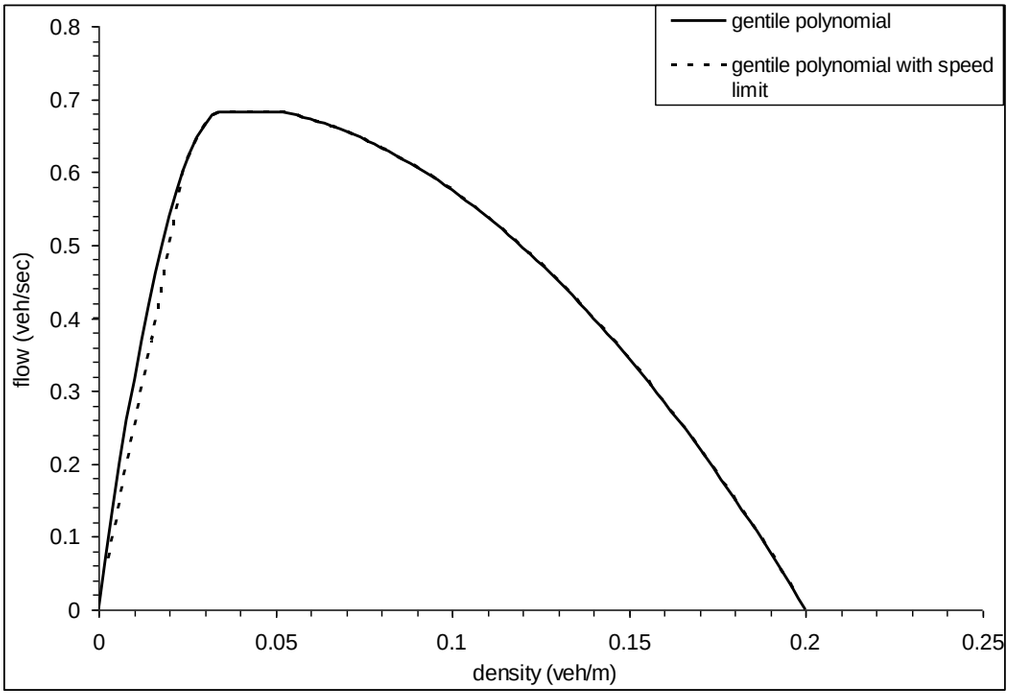


Figure 3.10: Effects of a speed limit on the shape of Gentile fundamental diagram.

Usually, rather than giving the link capacity and let the parameter γ to set subsequently the shape of the fundamental diagram, it is preferable for modelling purposes to set directly γ. According to this, we can define the following equations, where two independent shape parameters $\gamma°$ and γ^{+} are introduced to enhance model versatility:

$$q°(k) = \begin{cases} Q \cdot (1 - [1 - k \cdot V / Q / \gamma°]^{\gamma^{\circ}}) & k \le \gamma° \cdot Q / V \\ Q & \text{otherwise} \end{cases} \qquad (59)$$

$$q^{+}(k) = \begin{cases} Q \cdot (1 - [1 + (k - J) \cdot W / Q / \gamma^{+}]^{\gamma+}) & k \geq J - \gamma^{+} \cdot Q / W \\ Q & \text{otherwise} \end{cases} \quad (60)$$

The model can thus be formally described as:

$q(k) = min\{q°(k), q^{+}(k), Q\}$ Gentile polynomial model for GLTM. (61)

From (59)-(60) we can rewrite (53)-(56) for $q \in [0, Q]$ as follows:

$$k°(q) = \gamma° \cdot Q / V \cdot [1 - (1- q / Q)^{1/\gamma°}] \quad (62)$$

$$k^{+}(q) = J - \gamma^{+} \cdot Q / W \cdot [1 - (1- q / Q)^{1/\gamma+}] \quad (63)$$

$$w°(q) = V \cdot (1 - q / Q)^{1-1/\gamma°} \quad (64)$$

$$w^{+}(q) = -W \cdot (1 - q / Q)^{1-1/\gamma+} \quad (65)$$

$$v(q) = q / k(q) \quad (66)$$

$$dv(q)/dq = [1 - v(q) / w(q)] / k(q) \quad (67)$$

Note that the model capacity Φ of the link may be lower than the input capacity Q, depending on the shape parameters of the two branches, and is defined as:

$$\Phi = max\{q(k): k \in [0, J]\} \quad (68)$$

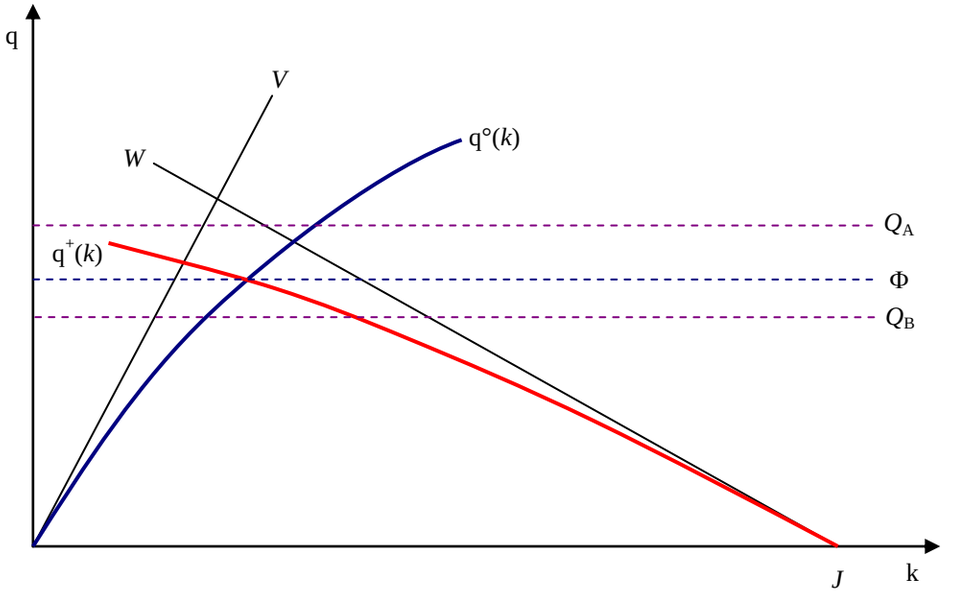


Figure 3.11: The fundamental diagram in practice.

Indeed, the two branches of the fundamental diagram do not necessarily joint smoothly into one curve and may intersect as depicted in Figure 3.11. Note that the effective capacity given by the intersection of the two branches is Φ. Whether an input capacity Q_A is given, this will not affect the model. On the other side, whether an input capacity Q_B is given, the resulting diagram is trapezoidal and $\Phi = Q_B$.
In order to avoid any doubts about the usage of Q, we recall that the link physical capacity is strictly related to the number of lanes and to the road characteristics: it shall not take into consideration any intersection, signal or bottleneck in the final section. This information should instead be used to evaluate the capacity at nodes together with turn capacity and other characteristics of the final intersection, as described in the following paragraph.

3.2.1.1 Final node delay model

Here we introduce a methodology to take into account the delay due to the presence of a traffic light or any other implicit loss of time at the final node of the link. Generally, green share and signal settings are considered as mutually exclusive alternatives to set turn capacity. When the traffic light is explicitly represented, the time-dynamic bottleneck introduced in (30) plays this role and the travel times are correctly computed. On the contrary, a relevant issue is how to consider properly the undersaturation delay at unsignalized intersections or at intersections where the signal timing is not explicitly considered and a green share is instead applied.
To this end, consider the link travel time t(q) already given in (68) and add to the legal travel time plus the unsaturated intersection delay. The latter can be constant or can be a function of the flow, as for signals:

$$t(q) = L / min\{S, v(q)\} + \delta + 0.5 \cdot C \cdot (1 - g)^2 / (1 - min\{q / (\psi \cdot Q) , g\}) \quad (69)$$

where*:*

L	link length
S	legal speed limit
δ	intersection delay (constant)
g	effective green share
C	cycle time
q	flow on the link
Q	input capacity
ψ	squeezing factor

ψ represents a squeezing phenomenon, i.e. a greater usage of road space in hypercritical conditions (e.g. flanking in more queues than the number of lanes), resulting in an increase of both the capacity and the jam density.
The effects of this new $k^{\#}(q) = q \cdot t(q) / L$ is depicted in Figure 3.12 sotto, for the following parameters:

param.	value	unit
Q	0.8	veh/h
V	130	km/h
J	0.2	veh/m
W	28.8	km/h
γ	1.6	
S	130	km/h
L	500	m
δ	5	s
g	0.5	
C	60	s
ψ	1	(veh/lane)

Table 3.1: Parameters of the fundamental diagrams in figure below.

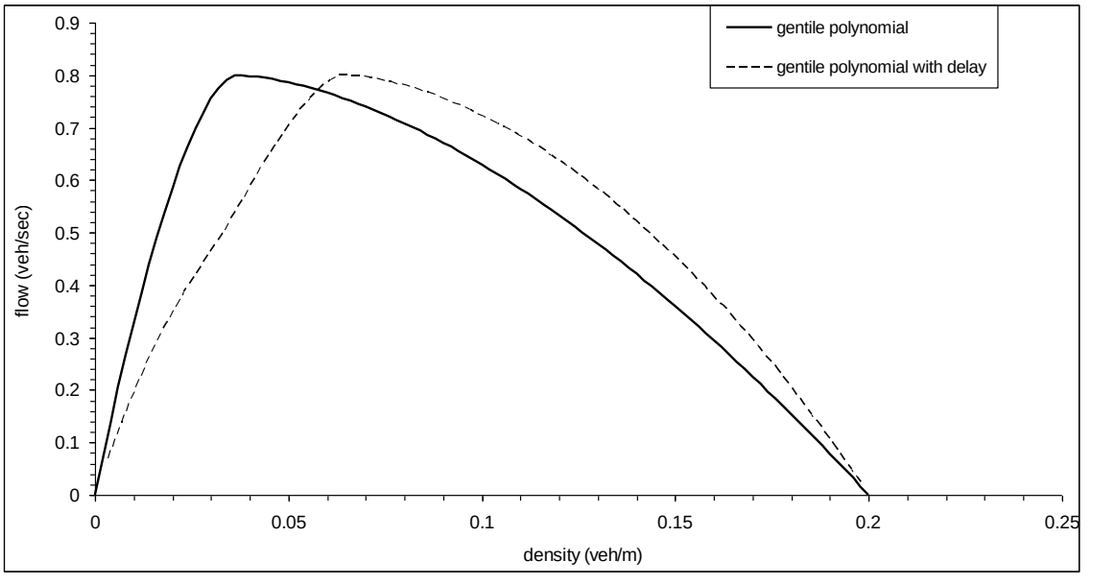


Figure 3.12: Effects of delay in Gentile fundamental diagram shape.

It is relevant to notice that in this case the delay applies both to the hypocritical and hypercritical branch of the fundamental diagram.
Here below we report the modified equations (62)-(67), when delay is considered:

$$v^{\#}(q) = L / t(q) \tag{70}$$

$$k^{\#}(q) \equiv q / v^{\#}(q) = q \cdot t(q) / L \tag{71}$$

$$dt(q)/dq = D_1 + D_2 \tag{72}$$

where:

$$D_1 = \begin{cases} - L / v°(q)^2 \cdot dv°(q)/dq & v°(q) < S \\ 0 & \text{else} \end{cases} \tag{73}$$

$$D_2 = \begin{cases} 0.5 \cdot C \cdot (1 - g)^2 / (1 - q / (\psi \cdot Q))^2 / (\psi \cdot Q) & q / (\psi \cdot Q) < g \\ 0 & \text{else} \end{cases} \quad (74)$$

$$w^{\#}(q) = 1 / [dk^{\#}(q)/dq] = L / [t(q) + dt(q)/dq \cdot q] \quad (75)$$

$$dv^{\#}(q)/dq = -L / t(q)^2 \cdot dt(q)/dq \quad (76)$$

having:

$$\begin{aligned} & dv°(q)/dq = \\ & = d[q / k°(q)]/dq = \\ & = k°(q)^{-1} - q \cdot k°(q)^{-2} \cdot dk°(q)/dq = \\ & = k°(q)^{-1} - q \cdot k°(q)^{-2} \cdot w°(q)^{-1} = \\ & = k°(q)^{-1} - v°(q) \cdot k°(q)^{-1} \cdot w°(q)^{-1} = \\ & = [1 - v°(q) / w°(q)] / k°(q) \end{aligned} \quad (77)$$

The underlying model consists in a sequence of the original link without additional delay and using the original fundamental diagram, plus a vertical queue at the end of the link behaving like a traffic light with minimum cycle. The latter means that we assume all vehicles in the vertical queue leaving the link in the next cycle and thus all of them suffer an average delay of half the red time, i.e. $C \cdot (1 - g)$. To simulate this we modify the entire fundamental diagram, thus obtaining the desired effect on travel time through an equivalent wave propagation.

It is clear that the presence of a vertical queue implies an additional density on the link, further than the one given by the regular oversaturation queue on the link. To avoid density on the link to be greater than the allowed one J (which would not allow to calculate a significant kinematic wave speed) we need to add a constraint. To calculate it explicitly through $w^{+}(q)$ we would need to solve its limit for $q \to 0$. Different approach, suppose the density to be at its limit value $k = J$, the link density is given by two components: one negative component due to the classical fundamental diagram shape ($dk(q)/dq = -q / W$) and one positive component introduced by the vertical queue ($dk(q)/dq = q \cdot (0.5 \cdot C \cdot (1 - g) + \delta) / L$). By imposing the former to be at least equal to the latter, we have:

$$L \geq W \cdot (0.5 \cdot C \cdot (1 - g) + \delta) \quad (78)$$

This grants even at maximum capacity the vertical queue to return a smaller density than the maximum allowed, as:

$$L \geq W \cdot (0.5 \cdot C \cdot (1 - g) + \delta) \geq Q \cdot (0.5 \cdot C \cdot (1 - g) + \delta) / J \quad (79)$$

being always $W \geq Q / J$ as tangential to the fundamental diagram.

3.2.2 Modelling of intersections and lanes

An intersection model more complex than the node model available in the original implementation of GLTM is presented. This allows to take explicitly into account approach lanes. These need to be correctly modelled for two practical reasons. First, from each lane only some turns are allowed and thus different lanes usually have different forward stars. Second, lanes have a separate storage capacity and vehicles queuing along one of them (e.g. due to a red light on the available manoeuvres) do not affect vehicles on other lanes, unless the lane spills back. The advantage of this method is a more accurate representation of turns, of travel times with respect to the final turn, of signal groups (if available) and thus stop and green times and of the impact of spillback due to queues on the pocket lanes.

Let $\boldsymbol{J}$ be the set of *junctions*, i.e. all nodes which need a more detailed modelling for simulation purposes. Let each junction j have a non-empty set $\boldsymbol{H}_j$ of *legs* which meet at the node and refer each to a specific link (eventually also to its return link). Let each leg h have a set $\boldsymbol{L}^+_h$ of *entry lanes* and a set $\boldsymbol{L}^-_h$ of *exit lanes*. Relevant attributes of each lane $l \in \boldsymbol{L}^+_h$ are its length L_l and its breadth B_l; for each exit lane $l \in \boldsymbol{L}^-_h$ length is not relevant as lane l does not represent a pocket lane. From each lane $l \in \boldsymbol{L}^+_h$ a set $\boldsymbol{T}_l$ of *lane turns* is allowed, each one toward a lane $l' \in \boldsymbol{L}^-_h$. The final configuration of the intersection is shown in Figure 3.13.

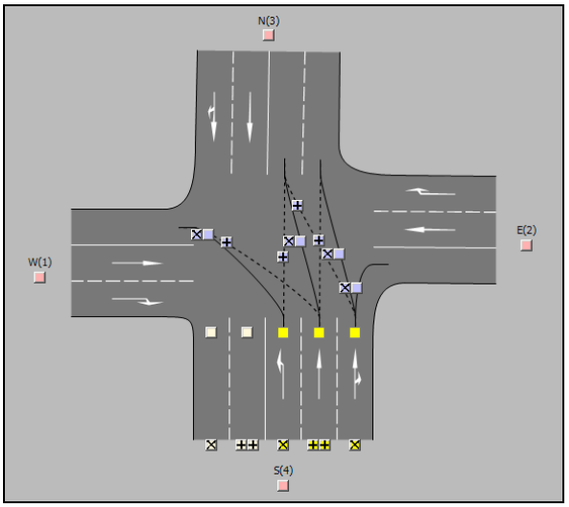


Figure 3.13: A modelled junction. Continuous line are allowed lane turns of southern leg; dashed lines are other possible manoeuvres the modeller can enable.

After introducing notation, here we describe how the junction model is implemented in the GLTM. GLTM network consists in monodirectional links. Thus, each leg is represented by one link if it has no entry or no exit lanes (i.e. $|\boldsymbol{L}^+_h| \cdot |\boldsymbol{L}^-_h| = 0$) by two separate links otherwise, one entering the junction, the other leaving it. Trivially, the link cannot have no lanes ($|\boldsymbol{L}^+_h| + |\boldsymbol{L}^-_h| = 0$), as the link would not be a leg of the junction. As there is no particular advantage in considering distinct spillback on the tail lanes of a link, multiple exit lanes of the same leg can be neglected and modelled as a

whole with the leg link without particular limitations. Therefore, this specific case will be not considered and we will assume each leg to have at most one exit lane. In Figure 3.14 the same junction of Figure 3.13 is shown, while multiple exit lanes of northbound and southbound links are considered as a unique (wider) exit lane.

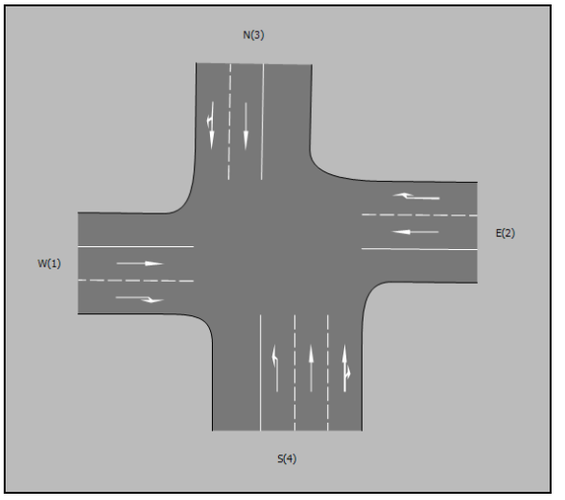


Figure 3.14: A modelled junction where exit lanes of each leg have been aggregated.

On the other side, entry lanes need to be distinguished as previously said. Thus distinct lanes are modelled as distinct links diverging from the leg link; furthermore, the lane length determines the point along the leg where the lane starts. An explanatory scheme is given in Figure 3.15.

Modelling approach lanes as distinct arcs raises the issue of establishing what is the maximum capacity of lane links. This can be set to several alternative values, according to the modeller choices. We report the following:

- equally sharing the link total capacity among all of them;
- sharing the link total capacity among all of them proportionally to the lane breadth;
- equal to the capacity of one leg lane.

To understand the last choice it is relevant to notice that the number of approach lanes is always greater or equal to the number of lanes of leg, as an effect of the pocket lanes. This often occurs in urban networks, where pocket lanes are used to increase the storage capacity immediately upstream a traffic light. It is important to remark that according to this the outflow of a road can be temporarily greater than its nominal capacity, until the queues stored on the lanes are totally cleared.

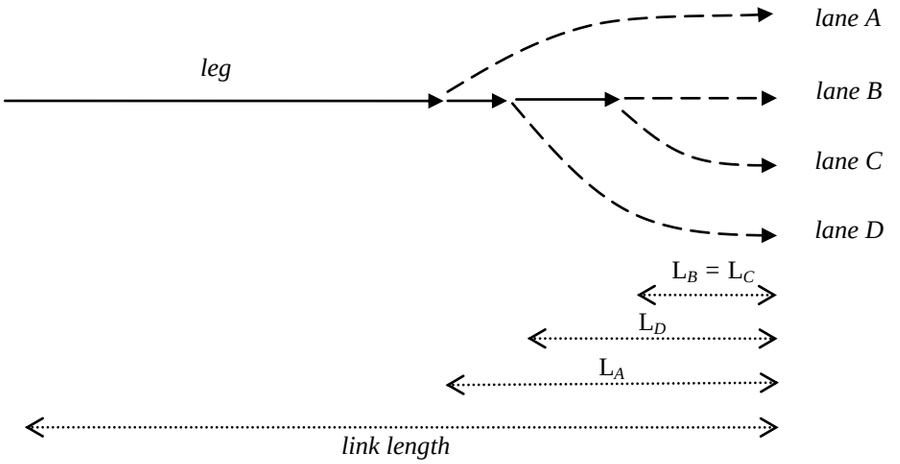


Figure 3.15: Junction schematic representation. Full lines are leg links, dashed lines are lane links.

3.2.3 The conflict area model

As it was already recalled in [5], vehicle manoeuvres may conflict in the following cases:

- merging (enter the same link);
- diversion (exit from the same link);
- crisscross with different origin link and destination link.

In practice, we reduce these conflicts to two types: the first two, where vehicles share the capacity of the upstream or of the downstream link, and the last one, where they share the space of their intersection point. The more vehicles mutually interact the more these effects become relevant and cannot be neglected for a realistic traffic representation. This means that such effects become significant wherever traffic density increases, i.e. in congested conditions and more generally in urban areas.

In most macroscopic traffic models, intersections are space elements with no dimension, i.e. no time or cost is spent by the users to cross it. At most, sometimes a turn delay is considered. Due to the common assumption of cost separation among links, usually models neglect the reciprocal influence among crisscrossing flows, even if they cross the same intersection at the same time. Some models consider a total capacity for intersection nodes, with the unrealistic assumption that the total volume crossing the node contributes to its impedance and the impedance is the same for all flows. In real traffic, vehicles interact only on specific conflict points between two conflicting movements and are delayed proportionally to the opposite flow volume. The junction model introduced above is considerably detailed but still does not handle the addressed problem. Lane turns are dummy connectors enabling vehicles to flow physically from the upstream link to the downstream link, but they have no physical meaning and do not affect flows on other turns. The model is now further extended to take this into account. Let a new object *conflict area* be introduced in the model. A conflict area $c \in \boldsymbol{C}(j)$ is a

point in junction j where two lane turns t_1 and t_2 intersect. In the example given in Figure 3.13, we can distinguish 11 different conflict areas, as shown in Figure 3.16. Note that conflict areas E, F and G are physically very close but they are distinct objects between couples of turns.

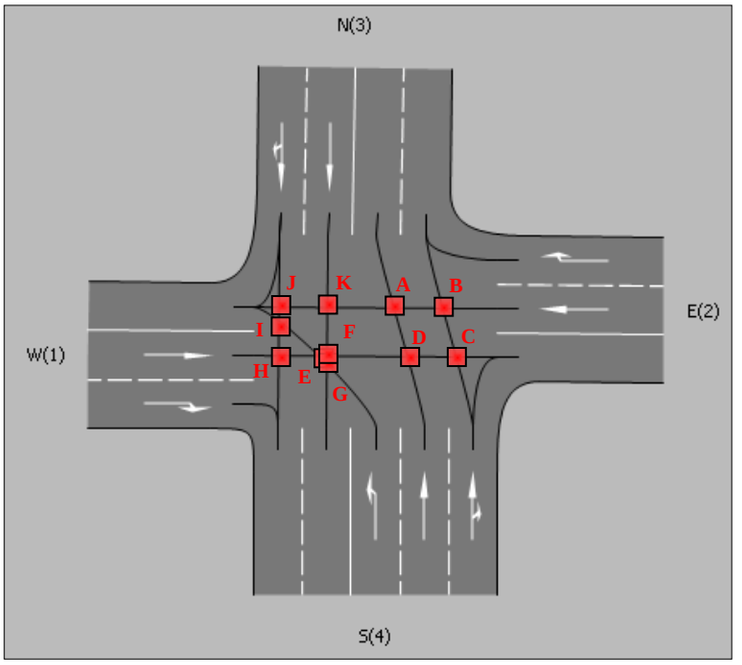


Figure 3.16: Conflict areas among conflicting manoeuvres of an intersection.

Each user performing a lane turn t will cross a sequence $\boldsymbol{C}_t$ of conflict areas before accomplishing its manoeuvre. Each lane turn t can be represented by a dummy connector only if $\boldsymbol{C}_t$ is empty. Otherwise, it is split in several links linking the given sequence $\boldsymbol{C}_t$ and the upstream and downstream lane links. A minimum length is given for all of them as we do not want to introduce sensible modifications on travel distance and time. Each conflict area $c \in \boldsymbol{C}_t$ is represented by one link with two upstream and two downstream links of the two conflicting manoeuvres. Figure 3.17 sotto shows a schematic representation for one lane turn crossing two conflict areas. Some of the sets previously introduced for the junction model are illustrated below for the given example:

j	junction index
$\boldsymbol{H}_j = \{j.1, j.2, j.3, j.4\}$	
$\boldsymbol{L}_{j.1} = \{j.1.1, j.1.2, j.1.3\}$	$j.1.1$ is the exit lane
$\boldsymbol{L}_{j.4} = \{j.4.1, j.4.2, j.4.3, j.4.4\}$	$j.4.1$ is the exit lane
$\boldsymbol{T}_{j.4.1} = \{\ \}$	exit lanes have no turns on the junction

$\boldsymbol{T}_{j.4.2}$ = {*j.4.2.4-1*} left: link 4 ► link 1
$\boldsymbol{T}_{j.4.3}$ = {*j.4.3.4-3*} straight: link 4 ► link 3
$\boldsymbol{T}_{j.4.4}$ = {*j.4.4.4-3, j.4.4.4-2*} straight: link 4 ► link 3, right: link 4 ► link 2
$\boldsymbol{C}_{j.1.2.1\text{-}2}$ = {*D*}
$\boldsymbol{C}_{j.2.2.2\text{-}1}$ = {*A*}
$\boldsymbol{C}_{j.4.3.4\text{-}3}$ = {*D, A*}

In Appendix at section 9.1 is described how the network is built and links are generated.

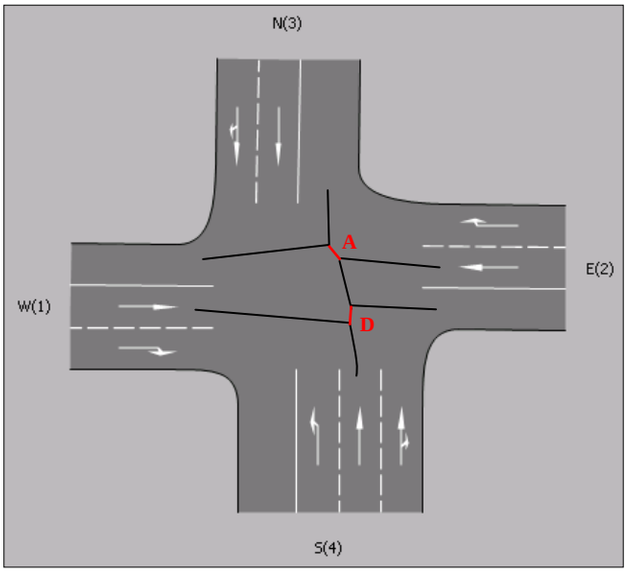


Figure 3.17: Dummy links introduced to model conflicting manoeuvres. Red links are related to the conflict areas.

For the introduced manoeuvre and conflict area links the link model described in section 3.1.3 holds regularly. However, while on mergings between two conflicting turns in a conflict area link the node model in 3.1.4 satisfies the requirements, it does not address the problem of flow diversion from the conflict area link in the two manoeuvre arcs. The problem is to assess the splitting rates p_{ab} between the manoeuvres in the forward star of the conflict area. This becomes simpler considering that the two flows actually do not merge: all the flow from the upstream manoeuvre link a is directed to the downstream link of manoeuvre t_1 and the same for t_2. Thus, introducing a violation of to the FIFO rule, the splitting rate p_{ab} is calculated as the proportion among the two flows from t_1 and t_2 on the conflict area link itself. Of course, the smaller the time interval τ, the smaller the error introduced in the FIFO rule. Moreover, (33) allows to take into account the priority π_{tc} for each lane turn t in the merging of conflict area c. In this way yield rules or balanced behaviours can be significantly simulated. The propagation function on conflict areas is described in Appendix at section 9.2.

3.2.4 The lane changing model

In the junction model introduced in 3.2.2 lanes are modelled as separate links diverging from the main trunk. This produces the representation of separate storage of vehicles addressed to different manoeuvres and thus different lanes. According to the node model in 3.1.4 sopra, in particular to condition (38), the effect of a lane link in spillback would result in a total lock of the upstream leg link until the lane spillback conditions falls. In reality, lane changing allows flow improvement according to two distinct phenomena:

- pre-emptive lane selection and anticipation of manoeuvre before the head of the leg link;
- jam density compression, as by factor ψ in (74), and subsequent sneaking.

The result is an average increase of the flow of a given manoeuvre *ab* at a node $x \in \boldsymbol{N}$, having $a \in BS(x)$ and $b \in FS(x)$, regardless the spillback conditions of any link $b' \in FS(x)-\{b\}$. In the original formulation from (38) we have:

$$y_{ab} = d_{ab} \cdot \rho_a \tag{80}$$

We extend the given equation by adding a new term which takes into account the above phenomena:

$$y_{ab} = d_{ab} \cdot \rho_a + \upsilon_a \cdot (min\{d_{ab}, n_{ab}\} - d_{ab} \cdot \rho_a) \tag{81}$$

where υ_a is called *sneaking factor* of link *a*. Actually, the sneaking factor allows a share of the flow constrained by the spillback of other manoeuvres to accomplish the wished manoeuvre *ab*. Note that in case *b* is the link in spillback, the model keeps consistent because:

$$n_{ab} = 0 \Rightarrow \rho_a = 0, y_{ab} = 0 \tag{82}$$

whatever the value of υ_a. Generally, we can expect υ_a to be proportional to three distinct factors:

- driver aggressiveness, which we could define *environmental sneaking factor* $\ddot{\upsilon}$;
- the number of lanes N_a of the upstream link *a*, as the greater it is, the more the queue will be able to spill back on the upstream link without affecting significantly other manoeuvres;
- the squeezing factor ψ already introduced in (69).

Thus, we introduce the following relation:

$$\upsilon_a = 1 - \frac{1}{1 + \ddot{\upsilon} \cdot (N_a \cdot \xi - 1)} \tag{83}$$

Advantages of (83) are:

$$\ddot{\upsilon} = 0 \Rightarrow \upsilon_a = 0 \tag{84}$$

$$\ddot{\upsilon} \to \infty \Rightarrow \upsilon_a = 1 \tag{85}$$

$$N_a \cdot \psi = 1 \Rightarrow \upsilon_a = 0 \tag{86}$$

$$N_a \cdot \psi \to \infty \Rightarrow \upsilon_a = 1 \qquad (87)$$

which is exactly what expected. Finally, we remark that when the upstream link has just one lane and no more than one vehicle per lane is allowed even in jam conditions we can expect that no actual increase is obtained as no overtaking is possible. In Table 3.2 sotto the value of υ_a as a function of $N_a \cdot \psi$ and ϋ are given.

$N_a \cdot \psi$ \ ϋ	**0.1**	**0.2**	**0.3**	**0.5**	**1**	**5**	**10**
1	0	0	0	0	0	0	0
2	0.09	0.17	0.23	0.33	0.5	0.83	0.91
3	0.17	0.29	0.38	0.5	0.67	0.91	0.95
4	0.23	0.38	0.47	0.6	0.75	0.94	0.97

Table 3.2: Link sneaking factor υ_a value over different values of the number of lanes N_a and of the environmental behaviour ϋ.

3.2.5 The multicommodity model

Due to its fast implementation, the GLTM is a suitable tool for real-time applications, such as ITS. Multimodality is often an essential specification in ITS tools. In general, representing mixed traffic is one of the main issues of dynamic simulators and this aspect has a relevant impact also on the reliability of the optimization results. Yet before the year 2000 this topic had been tackled in the KWT macroscopic models: in 1998 Lebacque *et al.* in [43] dealt with the topic of introducing buses and in 1999 Hoogendoorn and Bovy ([38]) showed a multiclass and multilane macroscopic model. Today, multiclass macroscopic models inspired to the LTM are one of the topics of Logghe's research (see e.g. [48]). Here we introduce how multiple modes can be represented by extending the previous GLTM model, where only one fundamental diagram is assumed, like in the original LWR theory. We will refer to the case of two classes, which can always be easily generalized to any number of classes.

Multiclass models in KWT usually assume that the fundamental diagrams of all the classes share a set of common characteristics, in order to satisfy some theoretical criteria. For example, in [47] the several fundamental diagrams are supposed to be similar, except if scaled by a proportionality factor α given by the vehicle class lengths:

$$\alpha = J_1 / J_2 = Q_1 / Q_2 = K_1 / K_2 \qquad (88)$$

This enables to carry out an easier and consistent computation, according to the scaling factor, but it implies a wide simplification. The example given in [47] is illustrated in Figure 3.18 sotto:

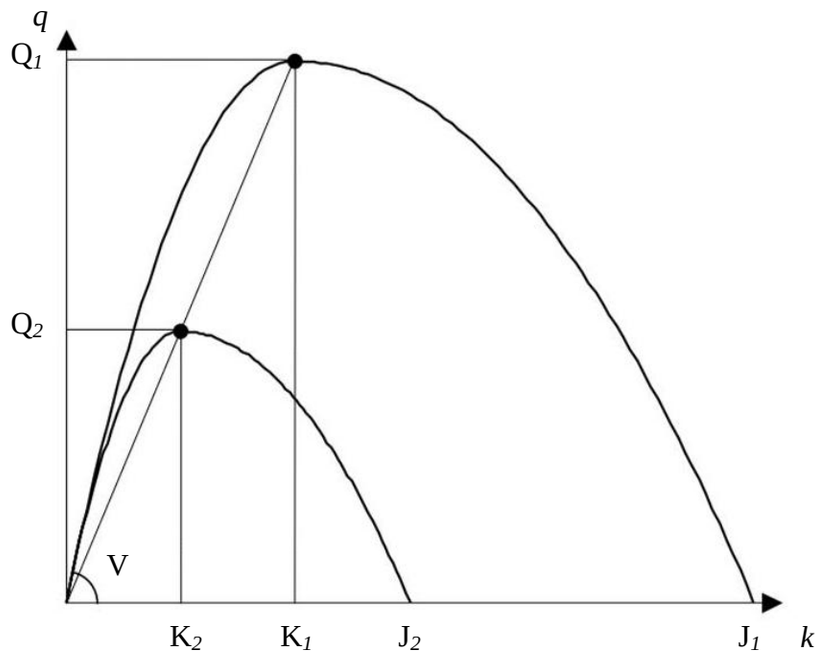


Figure 3.18: Similarly shaped fundamental diagrams (Logghe and Immers, 2003).

Note that in this case we also have $V_1 = V_2$, which is often not desirable. In other works, a weaker condition is the similarity of the sole hypercritical branches, by a scaling factor α:

$$\alpha = J_1 / J_2 \tag{89}$$

This means that all vehicle classes have the same jam wave speed value W_i and curvature γ^+ of the hypercritical branch. This prevents overtaking in hypercritical conditions (e.g. in queue). As the freeflow speed V_i is not required to be the same for all classes, different maximum flows Q_i are subsequently obtained. Other times, fundamental diagrams are required to be triangular, in order to keep the consistency of the shockwave speed even if propagating through a traffic mixtures which changes along the link.

In fact, the differences among vehicle classes which are expected to occur in traffic are the following:

- faster vehicles overtake slower vehicles in hypocritical conditions (e.g., cars and heavy ground vehicles);
- awkward vehicles may reach critical density faster than other vehicles, even faster than with respect to the length ratio, regardless their length (e.g. HGVs);
- vehicles can have a different perception of flow state and slower their hypocritical speed unproportionally to others, i.e. one can have a linear

hypocritical branch and another a parabolic curvature (e.g., motorbikes and old-aged car drivers);

- vehicles with a great agility can overtake other vehicles even in hypercritical conditions (e.g., motorbikes and cars);
- the jam condition does not hold for all vehicles contemporaneously (e.g., scooters and cars in urban jams).

According to all the conditions enumerated above, we want to generalize the models proposed above, extending the case to any shape of the fundamental diagram, regardless of proportions and curvatures both in the hypocritical and in the hypercritical branches.
Let us introduce ***I*** as the set of vehicle classes, which contains for the sake of simplicity two classes; we will address these classes with numbers 1 and 2. Note that all vehicles of a given class are mutually indistinguishable within any of the behavioural models; this means that they need to behave identically both along links (i.e. having the same vehicle type and driver behaviour) and at nodes (i.e. route choices). This would yield the modeller to define classes respectively by vehicle type (modes) and destination (users). In large scale applications the latter is often neglected as the number of classes would increase significantly.
According to equations (59) and (60), fundamental diagrams are respectively given by $q_1°(k)$ and $q_1^+(k)$ for class 1, $q_2°(k)$ and $q_2^+(k)$ for class 2, as depicted in Figure 3.19:

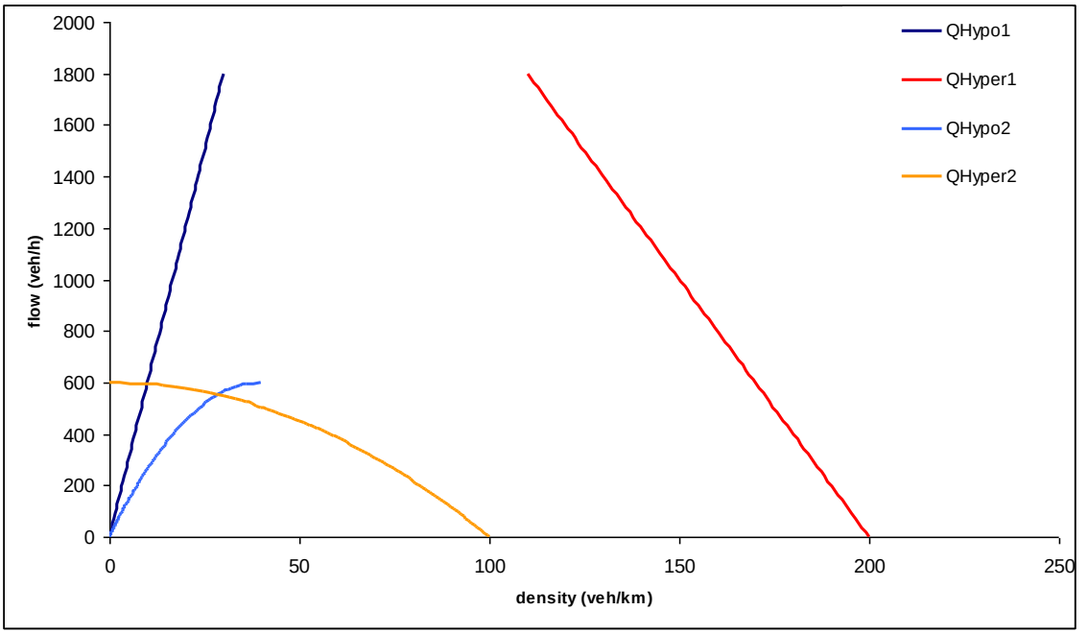


Figure 3.19: Fundamental diagrams with generic shape for two different vehicle classes.

param.	class 1	class 2	unit
Q	1800	600	veh[1]/h
V	60	30	km/h
J	200	100	veh[2]/km
W	20	12	km/h
γ°	1	2	
γ^+	1	2	
ψ	1	1	veh/lane

Table 3.3: Parameters of the fundamental diagrams in Figure 3.19.

Class 1 has a trapezoidal fundamental diagram with linear branches ($\gamma_1^\circ = \gamma_1^+ = 1$) and a maximum capacity of 1800 veh/h; class 2 has two parabolic branches ($\gamma_2^\circ = \gamma_2^+ = 2$) and the resulting maximum flow is about 550 veh/h, despite the input capacity of 600 veh/h, as the two branches intersect before reaching it. Let introduce the following parameters:

$\alpha_i = Q_{PCU} / Q_i$ ratio with respect to equivalent flow (90)

$\beta_i = J_{PCU} / J_i$ ratio with respect to equivalent density (91)

where *PCU* (Passenger Car Unit) is the car-equivalent unit. It is immediate to notice that β_i is the equivalent of the parameter α usually used in the most of the works cited above. As the fundamental relation $q = k \cdot v$ holds for any class, $\alpha_i = \beta_i$ represents the vehicle type i to have the same manageability than a *PCU*, additionally to the different length (both longer or shorter). On the contrary, $\alpha_i > \beta_i$ indicates that the vehicle occupies a different space, whatever its length. Whether $\alpha_i = \beta_i$ and $\gamma_1^\circ = \gamma_1^+ = \gamma_2^\circ = \gamma_2^+$ the model is equivalent to the one in [47] and in other works.
Below we introduce the modifications which occur both in the link and in the node model introduced in Chapter 3.1. Its theoretical aspects have been developed in recent studies and they are still under evaluation, so results in Chapter 6 sotto do not include the multiclass extension.

3.2.5.1 The propagation of multicommodity flows

Cumulative inflows $F(\tau)$ and outflows $E(\tau)$ become now:

$$F_i(\tau) = N_i(0,\tau) = {}_0\int^\tau f_i(\zeta) \cdot d\zeta \quad \forall\, i \in \boldsymbol{I} \tag{92}$$

$$E_i(\tau) = N_i(L,\tau) = {}_0\int^\tau e_i(\zeta) \cdot d\zeta \quad \forall\, i \in \boldsymbol{I} \tag{93}$$

As stated in [47], the traffic conservation law of the LWR model holds both for total traffic and for each distinct class:

$$\frac{\partial k(x,\tau)}{\partial \tau} = \frac{\partial q(x,\tau)}{\partial x} \tag{94}$$

$$\frac{\partial k_i(x,\tau)}{\partial \tau} = \frac{\partial q_i(x,\tau)}{\partial x} \forall i \in I \tag{95}$$

[1] Maximum capacity Q is expressed in vehicles of the given class.
[2] Jam density J is expressed in vehicles of the given class.

In particular, (95) introduces the generation of distinct class shockwaves at flow discontinuities, which propagate as by (3) and (6) according to the class fundamental diagram slope. The speed of each kinematic wave considered in (22), is a function of the flow at the point of discontinuity (in GLTM the inflow from the tail node). The input inflow cannot be $f_i(\tau)$, as this would neglect the traffic of other classes. On the contrary, it is:

$$u_i(x,\tau) = \tau + x / w_i^{\circ}(\tilde{f}_i(\tau)) \quad \forall\, i \in \boldsymbol{I} \tag{96}$$

where:

$$\tilde{f}_i(\tau) = \sum_{h \in \boldsymbol{I}} f_h(\tau) \cdot \alpha_h / \alpha_i \quad \forall\, i \in \boldsymbol{I} \tag{97}$$

i.e. a class-equivalent inflow $\tilde{f}_i(\tau)$ is used as an input for the fundamental diagram function $w_i^{\circ}(q)$. Similarly:

$$z_i(x,\tau) = \tau - x / w_i^{+}(\tilde{e}_i(\tau)) \quad \forall\, i \in \boldsymbol{I} \tag{98}$$

where:

$$\tilde{e}_i(\tau) = \sum_{h \in \boldsymbol{I}} e_h(\tau) \cdot \alpha_h / \alpha_i \quad \forall\, i \in \boldsymbol{I} \tag{99}$$

i.e. a class-equivalent outflow $\tilde{e}_i(\tau)$ is used as an input for the fundamental diagram function $w_i^{+}(q)$. Note that while most of multiclass models operate on one "equivalent" fundamental diagram for the whole traffic mix, $w_i^{\circ}(q)$ and $w_i^{+}(q)$ are respectively the wave speed on the hypocritical and hypercritical branches of the class fundamental diagrams and this allows to take into account the specific characteristics of each class fundamental diagram (curvature, density, speed, etc.).
Differently from the equivalent inflow and outflow, class kinematic waves propagate class cumulative flows, i.e.:

$$\hat{H}_i(x,\tau) = F_i(\tau) + f_i(\tau) \cdot x \cdot [1 / w_i^{\circ}(\tilde{f}_i(\tau)) - 1 / v_i^{\circ}(\tilde{f}_i(\tau))] \quad \forall\, i \in \boldsymbol{I} \tag{100}$$

$$\hat{G}_i(x,\tau) = E_i(\tau) + e_i(\tau) \cdot x \cdot [-1 / w_i^{+}(\tilde{e}_i(\tau)) + 1 / v_i^{+}(\tilde{e}_i(\tau))] \quad \forall\, i \in \boldsymbol{I} \tag{101}$$

and the NLMP applies separately for each class.

3.2.5.2 The multicommodity link model

The link model generates the input for the node model, i.e. the sending flows and the receiving flows. In this case, we have to start from class waves:

$$S_i(\tau) = H_i(\tau) - E_i(\tau) \quad \forall\, i \in \boldsymbol{I} \tag{102}$$

$$R_i(\tau) = G_i(\tau) - F_i(\tau) \quad \forall\, i \in \boldsymbol{I} \tag{103}$$

Note that while $S_i(\tau)$ represents the quantity of vehicles of class i that reached the end of the link and that are ready to go out, $R_i(\tau)$ represents the quantity of space along the link which was freed by the vehicles of class i which exited the link before time τ and propagated backward until the start of the link. So, before applying (30) and (31) and

comparing respectively with $\eta(\tau) \cdot \Phi$ and $\mu(\tau) \cdot \Phi$, we need to sum all class quantities up to the common unit measure of *PCU* (we assume Φ to be given in veh_{PCU}/h):

$$S^{EQ}(\tau) = \sum_{i \in \boldsymbol{I}} S_i(\tau) \cdot \alpha_i \quad (104)$$

$$R^{EQ}(\tau) = \sum_{i \in \boldsymbol{I}} R_i(\tau) \cdot \beta_i \quad (105)$$

Note that in (104) we use α_i because it is the equivalent factor for flows, while for (105) we use β_i because it is the equivalent factor for road occupancy by vehicles. Thus, given the equivalent vertical queue and physical storage respectively $S^{EQ}(\tau)$ and $R^{EQ}(\tau)$, we have:

$$s^{EQ}(\tau) = min\{S^{EQ}(\tau) / d\tau + dH^{EQ}(\tau) / d\tau \, , \eta(\tau) \cdot \Phi\} \quad (106)$$

$$r^{EQ}(\tau) = min\{R^{EQ}(\tau) / d\tau + dG^{EQ}(\tau) / d\tau \, , \mu(\tau) \cdot \Phi\} \quad (107)$$

The FIFO rule is then applied to the exit flows in the following way:

$$s_i(\tau) = S_i(\tau) \cdot (s^{EQ}(\tau) / S^{EQ}(\tau)) \quad \forall \, i \in \boldsymbol{I} \quad (108)$$

On the contrary, receiving capacity does not need to be distinct per class, as the road space let free by a vehicle of class 1 can be occupied without limitations by a vehicle of class 2, in accordance with the occupancy factor β_i, and vice versa.
Note that together (104), (106) and (108) do not allow a realistic FIFO rule, as in the vertical queue vehicles mix together and they are released proportionally to the class mix. In Chapter 3.2.5.3 sotto we introduce the node model extension and in the following Chapter 3.2.5.4 we introduce a model extension to represent more realistically multiclass queuing.

3.2.5.3 The multicommodity node model

We here remind that the node model is separable in time, so we will drop for the sake of readability index τ. The introduction of different classes in the GLTM node model involves at first the turn probabilities to be distinct by class as well. This means that, given a diversion node $x \in \boldsymbol{N}$, a link $a \in BS(x)$ and a link $b \in FS(x)$, the splitting rate $p_{ab\,i}$ expresses the probability that the next link of the path is b for vehicles of class i coming from link a. So, the turn demand flow $d_{ab\,i}$ of class i of turn ab is given by:

$$d_{ab\,i} = s_{a\,i} \cdot p_{ab\,i} \quad \forall \, i \in \boldsymbol{I} \quad (109)$$

On the other side, given a merging node x, the receiving capacity is not needed to be split among classes, as stated in previous chapter, and can be calculated by (34)-(35) as usual, while (38) needs:

$$d_{ab} = \sum_{i \in \boldsymbol{I}} d_{ab\,i} \cdot \alpha_i \quad (110)$$

Given the minimum flow share ρ_a, which is independent by class index i, $y_{ab\,i}$ can be computed as:

$$y_{ab\,i} = d_{ab\,i} \cdot \rho_a / d_{ab} \quad \forall \, i \in \boldsymbol{I} \quad (111)$$

which is the input to calculate the class cumulative flows $F_{a\,i}$.

3.2.5.4 The multicommodity queue model

Determining class sending flows through (108) actually violates the FIFO rule in case of inhomogeneous arrival times for distinct classes. To comprehend this, we can model the vertical queue into several buckets of the same dimension. Given:

Z^{EQ}	vehicles to be added to the queue (veh^{PCU})
Θ	set of buckets
κ	dimension of each bucket (veh^{PCU})
$S^{EQ}(\tau,\theta)$	vehicles in θ-th bucket of the equivalent vertical queue at time τ (veh^{PCU})
$\sigma_i(\tau,\theta)$	share of vehicles of class i in θ-th bucket at time τ

we have:

$$S^{EQ}(\tau,\theta) = \begin{cases} \kappa & \theta \le Z^{EQ} \setminus \kappa \\ Z^{EQ} \bmod \kappa & Z^{EQ} \setminus \kappa < \theta \le Z^{EQ} \setminus \kappa + 1 \\ 0 & \theta > Z^{EQ} \setminus \kappa + 1 \end{cases} \quad \forall\ \theta \in [\theta_0 + 1, n\Theta] \tag{112}$$

$$Z^{EQ} = \sum_{i \in \boldsymbol{I}} \mathrm{d}H_i(\tau)/\mathrm{d}\tau \cdot \alpha_i \tag{113}$$

$$\theta_0 = S^{EQ}(\tau\text{-}1) \setminus \kappa \tag{114}$$

$$\sigma_i(\tau,\theta) = \mathrm{d}H_i(\tau)/\mathrm{d}\tau \cdot \alpha_i / Z^{EQ} \quad \forall\ i \in \boldsymbol{I},\ \theta \in [\theta_0 + 1, n\Theta] \tag{115}$$

The node model in this case must be iterated for each $\theta \in \Theta$. Note that the number of buckets is limited by the relation:

$$n\Theta = \mathrm{L} \cdot \mathrm{J} / \kappa \tag{116}$$

A smaller number of buckets can be given, whether queue is supposed to be contained within a given maximum length.
During each iteration of the generic bucket θ it is:

$$s_{a\,i}(\theta) = \sigma_i(\tau,\theta) \cdot S^{EQ}(\tau,\theta) / \alpha_i \quad \forall\ i \in \boldsymbol{I} \tag{117}$$

while (109)-(111) hold, per bucket. The node model stops when $r^{EQ}(\tau)$ is null or $s^{EQ}(\tau)$ is empty.
The proposed model correctly addresses the FIFO problem, yielding to a totally correct representation of the rule, for $\kappa \to 0$.

3.2.6 The externality model

Subsequent to the current congestion levels on traffic networks, externality assessment, monitoring and reduction are some of the issues recently raised in the European general traffic management ([41]). From the point of view of assessment, which involves traffic models, this can be done in several ways. Some of these avoid traffic models and

extends spot measures of flows and emissions to the whole area of study. An example of this is given in [21]. Other studies introduced externality models in microsimulators, as these return a lot of useful information about every single vehicle: speed, acceleration, deceleration, type, etc. A microsimulation with externality calculation is presented in [60]. A macrosimulation assignment model considering externalities within its objective function for convergence is presented in [74]. Then, in [73], this model was used to extend the study to the effects of control strategies on vehicle emissions, which goes somehow in the direction of another issue which can be tackled with the methodology proposed in this work.
While in [73] it is possible to evaluate the effects of a given signal strategy on vehicle emissions, according to the resulting equilibrium, we want to include the minimization of emissions within the possible objectives of a signal control optimization strategy. As a result of this, we extended the present traffic model including an externality model.

3.2.6.1 The COPERT IV methodology

The COPERT IV methodology is widely considered the European standard to evaluate vehicle emissions. The model expresses the emission factor of several pollutants produced by a vehicle per travelled kilometre as a function of the vehicle speed and of some parameters indentifying the vehicle type, such as: age, engine, fuel, cold start, climatic conditions, road slope [55].
An example of the Italian fleet composition for the car mode in 2005 is provided below:

mode	fuel	engine cap.	norm.	vehicle type code	number of vehicles
Car	Gasoline	up to 1400	EURO-0+1	GC<1.4E1	8,744,642
			EURO-2	GC<1.4E2	5,045,470
			EURO-3	GC<1.4E3	3,681,425
			EURO-4	GC<1.4E4	1,093,192
		1401 – 2000	EURO-0+1	GC1.4-2.0E1	3,044,050
			EURO-2	GC1.4-2.0E2	1,628,535
			EURO-3	GC1.4-2.0E3	892,083
			EURO-4	GC1.4-2.0E4	241,748
		over 2000	EURO-0+1	GC>2.0E1	152,118
			EURO-2	GC>2.0E2	104,259
			EURO-3	GC>2.0E3	124,687
			EURO-4	GC>2.0E4	47,771
	Diesel	up to 2000	EURO-0+1	DC<2.0E1	1,006,434
			EURO-2	DC<2.0E2	1,968,463
			EURO-3	DC<2.0E3	3,988,143
			EURO-4	DC<2.0E4	1,023,699
		over 2000	EURO-0+1	DC>2.0E1	461,451
			EURO-2	DC>2.0E2	466,282
			EURO-3	DC>2.0E3	811,078
			EURO-4	DC>2.0E4	82,864

Table 3.4 – Italian car fleet composition (2005, Ministero dei Trasporti Italiano).

Given:

$\boldsymbol{I}$	set of classes
$\boldsymbol{M}_i$	set of vehicle types of class i
χ_m	share of vehicles of type m in the fleet $\boldsymbol{M}_i$
$\boldsymbol{E}$	set of pollutants

$\boldsymbol{A}$	set of links
$EF_{\hat{e}}^{m}(v)$	emission factor of externality $\hat{e}$ by a vehicle of type m at speed v (g/km)
$VK_{i\,a}$	production rate (veh · km) of class i on link a

The total emissions $PE_{\hat{e}\,a}$ of the pollutant $\hat{e}$ on link a is given by:

$$PE_{\hat{e}\,a} = \sum_{i \in \boldsymbol{I}} \left(\sum_{m \in \boldsymbol{M}i} EF_{\hat{e}}^{m}(v) \cdot \chi_m \right) \cdot VK_{i\,a} \quad \forall\, \hat{e} \in \boldsymbol{E},\, a \in \boldsymbol{A} \tag{118}$$

The emission factor of a vehicle of class i is thus calculated as the weighted average of emission factors of all vehicle types of class i at an average speed v, by the given share of that type within the vehicle fleet $\boldsymbol{M}_i$, retrieved from statistical data of the local area.
The function $EF_{\hat{e}}^{m}(v)$ of each emission factor depends on several coefficients, calibrated during the original development of the COPERT project. Below, an example of formula and parameters for vehicle type "*Gasoline car Euro3 cc1.4-2.0*" is provided:

EF	Formula (g/km)	a	b	c	d	e
CO	(a+c*V+e*V^2)/(1+b*V+d*V^2)	7.17E+01	3.54E+01	1.14E+01	-2.48E-01	0.00E+00
FC	(a+c*V+e*V^2)/(1+b*V+d*V^2)	2.17E+02	9.60E-02	2.53E-01	-4.21E-04	9.65E-03
HC	(a+c*V+e*V^2)/(1+b*V+d*V^2)	5.57E-02	3.65E-02	-1.10E-03	-1.88E-04	1.25E-05
NOx	(a+c*V+e*V^2)/(1+b*V+d*V^2)	9.29E-02	-1.22E-02	-1.49E-03	3.97E-05	6.53E-06
PM	(a+c*V+e*V^2)/(1+b*V+d*V^2)	1.19E-03	0.00E+00	0.00E+00	0.00E+00	0.00E+00

Table 3.5 – Emission factor for vehicle type "Gasoline car Euro3 cc1.4-2.0".

The figure below depicts the trend of the emission factors as a function of the average speed of the following five pollutants for the vehicle type "*Gasoline Car 1.4-2.0cc*":

- fuel consumption (FC);
- carbon monoxide (CO);
- volatile hydrocarbon (e.g.: benzene) (HC);
- nitrogen oxides (NOx);
- carbon dioxide (CO2).

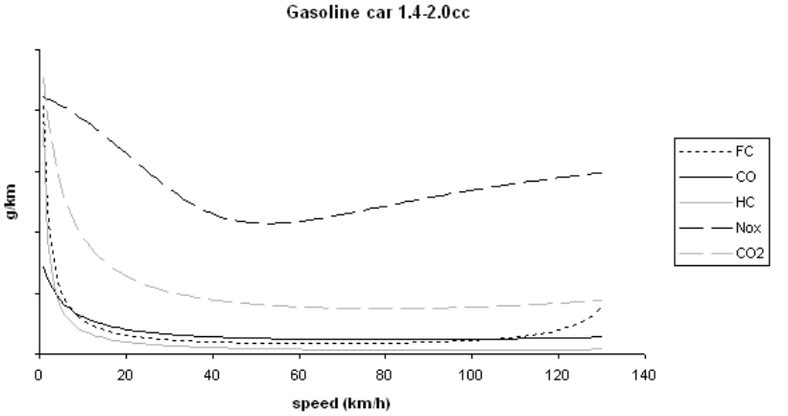


Figure 3.20: Emission profile for vehicle type "Gasoline car Euro3 cc1.4-2.0".

It is relevant to remark that the generic curve is quasiconvex and it typically has one global minimum. Furthermore, the speed corresponding to the minimum emission is not

the same for each function. Thus, when performing an optimization with respect to a combination of pollutant emissions, some multi-criteria approach is required. In the section dedicated to numerical results, we will show how synchronization parameters can influence the traffic emissions calculated by the dynamic model.

3.2.6.2 The implementation of the COPERT IV in dynamic simulation

Emissions produced by the traffic flows are calculated *ex post* of the simulation. With reference to the link model described in 3.1.3 sopra, the vehicles along the link can be split in two sets, those travelling in hypocritical state (freeflow) and those in a hypercritical states (queue). We use the same notation given in 3.1.3 and for the generic link *a* we drop the index *a* for the sake of simplicity.
Given:

L	link length of link *a*
$S(\tau)$	vertical queue of link *a* at time τ
$R(\tau)$	storage capacity of link *a* at time τ
$Q(\tau)$	queue length of link *a* at time τ
$U(\tau)$	number of vehicles on link *a* at time τ

From a theoretical point of view, the queue length $Q(\tau) = L - x$ is the position at time τ of the shockwave which separates the two flow states, with respect to the final point. I.e., in section *x* of link *a* the hypocritical and hypercritical cumulative flows are equal. This can be calculated by imposing the equality between (23) and (26):

$$H(x,\tau) = G(x,\tau) \tag{119}$$

To avoid the computational effort we apply the following linear approximation: the queue length $Q(\tau)$ is fairly given by the following:

$$Q(\tau) = L / (R(\tau) / S(\tau) + 1) \tag{120}$$

Figure 3.21 sotto illustrates the geometrical meaning of the approach:

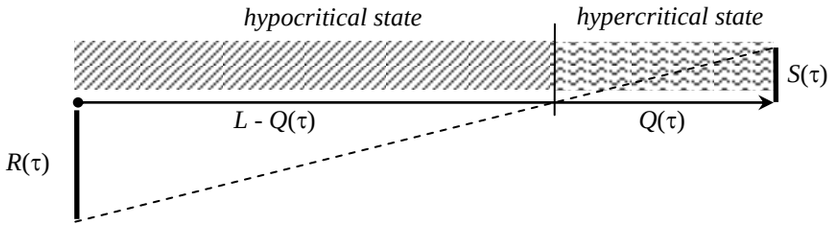


Figure 3.21: Geometrical approximated computation of the queue length.

So, it is possible to compute the average densities $k^{\circ}(\tau)$ and $k^{+}(\tau)$, respectively of vehicles in the freeflow state and in the queue state. From these, the two flow speeds $v^{\circ}(\tau)$ and $v^{+}(\tau)$ are given on the fundamental diagram. On one side, entering vehicles $f(\tau) \cdot d\tau$ are assumed to travel along the hypocritical region with a speed $v^{\circ}(\tau)$

corresponding to the average density of the hypocritical region $k°(\tau)$. On the other side, leaving vehicles $e(\tau) \cdot d\tau$ are assumed to have travelled along the hypercritical region with a speed $v^+(\tau)$ corresponding to the average density of the hypercritical region $k^+(\tau)$. The method keeps consistency in the strong assumption that the queue length is not significantly varying within the travel time of vehicles along the link.
We now provide the computation method for these densities. The number of vehicles on link a at time τ, $U(\tau)$, is given by:

$$U(\tau) = F(\tau) - E(\tau) \quad (121)$$

Not all $U(\tau)$ vehicles are travelling along the link, since $S(\tau)$ have already reached the end of the link and are waiting in the vertical queue. So, $U(\tau) - S(\tau)$ vehicles are currently travelling along the link, which is divided into two different flow state regions. Thus, only a share of these $U(\tau) - S(\tau)$ vehicles, equal to the share of the link which is not in queue i.e. $(L - Q(\tau)) / L$, is in hypocritical conditions:

$$U°(\tau) = (U(\tau) - S(\tau)) \cdot (1 - Q(\tau) / L) \quad (122)$$

The number of vehicles in a hypercritical state $U^+(\tau)$ can thus be obtained as:

$$U^+(\tau) = U(\tau) - U°(\tau) = S(\tau) + (U(\tau) - S(\tau)) \cdot Q(\tau) / L \quad (123)$$

i.e. the number of vehicles in the vertical queue, plus part of the vehicles which are still travelling along the link but are in its hypercritical state region.
We can then derive the two densities and speeds as:

$$k°(\tau) = U°(\tau) / (L - Q(\tau)) \quad (124)$$

$$k^+(\tau) = U^+(\tau) / Q(\tau) \quad (125)$$

$$v°(\tau) = v°(k°(\tau)) \quad (126)$$

$$v^+(\tau) = v^+(k^+(\tau)) \quad (127)$$

Based on this, we can thus distinguish two distinct emission factors for a link: one is produced by the hypocritical flow and the other by the hypercritical flow. The total pollution results from integrating the proposed emission model in time. According to (118), the total emission of pollutant $\hat{e}$ produced along the link is given by:

$$PE_{\hat{e}} = \int_\tau \sum_{i \in \boldsymbol{I}} \left(\sum_{m \in \boldsymbol{M}i} EF_{\hat{e}}^{m}(v°(\tau)) \cdot \chi_m \right) \cdot (L - Q(\tau)) \cdot f(\tau) \cdot d\tau +$$
$$+ \int_\tau \sum_{i \in \boldsymbol{I}} \left(\sum_{m \in \boldsymbol{M}i} EF_{\hat{e}}^{m}(v_a^+(\tau)) \cdot \chi_m \right) \cdot Q(\tau) \cdot e(\tau) \cdot d\tau \quad \forall\, \hat{e} \in \boldsymbol{E},\, a \in \boldsymbol{A} \quad (128)$$

where $(L - Q(\tau)) \cdot f(\tau) \cdot d\tau$ and $Q(\tau) \cdot e(\tau) \cdot d\tau$ are the vehicles per kilometre produced on link a during the infinitesimal time interval $[\tau,\tau+d\tau]$ by hypocritical and hypercritical flows, respectively.

4 A suitable formulation for the synchronization problem

We aim to determine optimal cycle, offset and green times, given a supply network and a time-varying demand with respect to a given objective. Signal settings are assumed to be invariant within the represented time interval: variation can be obtained by splitting the time horizon in several intervals and carrying out the optimization for each of them. The stage sequence of each intersection is given and is not subject of optimization. Differently from [6] and [7], route choice is here assumed to be inelastic and thus invariant with respect to the proposed solutions; paths are implicitly represented through given splitting rates, which can be supposedly time-varying within the simulation.
We introduce the following notation:

$\Omega(C, \boldsymbol{g}, \boldsymbol{o})$	objective function
$\boldsymbol{X}$	set of synchronized junctions
$\boldsymbol{S}_j$	set of stages of junction j
C	cycle time
g_j^s	green duration of stage s of junction j
o_j^s	offset of stage s of junction j
C_{min}	minimum cycle
C_{max}	maximum cycle
$g_{min\,j}^{\ s}$	minimum green duration of stage s of junction j
$g_{max\,j}^{\ s}$	maximum green duration of stage s of junction j
$I_j^{s,q}$	intergreen time between stage s and stage q of junction j

Here we refer as *green* to the effective green, which has been already introduced in Chapter 2 sopra. As generally intergreen time is defined between signal groups, we also define $I_j^{s,q}$ as:

$$I_j^{s,q} = max\{I_j^{p,r}: \lambda_{ps} = 1, \lambda_{rq} = 1\} \quad (129)$$

The signal setting optimization problem is the following:

$$min_{C,\,\boldsymbol{g},\,\boldsymbol{o}}\ \Omega(C, \boldsymbol{g}, \boldsymbol{o}) \quad (130)$$

$$C \in [C_{min}, C_{max}] \quad (131)$$

$$g_j^s \in [g_{min\,j}^{\ s}, g_{max\,j}^{\ s}] \quad \forall\, j \in \boldsymbol{X},\ s \in \boldsymbol{S}_j \quad (132)$$

$$o_j^s \in [0, C) \quad \forall\, s \in \boldsymbol{S}_j,\ j \in \boldsymbol{X} \quad (133)$$

$$o_j^s - (o_j^q + g_j^q) \geq I_j^{q,s} \quad \forall\, j \in \boldsymbol{X},\ s \in \boldsymbol{S}_j,\ q \in \boldsymbol{S}_j - \{s\} \quad (134)$$

$$\sum_{s \in \boldsymbol{S}j} g_j^s + (o_j^s - max\{o_j^q + I_j^{q,s}: q \in \boldsymbol{S}_j - \{s\}\}) = C \quad \forall\, j \in \boldsymbol{X} \quad (135)$$

(131)-(133) specify box constraints defining minimum and maximum values for all of the variables. (134) imposes stage q to satisfy the intergreen constraint with respect to

any of the other stages of the junction. (135) imposes the sum of all the green times g_j^s and the lost time due to intergreen to be smaller or equal to the cycle time. For the sake of simplicity, starting from now on we will refer to the resulting additional red time before stage s as A_j^s, which can be calculated as given in (135).

$$A_j^s = o_j^s - max\{o_j^q + I_j^{q,s}: q \in \boldsymbol{S}_j - \{s\}\} \quad \forall j \in \boldsymbol{X}, s \in \boldsymbol{S}_j \tag{136}$$

In the most common cases deduced from reality, the only active constraint in (136) is the one with the immediate previous stage.

Now we introduce some relaxations to reduce the problem complexity, i.e. essentially to reduce the number of variables. Let us suppose we are dealing with a synchronized path and that the junctions are subsequently ordered along this. For each signal j we define *main stage m(j)* the one allowing the manoeuvre along the defined path; any other stage is called *secondary stage*. Whether more than one stage allow the manoeuvre, we can set without loss of generality one of these as the main stage and the other(s) as secondary stages.

Given the main stage green time, we define *residual green* Y_j of junction j as the difference between the cycle time and the sum of the main stage green $g_j^{m(j)}$, the minimum green of each secondary stage $g_{min\,j}^s$ and the lost time A_j^s. In practice, the residual green is the unused green time to be assigned, given the main stage green duration. Thus, the green times of secondary stages can be obtained as the minimum green of each stage plus the proper share of the residual green according to a given proportion ι_j^s. Then, as we assumed the stage sequence to be fixed, secondary offsets come subsequently from stage green times and additional reds.

Thus, we have:

$$Y_j = C - [(g_j^{m(j)} + A_j^{m(j)}) - \sum_{s \in \boldsymbol{S}j - \{m(j)\}} g_{min\,j}^s + A_j^s)] \quad \forall j \in \boldsymbol{X} \tag{137}$$

$$g_j^s = min\{g_{min\,j}^s + \iota_j^s \cdot Y_j, g_{max\,j}^s\} \quad \forall j \in \boldsymbol{X}, s \in \boldsymbol{S}_j - \{m(j)\} \tag{138}$$

$$o_j^s = (o_j^{s-1} + g_j^{s-1} + A_j^{s-1}) \bmod C \quad \forall j \in \boldsymbol{X}, s \in \boldsymbol{S}_j - \{m(j)\} \tag{139}$$

Whether the constraint (138) becomes active with respect to the maximum green, the remaining green can be stored in a green "pool" and steps (137)-(138) can be repeated. The following constraints grant the non-emptiness of the solution set:

$$C_{min} \geq \sum_{s \in \boldsymbol{S}j} (g_{min\,j}^s + A_j^s) \tag{140}$$

$$C_{max} \leq \sum_{s \in \boldsymbol{S}j} (g_{max\,j}^s + A_j^s) \tag{141}$$

These can be satisfied with a pre-emptive check-and-correct of (131).

(137)-(141) allow to reduce considerably the number of variables only to the common cycle time and the green and the offset of main stages of every synchronized junction. So, problem complexity drops from $2 \cdot n\boldsymbol{X} \cdot \varsigma + 1$, where ς is the average number of stages per junction, to just $2 \cdot n\boldsymbol{X} + 1$. So for sake of readability from now on we will refer to $g_j^{m(j)}$, $o_j^{m(j)}$, $g_{min\,j}^{m(j)}$, $g_{max\,j}^{m(j)}$ simply as g_j, o_j, $g_{min\,j}$, $g_{max\,j}$. Thus, each solution is given by a tuple $(C, g_1, o_1, \ldots, g_n, o_n)$.

The proposed formulation is tailored on a sequence of junctions, so it is particularly suitable for arterial synchronization. Nevertheless, also junctions not belonging to the path can be added to the problem. For example, simultaneous optimization of several corridors is possible, considering the main stages of each corridor. Whether some corridors intersect, we will need the common junctions to occur only once, thus assigning their main stages to one corridor and letting the heuristic above to define the timings of remaining corridor(s) on the shared junctions. Generally, a whole junction network can be optimized, even if it is not possible to define any synchronization path, just defining a sequence of all junctions: the given order in this case will not have a significant meaning. Without any loss of generality we will refer to one single corridor.

4.1 Non-linear optimization through genetic algorithm

As we have seen in 2.1 sopra, several methodologies can be used to estimate the value of the objective function for a given set of input parameters (the signal settings C, $\boldsymbol{g}$, $\boldsymbol{o}$). In our case, the estimation is performed by the simulator and then returned to the optimization routine. We aim to find optimal signal settings for a given set of junctions of a network for a given transport demand, optimizing a predefined objective function, for example the total delay. Due to the complex interactions of traffic flows in the network, usually the synchronization problem is a non-convex problem. Thus, finding a global optimum is not granted by the available optimization techniques. Moreover, the estimation of the objective function - whatever is chosen - is performed by the dynamic simulation model introduced in Chapter 3.1 sopra: in fact, no derivative information is available. Following the research stream introduced in 2.1 sopra, we mean to perform the optimization through a genetic algorithm. These are considered effective methods to determine sub-optimal solutions in *black-box* problems like the one considered here. This means that there is no warranty of reaching a global nor local optimum, but good-enough solutions are generally obtained in this way. The methodology proposed here extends the one already given in [34].

Genetic algorithms belong to the set of optimizers often applied in problems without derivative information or with non-convex objective function; in the latter derivatives would only have the effect to drive the optimization toward local optima. Their methodology determines sub-optimal solutions through a heuristic exploration of the space of solutions. It imposes to evaluate the many solution points selected during the algorithm, i.e. to assess a performance index, in this case the total travel time. Therefore, it is desirable that the fitness function requires considerably low computational times. Considering the trade-off between traffic model accuracy and its efficiency, the proposed traffic model is the best solution found.

Genetic algorithms take their name from Genetics, due to the parallelism among the general underlying idea and the biologic phenomena studied by the latter. In fact, they borrow many terms from this discipline. First, a set Б of feasible solutions is given, called *initial population*, composed by *n*Б solutions. Each solution is called *individual* and it is described by the values of optimization variables, called the *genes* of the given individual, which compose its chromosome complement З. Here every individual is a

set of signal setting data for the whole set of junctions, i.e. its genes are the cycle time and the green and offset of all signalized junctions and $n\mathfrak{Z} = 2 \cdot n\boldsymbol{X} + 1$. Each iteration is called *generation*, at the end of which only a subset of individuals are selected to breed the next generation. Individuals are generally selected for the next breed proportionally to a degree of suitability, given by a *fitness function*. The fitness function is the objective function of the optimization problem. Thus, similar to the Darwinian principle of the *survival of the fittest*, each generation a set containing individuals (generally) better than the previous is obtained. In other words, only solutions more capable to minimize traffic delays are allowed to "survive". Further details about genetic algorithms can be found in [4].

4.1.1 The initial population

The composition of the initial population set strongly affects the convergence speed of genetic algorithms. As the stop criteria is usually triggered when the algorithm does not manage to find significant improvements from the point of view of the population fitness, the worse the solutions are the longer the algorithm should be able to keep on working.

The initial population is often created by selecting $n\text{Б}$ random feasible solutions. As we have said above, this is not desirable. On the contrary, we breed it with maximum bandwidth solutions. As shown in [3], maximum bandwidth solutions are not optimal solutions for the minimum delay problem; despite this, they allow to start from quite good solutions. In fact, we can reasonably expect that any of the feasible objectives of the synchronization problem is improved by an improvement of the vehicle progression (total delay, throughput, externalities, …). The maximum bandwidth solutions are created through the MAXBAND algorithm based on equivalent systems from [56], further developed in [57]. As for all maximum bandwidth algorithm, this returns the optimal offsets for a given cycle time and the green splits of consecutive junctions along a path.

To gather $n\text{Б}$ solutions, different sets of input variables are evaluated. First, the cycle feasibility interval $[C_{min}, C_{max}]$ is split into equivalent intervals equal to:

$$\Delta C = (C_{max} - C_{min}) / (n\text{Б} /3) \tag{142}$$

Then, for the i-th cycle time value:

$$C_i = C_{min} + i \cdot \Delta C \tag{143}$$

the following three green split solutions are considered, for every junction:

a. pre-defined green splits (e.g., actual signal settings);
b. maximum green to the main stage (consistently to constraints);
c. green splits from Webster equisaturation rule (see Chapter 2.1 sopra).

Thus $n\text{Б}$ maximum bandwidth individuals are generated, then their travel time is evaluated by the traffic model, finally they are added to the initial population set.

4.1.2 The generation of new solutions

Starting from the initial population, genetic algorithms explore the feasible set generating new solutions through heuristic procedures and evaluating them. New individuals are obtained through *mutation* and *crossover* operations. Technicalities about these operations depend on the coding of variables: earlier algorithms relied on binary coding while later continuous values were assumed. We will not go into this, as this is not in the aims of this contribution. The proposed algorithm assumes that variables are real numbers, with explicit coding.

A mutation step consists in selecting one individual and modifying one or more of its genes. In the given algorithm, genes C, g_j and o_j, are expressed as integer numbers: each generic variable з$_j$ is randomly increased or decreased of a quantity between 0 and з$_{rng}$, where з$_{rng}$ is the maximum allowed variation. The mutated value is then adjusted to satisfy the boundary constraints.

In genetic algorithms mutation is usually performed by mutating each existing solution with a very low probability p_{mut} (usually $p_{mut} \in [0.001, 0.05]$). So the previous solution is lost. Differently, we produce a new set Ш of nШ new individuals by randomly selecting for each ш∈Ш an individual б∈Б and mutating each of its genes with probability p_{mut}. Thus during mutation nШ mutated copies of original individuals are spawned. This allows not to loose previous solutions and contemporaneously to control the mutation rate through the nШ parameter.

Let $u(x,y)$ be the function returning a random number between x and y with uniform distribution, we have:

$$\text{ш}(C) = \begin{cases} \min\{\max\{C_{min}, \text{б}(C) + u(-1,1) \cdot C_{rng}\}, C_{max}\} & u(0,1) < p_{mut} \\ & \forall\, j \in \boldsymbol{X} \\ \text{б}(C) & \text{otherwise} \end{cases} \quad (144)$$

$$\text{ш}(g_j) = \begin{cases} \min\{\max\{g_{min\,j}, \text{б}(g_j) + u(-1,1) \cdot g_{rng}\}, g_{max\,j}\} & u(0,1) < p_{mut} \\ & \forall\, j \in \boldsymbol{X} \\ \text{б}(g_j) & \text{otherwise} \end{cases} \quad (145)$$

$$\text{ш}(o_j) = \begin{cases} (\text{б}(o_j) + u(-1,1) \cdot o_{rng}) \bmod C_{mut} & u(0,1) < p_{mut} \\ & \forall\, j \in \boldsymbol{X} \\ \text{б}(o_j) & \text{otherwise} \end{cases} \quad (146)$$

A crossover step consists in selecting two target individual, denoted as *parents*, to generate one (two in some implementations) *son* individual(s). As for mutation, we spawn a new set Ж of nЖ individuals generated through crossover operations. For each generic gene з$_j$, the son-individual ж∈Ж randomly inherits the gene either from parent-individual б$_a$∈Б or from the other parent б$_b$∈Б:

$$ж(з_j) = \begin{cases} б_a(з_j) & u(0,1) < p_{crx} \\ б_b(з_j) & \text{otherwise} \end{cases} \quad (147)$$

where usually $p_{crx} = 0.5$ to make the choice balanced among the two parents. Checks are performed to produce feasible solutions. In our implementation, each crossover step two "mirror" sons σ_c and σ_d, are generated, i.e.:

$$ж_c(з_j) = б_a(з_j),\ ж_d(з_j) = б_b(з_j) \quad u(0,1) < p_{crx} \quad (148)$$

$$ж_c(з_j) = б_b(з_j),\ ж_d(з_j) = б_a(з_j) \quad \text{otherwise} \quad (149)$$

Individuals are selected as parents with a probability proportional to their fitness function: this is because better solutions are reasonably expected to breed individuals more fitting than the worst ones. In practice, inheriting a gene $з_j$ means assuming something about the signal behaviour of *j*-th junction (except for the 0-th gene, the cycle time). Thus, taking one junction traffic signal setting from a much better solution is supposed to return a better behaviour of that junction at least. Nevertheless, allowing a non-null probability to select worse individuals falls within the criteria of exploration of the solution set.

Similarly to the mutation process, generating the new set Ж allows to keep parent individuals and to introduce the parameter *n*Ж as crossover evolution rate.

When the mutation and crossover phases are done, a subset $Б^{i+1}$ of individuals is selected to produce next generation set, while elements of its complement-set to the previous generation set $Б^i$ are lost. In practice:

$$Б^{i+1} = \Sigma(Б^i \cup Ш^i \cup Ж^i) \quad (150)$$

where Σ is the survival function. The individual survival usually is performed randomly, assessing survival probability proportionally to individual fitness value. *Elitism* gives surviving probability equal to 1 to the best ζ solutions: this grants the best found solutions to be never lost.

5 A tool for signal synchronization: TOSCA

The methodologies introduced so far above led to their integration into a unique software package called TOSCA (Timing Optimization under Spillback Congestion along Arterials). TOSCA incorporates both the GLTM and the genetic algorithm. The input data (options, transport system, optimization) can be stored in a MySQL or PostgreSQL database, compatible with the structure described in Attachment A. Usually a modeller needs not only the input data but a more versatile tool, such as a network editor. Considering this, TOSCA also integrates into VISUM ([61]), a transportation planning software for network and demand modelling. The VISUM package has been chosen due to the high level of detail of its junction modelling. The figures in Chapter 3.2.2 sopra were taken from its junction editor.

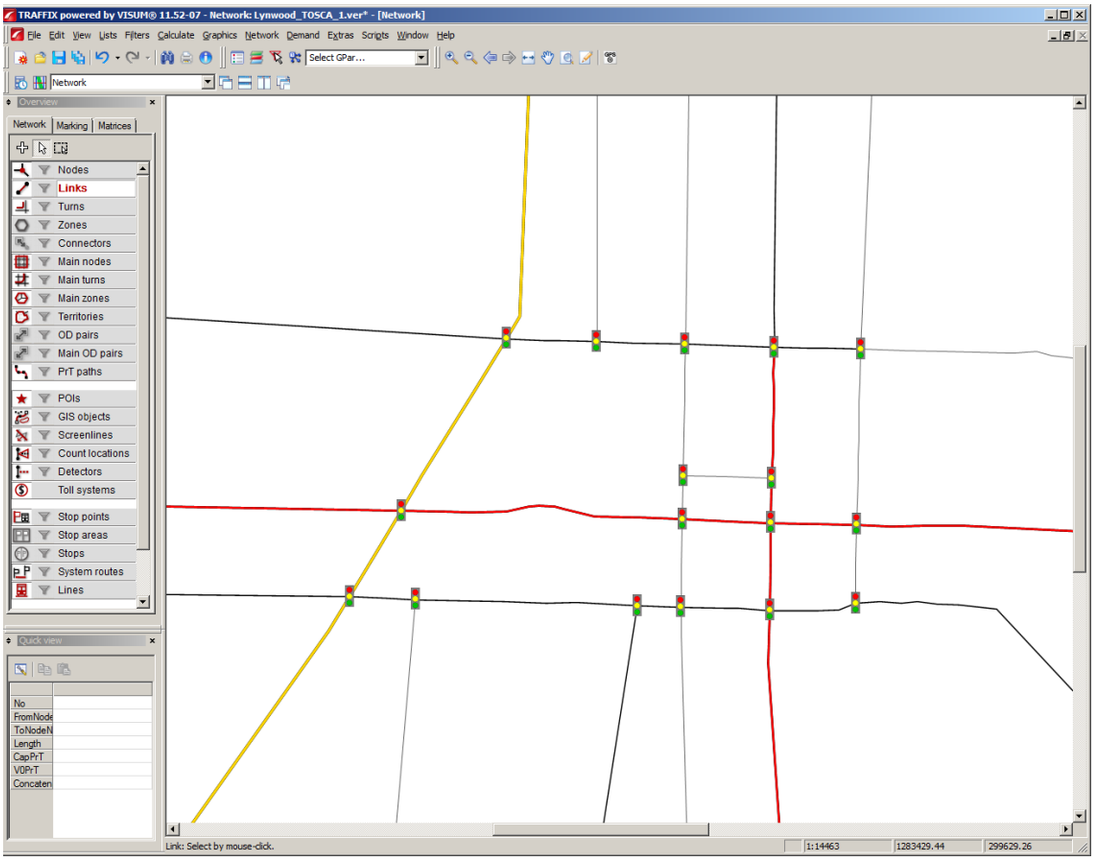
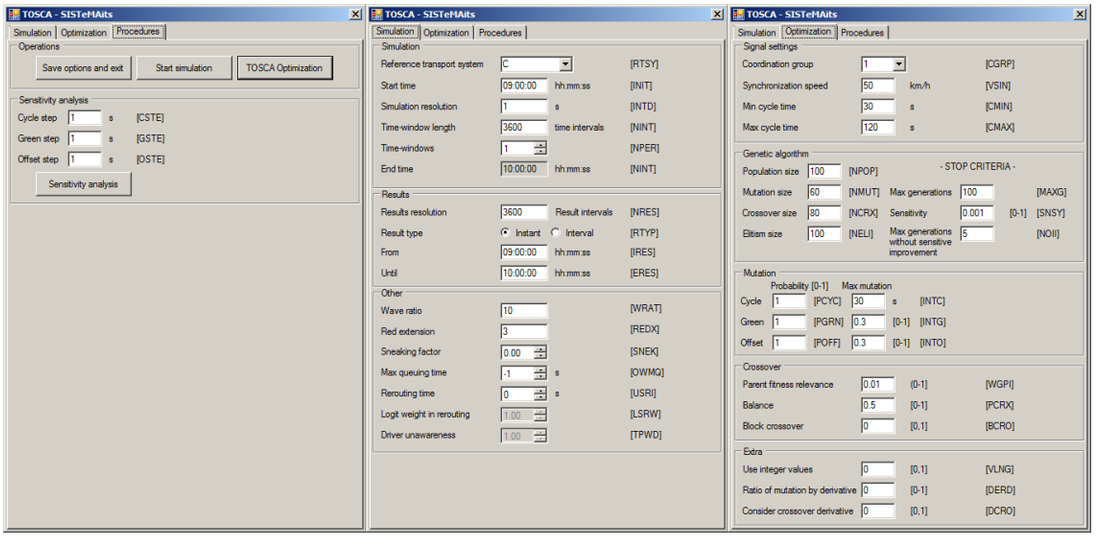


Figure 5.1: An example of the GUI of VISUM, with a signalized network.

VISUM allows external programming through Visual Basic scripts (vbscripts). Two vbscripts are part of the engine of TOSCA. The first one needs to be run just once per model to extend the VISUM model to comprehend deeper aspects, for example the conflict areas and their respective priority or to specify the main stage of each signal controller. The other is the arbiter of the protocols among VISUM and TOSCA.

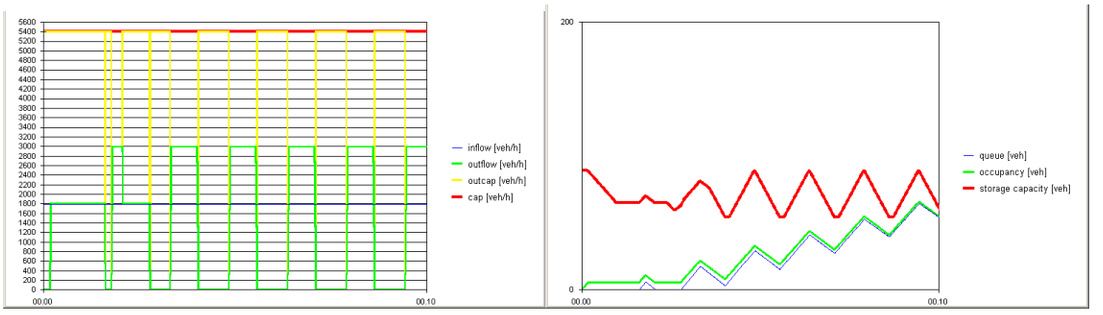
Communication take place through text files: VISUM can export the complete network and demand databases into text files and TOSCA has a module which imports them, according to VISUM standards. To accomplish the file data exchange a suitable GUI has been developed (Figure 5.2). From the procedure GUI (Figure 5.2a) the user can decide the procedure to execute.



a. **b.** **c.**

Figure 5.2: The GUI of TOSCA: procedure (a), simulation (b) and optimization (c) consoles.

On one hand, as TOSCA incorporates the GLTM, the dynamic network loading can be run alone, to have an insight of the dynamic behaviour of the system in the non-intervention scenario. The user can decide the time interval to simulate, the discretization of the simulation, the aggregation of its results and several parameters which influence the simulation, e.g. the environmental sneaking factor.



a. **b.**

Figure 5.3: The alternative result interface, displaying the dynamic chart of flow (a) and queue (b).

There are two ways to display the results of TOSCA simulation. First option, TOSCA generates a text file which can be (automatically, through the vbscript) reimported, to make the simulation dynamic results available in the VISUM interface. This allows the

VISUM modeller to operate with this data into the environment the most familiar to him. Second option, a rougher but more detailed interface can be launched by TOSCA at the end of the procedures (Figure 5.3). The greater level of detail consists in displaying graphs with the most of the link model variables: capacity, exit bottleneck, inflow, outflow, number of vehicles, queue, storage, travel time. By comparing these the analyst can truly go deep into the phenomena of traffic dynamic evolution.
On the other hand, the optimization can be performed on all the signal controllers which belong to the given coordination group. Control parameters of the genetic algorithm can be set, such as the population dimension, mutation and crossover rate, probabilities and ranges, stop criteria and other theoretical parameters used during experimental tests. As the simulator plays the role of a *black-box* for the genetic algorithm, it is also possible to choose the optimization objective function among a set of data the simulator can return. Currently, the available options are:

a. minimization of total travel time[1]: $\Omega(C, \boldsymbol{g}, \boldsymbol{o}) = \sum_{a \in \boldsymbol{A}} \nu_a \cdot t_a(C, \boldsymbol{g}, \boldsymbol{o})$;
b. maximization of total network throughput: $\Omega(C, \boldsymbol{g}, \boldsymbol{o}) = \sum_{z \in \boldsymbol{Z}} \nu_z \cdot E_z(\mathrm{T})$;
c. minimization of total externality mix: $\Omega(C, \boldsymbol{g}, \boldsymbol{o}) = \sum_{a \in \boldsymbol{A},\, \hat{e} \in \boldsymbol{E}} \nu_{\hat{e}} \cdot PE_{\hat{e}\,a}(C, \boldsymbol{g}, \boldsymbol{o})$.

Assuming a coefficient ν_a for link travel time allows to assess priority to some links of the network. For example, progression along an arterial can be maximized giving a positive weight to its links and 0 to all other links of the network. In the case of the zone coefficient ν_z, it can improve the throughput in some prior areas of the network. In the case of the externality coefficient $\nu_{\hat{e}}$, it allows to set a priority in the considered externalities.
While the genetic algorithm is running, a dynamic chart is displayed to the user. In this, the original solution (red), the best maximum bandwidth solution (green) and the overall best solution (violet) objective function values are plotted (Figure 5.4).

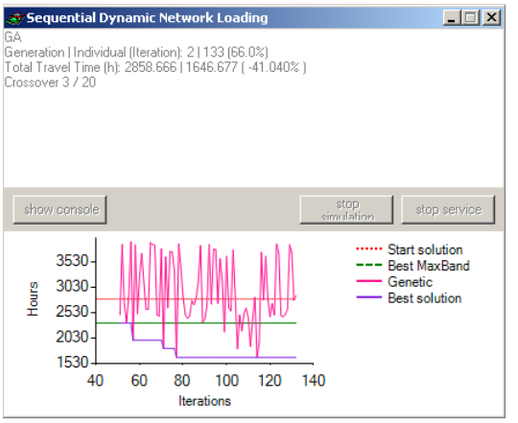


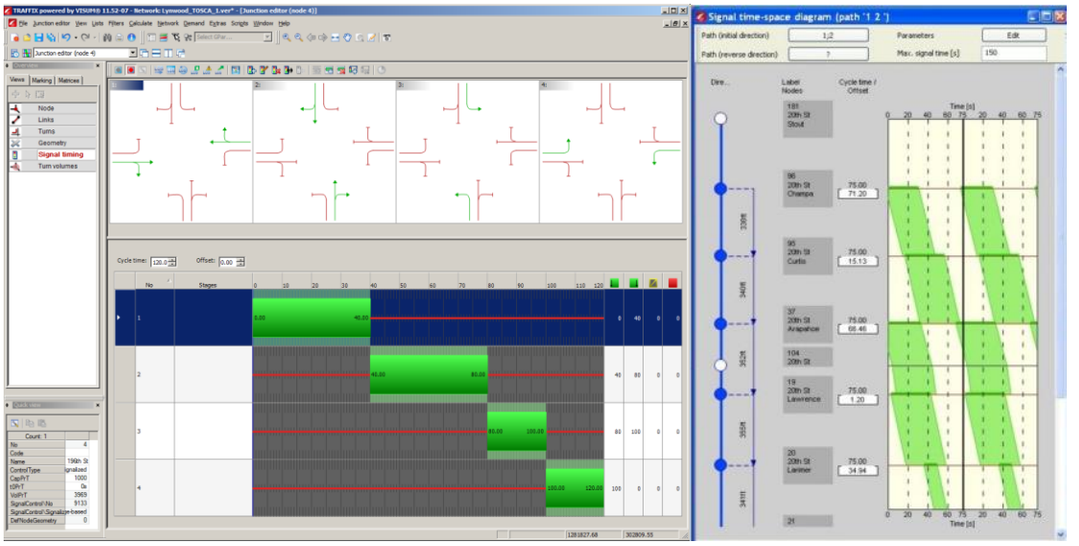
Figure 5.4: The control chart during the genetic algorithm run.

[1] The total travel time in the network of a vehicle can be expressed as the sum of the free-flow travel time on the non-signalized network, which is constant, plus an additional delay due to the interactions between vehicles and traffic signals. Thus, total delay minimization and total travel time minimization are equivalent problems.

Beside these, the travel time of each of the evaluated solutions is added (pink): when the algorithm reaches a probable global optimum the value of these start converging and no significant improvements are found. In this case, the user can stop the algorithm manually, without waiting for the satisfaction of any of the stop criteria.
When the optimization stops, three solutions are then exported into the database, in terms of tables with signal IDs and timings. These are:

- the original solution;
- the best maximum bandwidth solution found;
- the best overall solution found by the genetic algorithm.

If the user works from VISUM, a text file containing the optimal signal settings according to the VISUM format is exported by TOSCA. Thus, the modeller can import into VISUM the optimal settings and overwrite the previous existing ones. In Figure 5.5a an overview of how the signal settings are displayed in VISUM is illustrated. Another advantage of reimporting signal data in VISUM is that there a time-space diagram feature is already available (see Figure 5.5b).



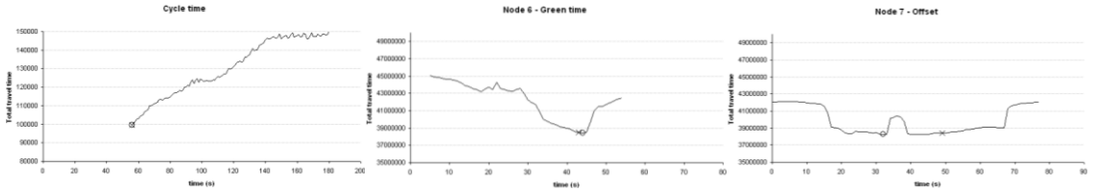
a. **b.**

Figure 5.5: The signal editor in VISUM: Stage timings (a, lower screen), stages (a, upper screen) and space-time diagram in VISUM (b).

The input-output architecture of TOSCA, for both cases of VISUM and database, is illustrated in Figure 5.7.
Finally, a sensitivity analysis routine has been added to TOSCA. Starting from the non-intervention hypothesis each of the variables, one by one, is increased/decreased of the input step (in seconds). This allows to plot the profile of the objective function along each of the dimensions of the problem, i.e. the cycle time and the green and the offset of each of the synchronized intersections. The routine can be used starting either from the original solution or from the genetic optimal solution. This allows several scopes:

a. to test the problem complexity, for scientific purposes;

b. to check the relevance of the several analyzed junctions in the overall system;
c. to check if the final solution is a local minimum or how far it is;
d. to perform a final hill-climbing adjustment procedure on the final sub-optimal solution.



a. **b.** **c.**

Figure 5.6: Three plot diagrams of the sensitivity analysis routine in TOSCA: cycle (a), green time (b) and offset (c).

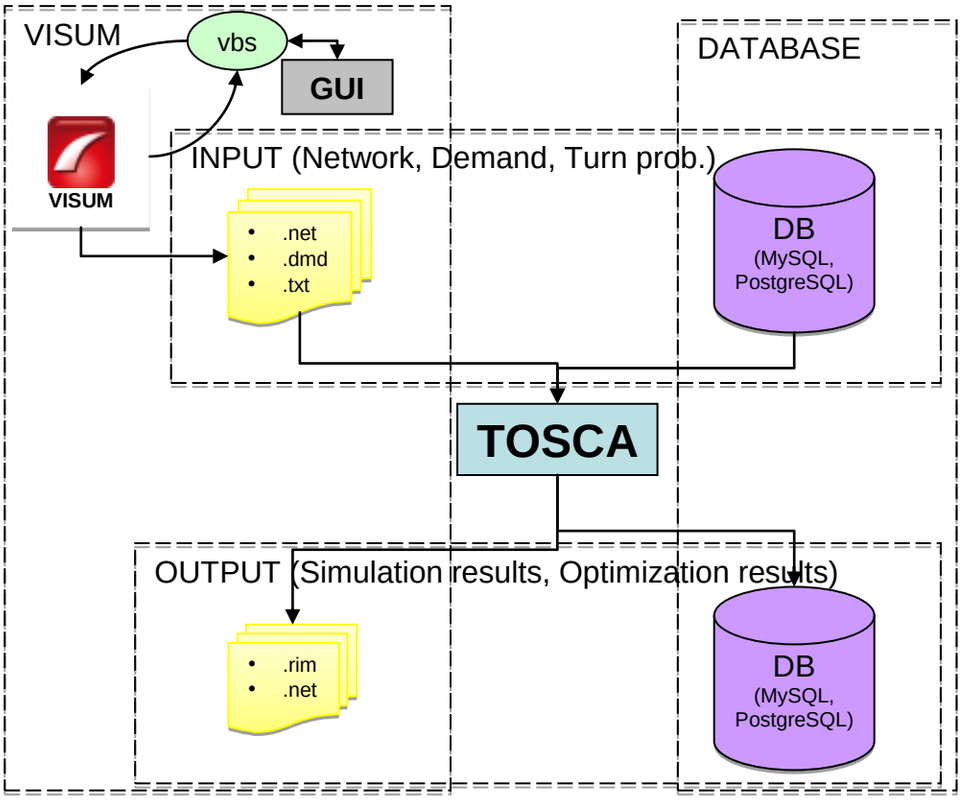


Figure 5.7: The architecture of TOSCA.

6 Results

Results are given in two sections. In the first section, results in the fitting among the traffic model and experimental data are illustrated. To retrieve the experimental data we chose to use a microsimulator. This choice has been done for several reasons. First, because the microsimulators allow to gather a large set of data, with any time discretization, without great efforts neither in time nor expense. Second, because we were interested in testing the goodness of the optimal signal setting solutions in a different context from the GLTM itself, which was already used by the optimization algorithm. Implementing on the field a signal setting solution to gather data about its goodness was unfeasible. So, the microsimulator has been used as the real-world of reference both for calibration of the traffic model and for comparison of the optimal signal settings. The microsimulator which was used is VISSIM ([62]), which already integrates with the output of VISUM (in terms of supply and demand data) and allows an easy switch of the work from one to the other.
The test network was Lynnwood, a city 20 km far from Seattle, US. The network was composed by 78 links, 16 zones and 17 signalized junctions, disposed in a grid where 5 distinct bidirectional corridors could be identified, one of them traversing all the others. The network is illustrated in Figure 5.1 sopra.
Most of the charts shown in current chapter were made through XRES, an Excel Workbook suitably developed to import and elaborate data from VISSIM, VISUM and TOSCA.

6.1 The calibration of the traffic model

The issue of calibrating the fundamental diagram parameters has been carried out by plotting the theoretical shapes of the fundamental diagram, according to the set of parameters, subject of the calibration and matching these with the microscopic data. It is possible to do the same with real field data. The resulting charts are illustrated in Figure 6.1 for a triangular-shaped fundamental diagram.

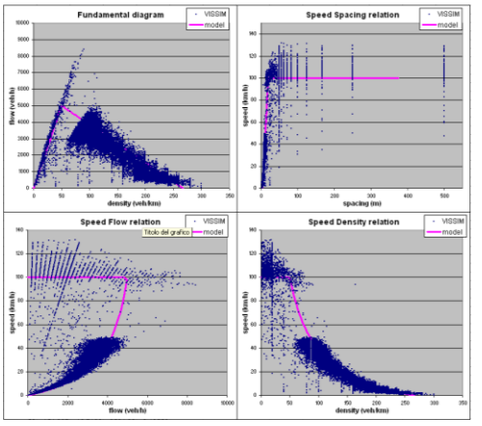
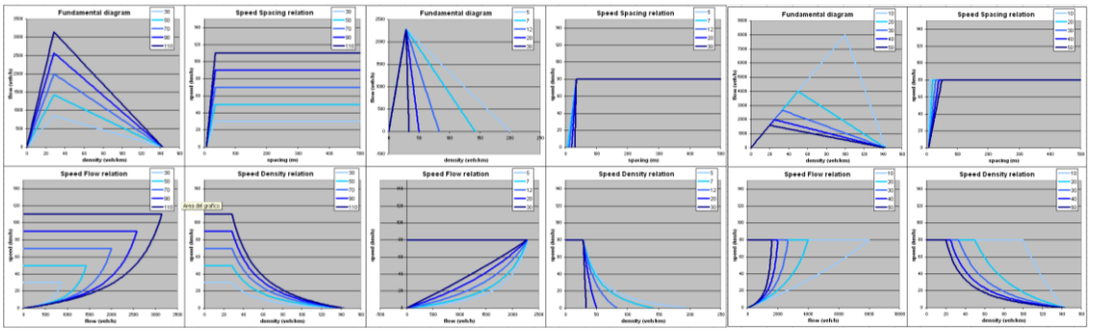


Figure 6.1: Matching field data with theoretical shapes of the fundamental diagram.

Four shapes are shown:

1. flow-density diagram (the most common), it is particularly suitable to set the jam wave speed W;
2. speed-spacing diagram, it is particularly suitable to set the critical density K;
3. speed-flow diagram, it is particularly suitable to set the jam density J;
4. speed-density diagram, it is particularly suitable to set the freeflow speed V.

The other parameters were derived from these. In Figure 6.2 it is illustrated how the parameters affect these shapes:



a. **b.** **c.**

Figure 6.2: The effects of different parameters on the fundamental diagram: V, W (a), J (b), K (c).

The conflict area model introduced in Section 3.2.3 sopra has been successfully tested on non-elementary networks (~100 links, ~20 zones, 3 hours simulation, 1 second time-interval). Now we are going to illustrate the results in the case of a left-turn manoeuvre intersecting an opposing straight crossing flow. The left turn flow is 2000 veh/h, close to the link and turn capacity of 2100 veh/h, while the straight crossing flow has a demand level which increases constantly, by 200 veh/h every 5 minutes. The illustrated charts are the detailed charts of the rough interface mentioned in Chapter 5 sopra.

In the non-conflict case (a) the flows do not interact, and the 2000 veh/h flow level is unaltered for all the one-hour period of simulation. By adding a conflict area among the two turns and assessing the same priority to them (b), the left-turn flow keeps its original level until the total flow is greater than the conflict area capacity (which, we remind, is the greater of the two link capacities). As soon as this occurs, the demand flow drops and the capacity is equally shared among the two flows. In fact, the straight turn flow grows until reaching the same share than the left turning flow.

In a real intersection, the straight movements usually have priority while the vehicles turning left are delayed until the flow rate of the straight movement allows them to cross it. So, we apply a null priority ($\pi_{ab} = 0$) to the left turn in the conflict area. In this case (c), the left turn flow drops down coupled with the increase of the straight flow, which is exactly what expected. The total flow is always equal to the conflict area capacity.

In Figure 6.3 sotto the charts of the three distinct cases above are given.

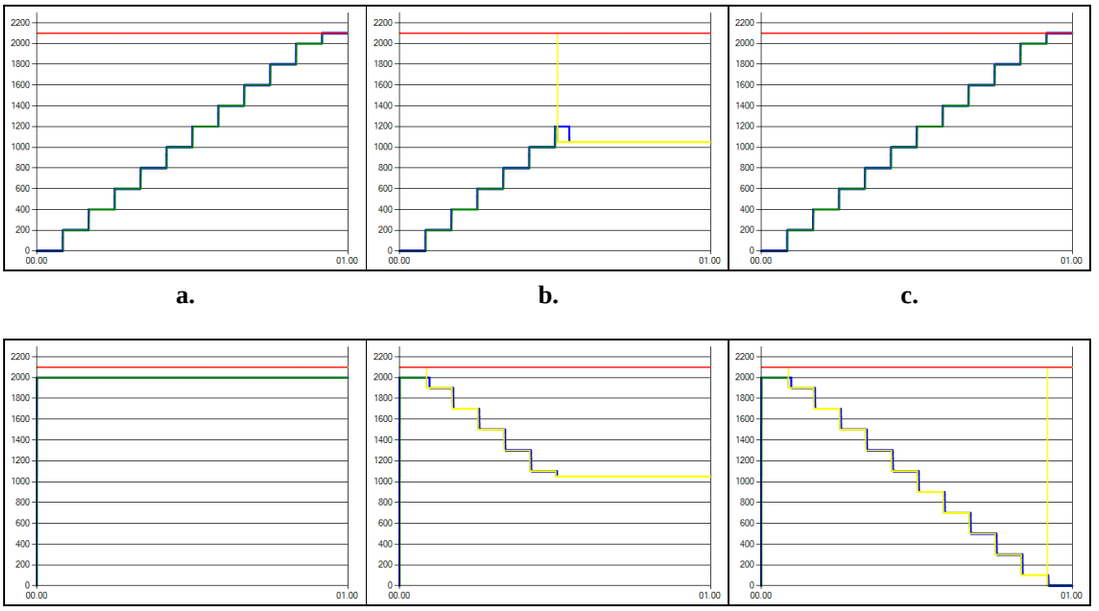


Figure 6.3: The saturation flows of a straight movement (upper charts) and a left turn movement (lower charts) in case of no conflict (a, d), conflict without priority (b, e), yielding left turn (c, f).

In Figure 6.4 the behaviour of the intersection among a straight movement and a left turn is illustrated. On the x-axis we have the priority flow, while on the y-axis we have the yielding flow. The red line shows that our model linearly approximates the phenomenon. The blue dots of the microsimulator show that in reality, a yielding flow is even more delayed by the priority rule than linearly. More, the yielding flow slows down (and thus decreases its saturation flow) even if the opposing flow is null, to safely approach the intersection. This is shown by the value of the intersection of the blue line with the vertical axis, which is smaller than the link capacity (2100 veh/h).

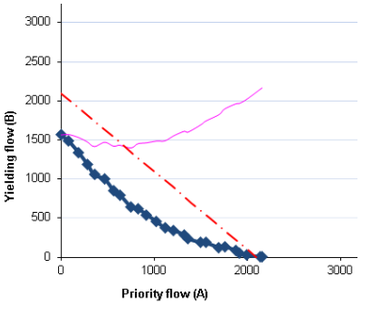
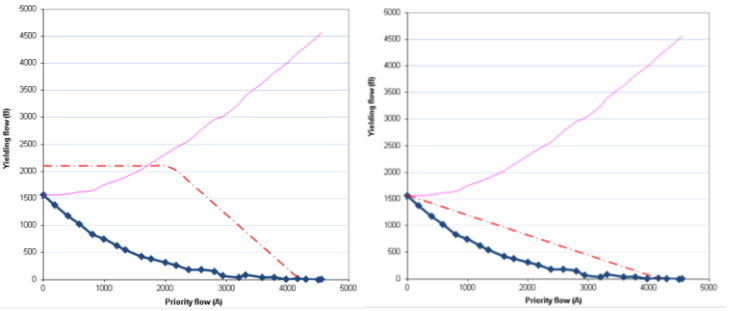


Figure 6.4: The saturation flow of a left turn movement crossing a straight movement with respect to the saturation rate of the latter.

Furthermore, when the two links have different capacity, an undesired behaviour is registered, which is illustrated in Figure 6.5a. The solution to correct both the latter and the zero saturation flow is to apply a turn capacity reduction to the delayed manoeuvre,

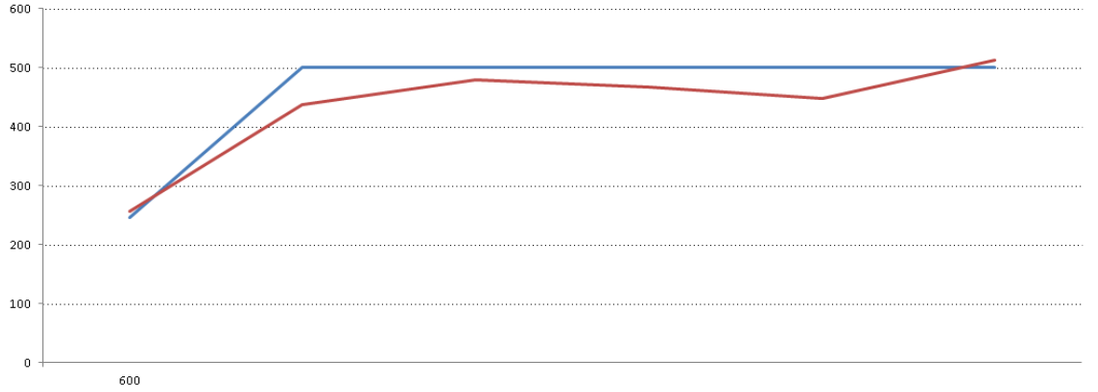
as an effect of the presence of a conflict. If none of them has priority, both will be reduced. The effects of this are shown in Figure 6.5b.



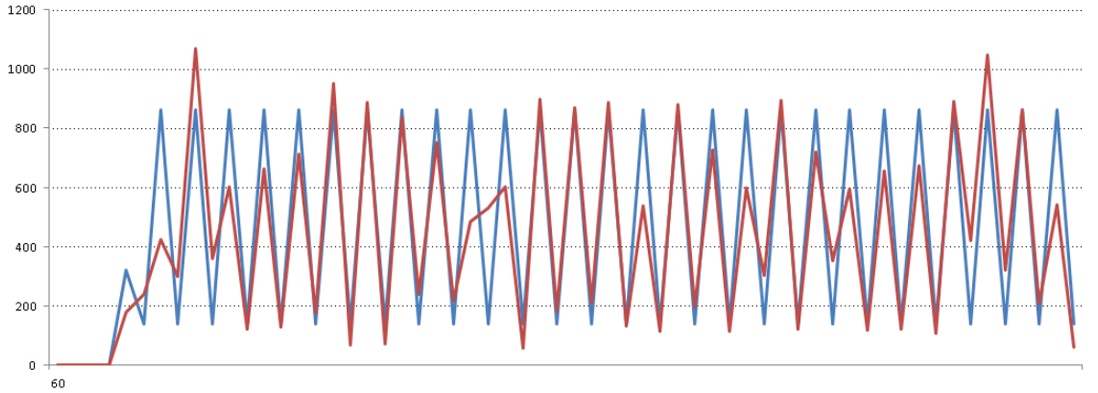
a. **b.**

Figure 6.5: The saturation flow drop of a delayed manoeuvre (a) and the result of the application of a capacity reduction coefficient (b).

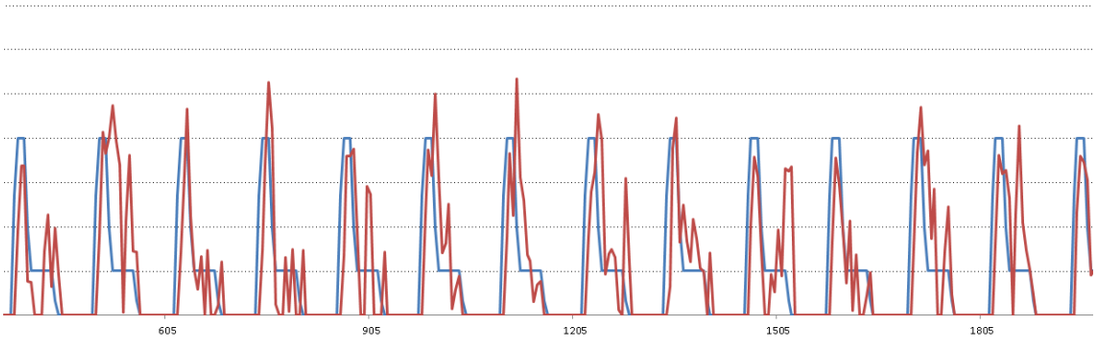
The overall effects of the calibrations above can be appreciated by giving a look at the link dynamics. In the following pages we see the comparison of the sensible variables (flow, speed and density) between the macroscopic and the microscopic model on a link in an analysis period of one hour, with an aggregation of 10 minutes (a), 1 minute (b), 5 seconds (c).



a.

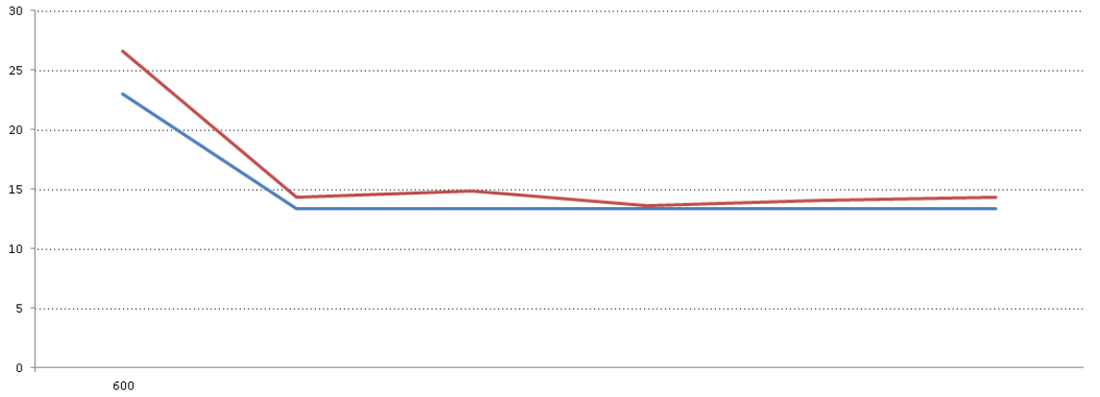


b.

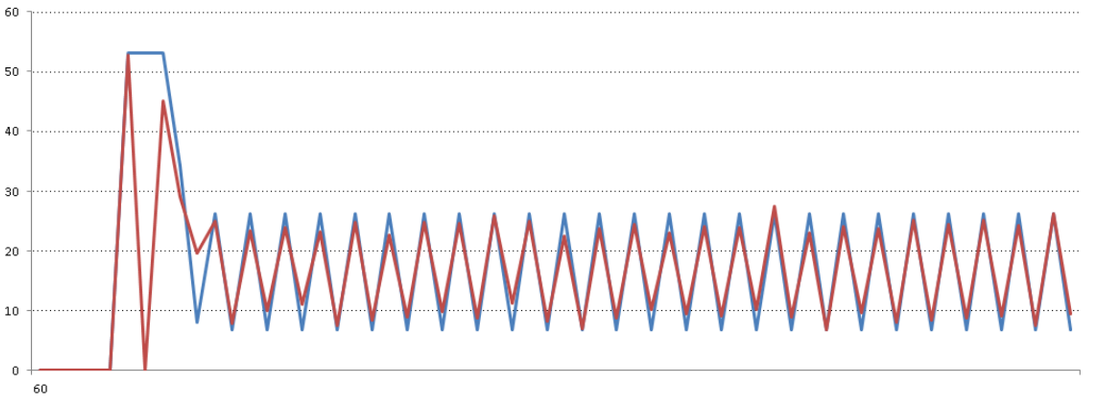


c.

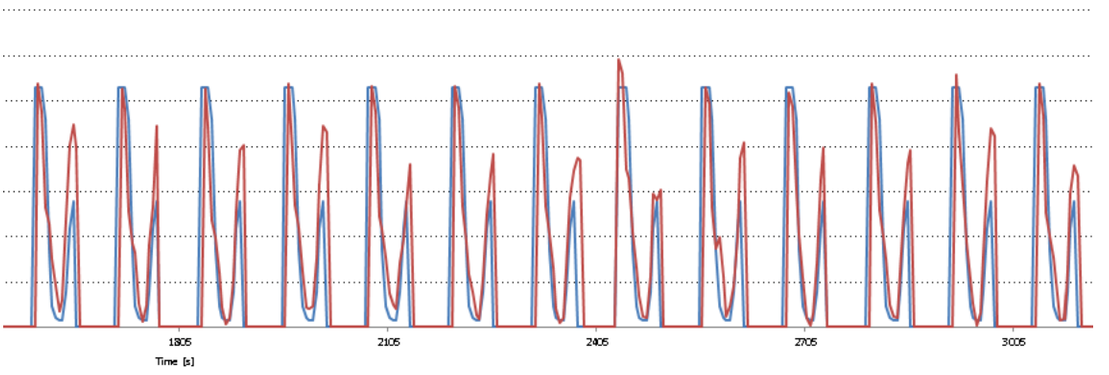
Figure 6.6: The dynamic evolution of the flow variable on a link after calibration. In blue the macrosimulator, in red the microsimulator. Aggregation by: 10 min. (a), 1 min. (b), 5 sec. (c).



a.

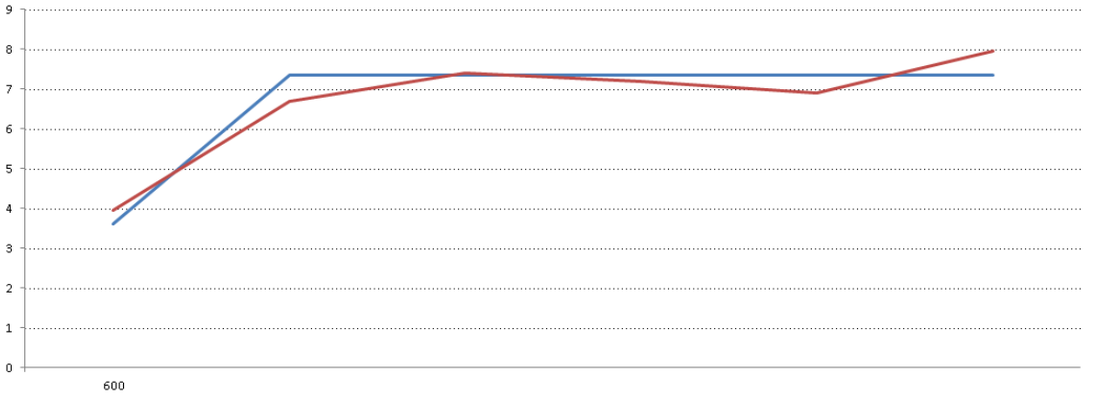


b.

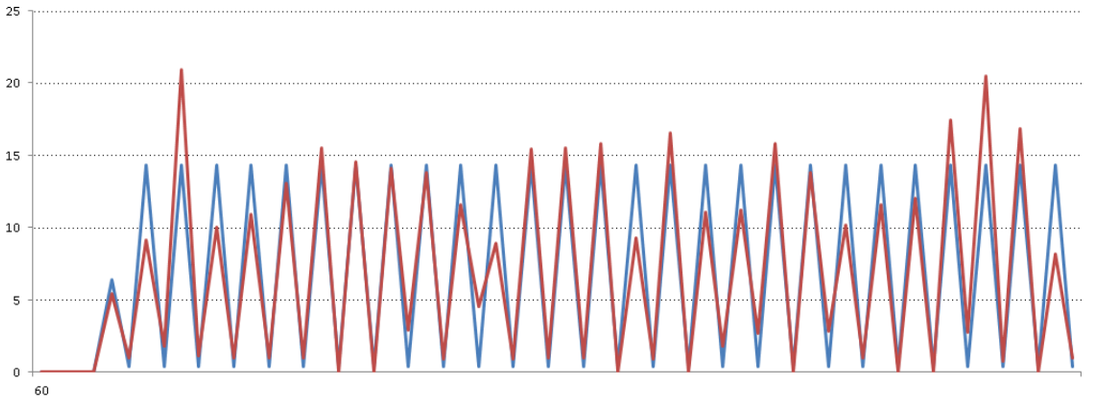


c.

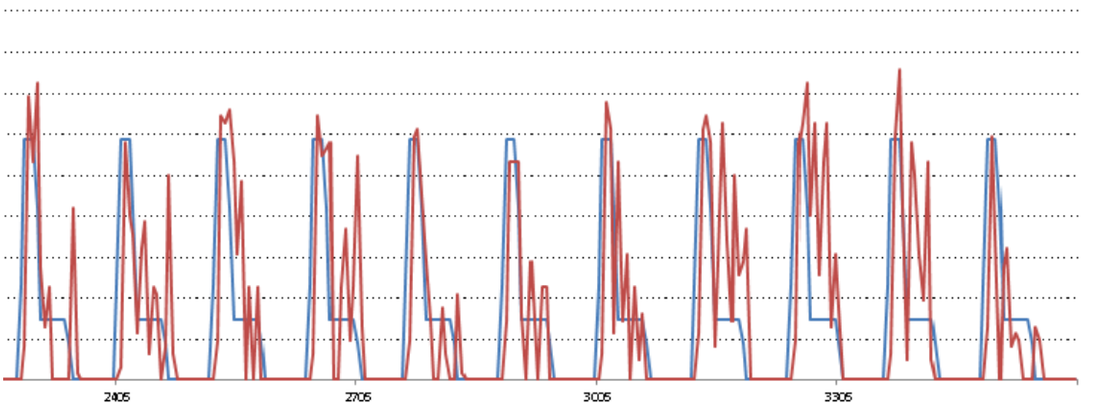
Figure 6.7: The dynamic evolution of the speed variable on a link after calibration. In blue the macrosimulator, in red the microsimulator. Aggregation by: 10 min. (a), 1 min. (b), 5 sec. (c).



a.



b.



c.

Figure 6.8: The dynamic evolution of the density variable on a link after calibration. In blue the macrosimulator, in red the microsimulator. Aggregation by: 10 min. (a), 1 min. (b), 5 sec. (c).

Results are even more evident if comparing the total amount of the variables over all links, with the maximum aggregation of 10 minutes (Figure 6.9). This returns an insight of the levels over the whole area and their deterioration in presence of congestion.

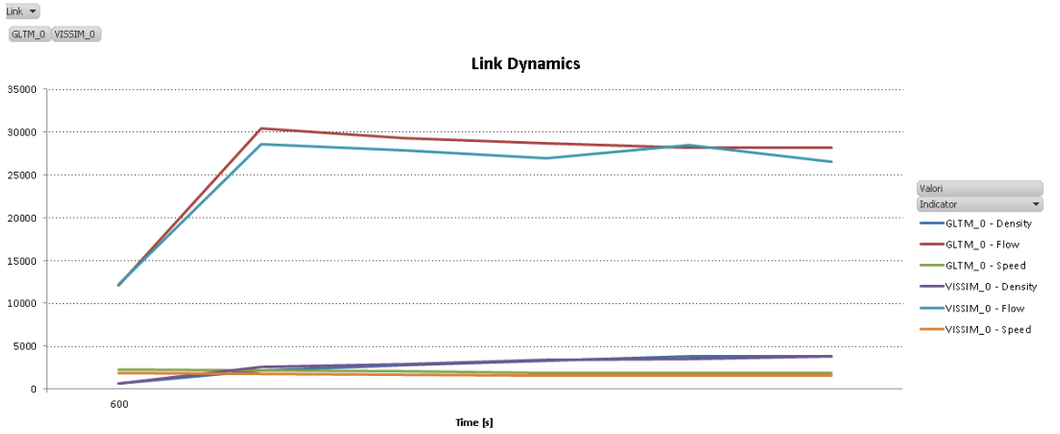


Figure 6.9: Comparisons among the total values of flow, speed and density over the whole area.

Figure 6.10 shows that on the links where congestion occurs, the instant when this is triggered, marked by the vertical red line, is correctly caught by the macrosimulation.

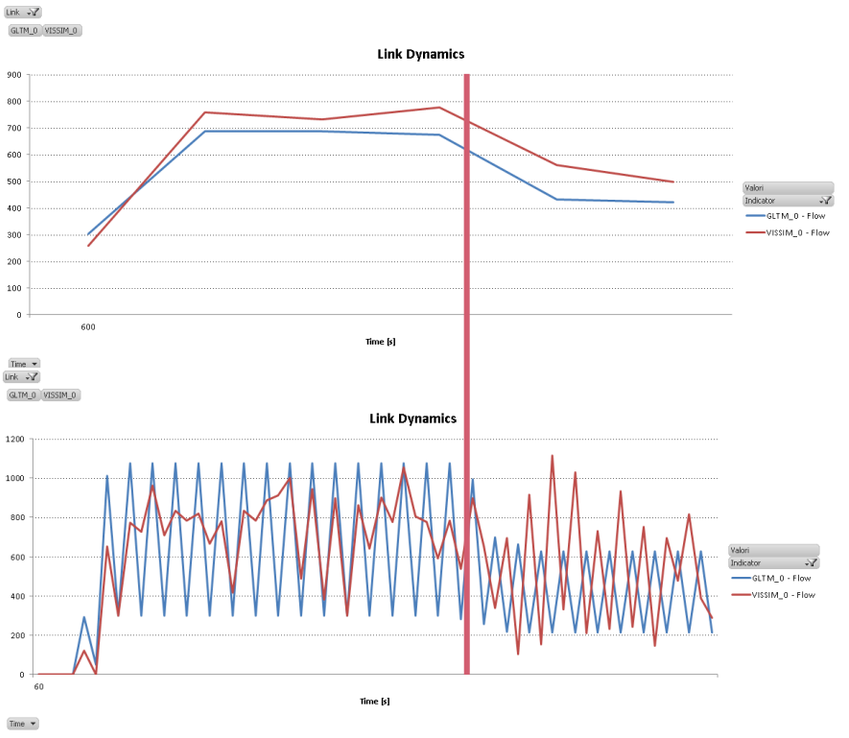


Figure 6.10: The instant when congestion raises up on the given link, depicted by the red line, is comparably the same both in the microsimulation and in the macrosimulation.

6.2 The signal setting optimization

With such fitting results between the microsimulator and the macrosimulator used for the optimization algorithm, we expect the same trend among the optimal solutions and the test network. In this case, the XRES Workbook has been used to import data of several microsimulations, each one implementing an alternative signal setting solution. The optimization has been evaluated through several indicators:

a. total travel time of all vehicles in the network;
b. throughput of the network;
c. travel time along predefined sections in the network (in real world this can be checked through plate-recognition cameras);
d. queue length;
e. number of stops.

Differently from what happened before calibrating the network, after calibration all the solutions proposed as optimal and then tested in the microsimulator showed a coherent improvement, both in the case of a single corridor and of the whole network. In the former case, the optimization finds good solutions within 10-20 minutes of run. This is coherent with our expectations: in the case of the corridor the maximum bandwidth solutions are already good solutions, thus a fastest convergence is expected. In this case, the most suitable indicator is the section travel time. In Figure 6.11 the chart shows the travel time along a given section, as a function of the entry time. Note that at the end of the chart values drop as the latest vehicles which entered the section did not exit before the end of the simulation, so values are missing.

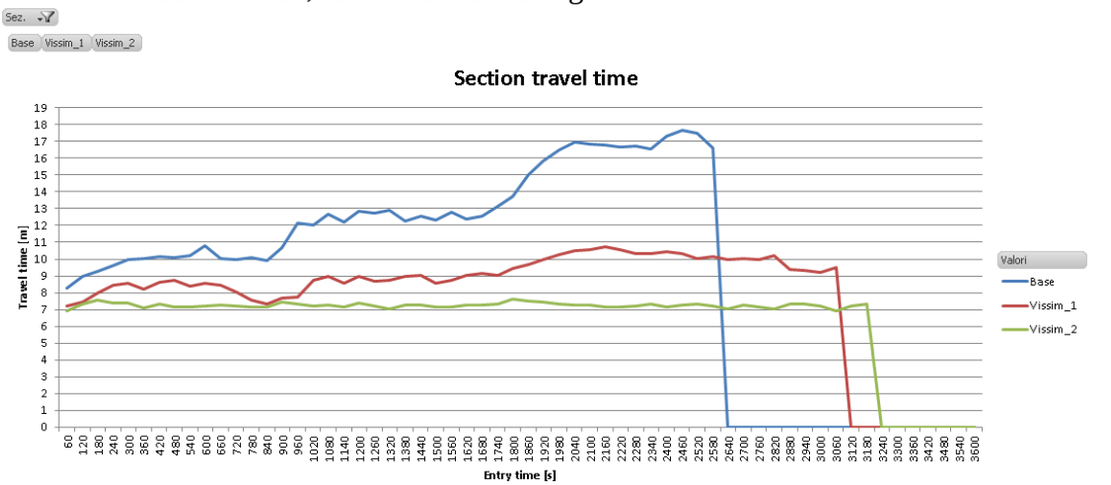


Figure 6.11: The travel time over a section of a corridor, as a function of the entry time. Lines depict the original solution (blue), and the optimal solutions considering the travel times of side approaches to the corridor (red) and neglecting them (green).

The value of the original solution is given in blue. It shows that current settings generate congestion, as the section travel time is constantly increasing. The red line refers to the optimization considering the total travel time over all the links of the subarea of the arterial corridor, comprehending the side approaches. The optimal solution found by the

algorithm significantly improved the section travel time of up to about the 45%. The vehicle progression along the corridor can be further improved by neglecting the travel time on side approaches, i.e. setting the travel time multiplier $\nu_a = 0$ for each of the links not belonging to the arterial. In this way, the objective function mostly coincides with the considered indicator. In this case (green line), the solutions improves the travel times along the corridor of up to the 60% but vehicles on the side approaches are probably penalized. This proved the effectiveness of the objective function coefficient in the proposed case of progression maximization along one arterial.
In the case of the synchronization of the whole network, the optimization can require more than 1 hour before stopping. In fact, the maximum bandwidth maximization in this case were less significant, thus worse solutions, thus the worse initial population set implies a slower convergence of the algorithm. Nevertheless, the final optimal solution reduced 10.69% the travel time of the total area and improved its throughput of 14.75%. Section travel time is not a suitable indicator in this case: global indicators are required to evaluate the optimization over a widespread area. Total travel time and throughput are significant indicators in this case.

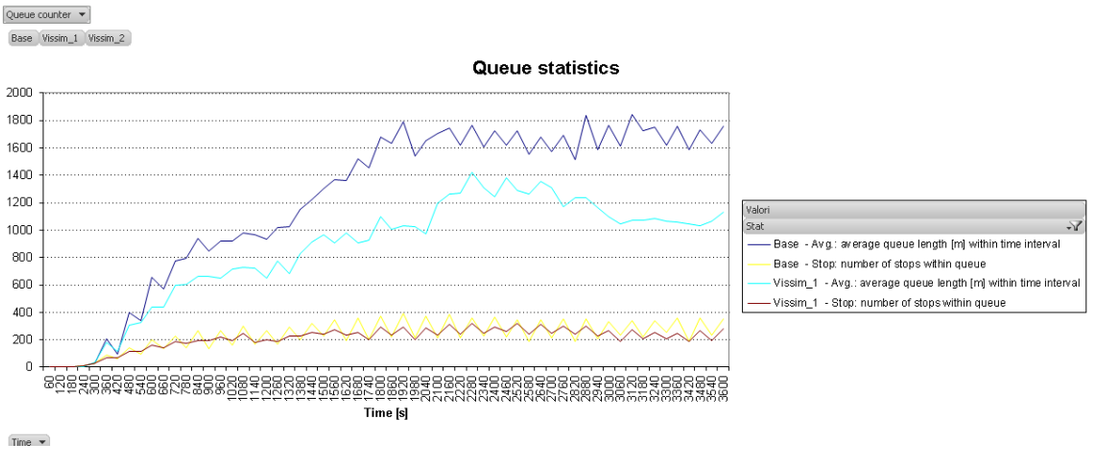


Figure 6.12: The total results from queue counters all over the network. Average queue length (blue and light blue) and number of stops (yellow and brown) are displayed.

The microsimulator allows to gather information about queue lengths and number of stops in the wished links in the network. We remark that considering individually these can lead to wrong assumptions, because a queue reduced in some spots could have just been displaced in some other place in the network. In this case, at most a total value over the whole network can be used. This is displayed in Figure 6.12: the sum of the average queue on all the queue counters over the entire network was reduced thanks to the proposed optimal solution for the network case.

7 Conclusions

The proposed methodology showed consistent results both in terms of traffic modelling and optimization results. The new models proposed to extend the GLTM significantly improved its representation of dynamic traffic phenomena, producing macroscopic results that are comparable to those of a microsimulation, without its computational efforts.

The proposed fundamental diagram equation allows a great versatility and addresses a large number of parameters, larger than the most common fundamental diagram equations. Nevertheless, the number of calibration parameters is much smaller than that of any microsimulator. As a consequence, the calibration process is easy and reliable. Moreover, it was proved that a calibrated model fits field data. The detailed junction model represents consistently the queuing along several approach lanes of a signalized junction and the separate spillback of the queues. The conflict area model introduced promising aspects in the representation of the complex interactions that occur among conflicting manoeuvres in a junction. In practice, the model could be suitably used to perform the following evaluations:

- the effective turn saturation flows;
- the obstructions to crossing movements which occur when in presence of queue spillback up to the upstream junctions;
- the effects of introducing a traffic light on a junction where the resolution of conflicts was previously left with the drivers' rationality;
- the stage sequence when in presence of permitted manoeuvres.

The presented lane changing model does not address the explicit lane representation, so trajectories of vehicles across several lanes are not available. However, it allows to fit the flow levels registered in real traffic when in presence of pre-emptive lane choice behaviour by the users according to the desired manoeuvre at the end of the link. The multicommodity model actually opens the GLTM to the world of ITS applications and extends all previous models to the case when more user classes are available in the system. Its definite generality permits to represent a large range of vehicle types and driver behaviours without introducing any of the strong constraints required by the most common multicommodity models in KWT. It directly allows to represent the urban traffic mix, including the case of smaller vehicles sneaking in the jam (e.g. scooters in car queues). A mixed queue model with scalable complexity has been proposed to allow either a realistic multiclass queuing or an efficient modelling of it. Finally, the externality model introduces the calculation of both emissions and safety indices in a dynamic macroscopic model, according to the COPERT IV European methodology.

The optimization algorithm has found to be suitable for several scopes and contexts, according to the models introduced above. The fitness of the traffic model grants a high truthfulness of the simulation and thus of the optimal solution. The conflict model represents the blocking back of junctions and allows to perform an optimization capable to avoid it. Moreover, it allows the optimization of the stage sequence if the problem formulation is further extended in order to take it into account. Finally, the externality

model allows to optimize the signal settings with respect to an objective of minimum emissions or maximizing safety indices.
Furthermore, the versatility obtained by integrating the whole of models into a commercial software such as VISUM allows to exploit the latest developments reached in the industrial world, the detail of the VISUM network model and gives significant visibility to the academic results obtained until now. A special effort was done in making of TOSCA a usable tool for both beginners without detailed notions of traffic theory and skilled modellers.

7.1 Future developments

The results obtained so far have already outlined what will be the future developments for the illustrated methodologies. On one hand, the traffic model can be further improved. The conflict area model can be fine-tuned through the capacity coefficients: scientific results should validate such hypothesis. A significant improvement in the link model has already started, investigating about the acceleration of traffic flows. Introducing constraints for the several classes, in terms of maximum and minimum allowed acceleration, would lead to a model that overcomes the limits of a first-order implementation of the KWT, reproducing more realistic vehicle trajectories, lost times and other second-order phenomena. Furthermore, the study of fundamental diagram shapes which consider the theoretical phenomena of the capacity drop and of the hysteresis would allow significant improvements both in the ITS applications through the GLTM and the optimization strategies taking these phenomena into account. Finally, it was noted that sometimes the experimental traffic data show convex fundamental diagrams. Convexity is a condition obtainable by the proposed polynomial model of fundamental diagram. It is interesting to investigate about its theoretical implications in the KWT.
On the other hand, the optimization engine could be further validated. An extension which promises significant improvements is the introduction of the distance-based elimination of individuals inside the population set. The introduction of a distance metric would allow to avoid keeping similar solutions with comparable values of fitness. This would significantly enhance the exploration of the feasible set and it possibly avoids the premature convergence to local minima of the problem. From a theoretical point of view it would be interesting to couple the simulator with alternative stochastic optimization methods, e.g. the tabu search and the swarm methods. A final comparison among the optimization results obtained by TOSCA and the optimal settings returned by the commercial software currently available would be a fascinating challenge for the academic community.

8 References

[1] Abu-Lebdeh, G., Benekohal, R.F., 2003, *Design and evaluation of dynamic traffic management strategies for congested conditions*, Transportation Research Part A 37, 109-127.

[2] Allsop, R.E., 1968, *Selection of Offsets to Minimize Delay to Traffic in a Network Controlled by Fixed-Time Signals*, Transportation Science 2, 1-13.

[3] Bavarez, A., Newell, G.F., 1967, *Traffic signal synchronization on a one-way street*, Transportation Science 1, 55-73.

[4] Bodenhofer, U., 2003, *Genetic Algorithms: Theory and Applications*, Johannes Kepler University, Lecture Notes.

[5] Cantarella, G.E., Vitetta, A., 2010, *La regolazione di intersezioni stradali semaforizzate. Metodi ed applicazioni*, Strumenti per l'analisi dei sistemi di trasporto 1035.13.

[6] Ceylan, H., Bell, M.G.H., 2005, *Genetic algorithm solution for the stochastic equilibrium transportation networks under congestion*, Transportation Research Part B 39, 169-185.

[7] Ceylan, H., Bell, M.G.H., 2004, *Traffic signal timing optimization based on genetic algorithm approach, including drivers' routing*, Transportation Research Part B 38, 329-342.

[8] Chang, T., Sun, G., 2004, *Modeling and Optimization of an Oversaturated Signalized Network*, Transportation Research Part B 38, 687-707.

[9] Cohen, S.L., 1983, *Concurrent Use of MAXBAND and TRANSYT Signal Timing Programs for Arterial Signal Optimization*, Transportation Research Record 906, 81-84.

[10] Cohen, S.L., Liu, C.C., 1986, *The Bandwidth-Constrained TRANSYT Signal-Optimization Program*, Transportation Research Record 1057, 1-7.

[11] Colombaroni, C., Fusco, G., Gemma, A., 2009, *Optimization of Traffic Signals on Urban Arteries through a Platoon-Based Simulation Model*, Proceedings of the 11th WSEAS International Conference on Automatic control, Modelling and Simulation, Istanbul, Turkey.

[12] Daganzo, C.F., 1994, *The cell transmission model: a dynamic representation of highway traffic consistent with hydrodynamic theory*, Transportation Research B 28, 269-287.

[13] Daganzo, C.F., 1995, *The cell transmission model, part II: network traffic*, Transportation Research Part B 29, 79-93.

[14] Daganzo, C.F., 1997, *Fundamentals of transportation and traffic operations*, Pergamon, Oxford, UK.

[15] Dazhi, S., Benekohal, R.F., Waller, S.T., 2006, *Bilevel Programming Formulation and Heuristic Solution Approach for Dynamic Traffic Signal Optimization*, Computer-Aided Civil and Infrastructure Engineering 21, 321-333.

[16] De Schutter, B., 2002, *Optimizing acyclic traffic signal switching sequences through an Extended Linear Complementarity Problem Formulation*, European Journal of Operational Research 139, 400-415.

[17] del Castillo, J.M., 2011, *Three new models for the flow–density relationship: derivation and testing for freeway and urban data*, submitted to Transportmetrica.

[18] Dion, F., Hellinga, B., 2002, *A rule-based real-time traffic responsive signal control system*

with transit priority: application to an isolated intersection, Transportation Research Part B 36, 325-343.

[19]Dion, F., Rakha, H., Kang, Y., 2004, *Comparison of delay estimates at under-saturated and oversaturated pre-timed signalized intersections*, Transportation Research Part B 38, 99-122.

[20]Donati, F., Mauro, V., Roncolini, G., Vallauri, M., 1984: *A Hierarchical Decentralised Traffic Light Control System. The First Realisation: 'Progetto Torino'*, in Proceedings of IFAC 9th World Congress.

[21]Ekstrom, M., Sjodin, A., Andreasson, K., 2004, *Evaluation of the COPERT III emission model with on-road optical remote sensing measurements*, Athmospheric Environment 38, 6631-6641.

[22]Feldman, O., Maher, M., 2002, *Optimization of traffic signals using a cell transmission model*, School of the Built Environment&Transport Research Institute.

[23]Friedrich, B., 2000, *Models for Adaptive Urban Traffic Network Control*, in Proceedings of the 8th International Conference of the European Working Group on Transportation, Rome.

[24]Fusco, G., Gentile, G., Meschini, L., Bielli, M., Felici, G., Cipriani, E., Gori, S., Nigro, M., 2007, *Strategies for signal setting and dynamic traffic modelling*, in Proceedings of TRISTAN VI, Phuket Island, Thailand.

[25]Gartner, N.H., 1989, *OPAC: Strategy for demand-responsive decentralized traffic signal control*, in Proceedings of IFAC Control, Computers, Communications in Transportation, Paris.

[26]Gartner, N.H., Assman, S.F., Lasaga, F., Hou, D.L., 1991, *A multi-band approach to arterial traffic signal optimization*, Transportation Research Part B 25, 55-74.

[27]Gartner, N.H., Hou, D.L., 1994, *Performance Evaluation of Multi-Band Progression Method*, 7th IFAC/IFORS Symposium on "Transportation Systems: Theory and Application of Advanced Technology", Tianjin, China.

[28]Gartner, N.H., Little, J.D.C., Gabbay, H., 1975, *Optimization of traffic signal settings by mixed-integer linear programming, Part I: the network coordination problem*, Transportation Science 9, 321-343.

[29]Gartner, N.H., Little, J.D.C., Gabbay, H., 1975, *Optimization of traffic signal settings by mixed-integer linear programming, Part II: the network synchronization problem*, Transportation Science 9, 344-363.

[30]Gentile, G., 2008, *The General Link Transmission Model for dynamic network loading and a comparison with the DUE algorithm*, Proceedings of the Second International Symposium on Dynamic Traffic Assignment, Leuven, Belgium.

[31]Gentile, G., Meschini, L., Papola, N., 2005, *Macroscopic arc performance models with capacity constraints for within-day dynamic traffic assignment*, Transportation Research Part B 39, 319-338.

[32]Gentile, G., Meschini, L., Papola, N., 2007, *Spillback congestion in dynamic traffic assignment: a macroscopic flow model with time-varying bottlenecks*, Transportation Research B 41, 1114-1138.

[33]Gentile, G., Papola, N., 2009, *The simplified theory of kinematic waves based on cumulative flows: application to macroscopic link performance models*, in Transportation

Systems Analysis: Models and Applications, Springler, 497-510.

[34]Gentile, G., Tiddi, D., 2009, *Synchronization of traffic signals through a heuristic-modified genetic algorithm with GLTM*, in Proceedings of XIII Meeting of the Euro Working Group on Transportation, Padua, Italy.

[35]Greenberg, H., 1959, *An Analysis of Traffic Flow*, Operations Research 7, 78-85.

[36]Greenshields, B., 1935, *A study of traffic capacity*, in Proceedings of the Annual Meeting of the Highway Research Board 14, 448-477.

[37]Hadi, M.A., Wallace, C.E., 1993, *Hybrid genetic algorithm to optimize signal phasing and timing*, Transportation Research Record 1421, 104–112.

[38]Hoogendoorn, S.P., Bovy, P.H.L., 1999, *Multiclass macroscopic traffic flow modelling: a multilane generalisation using gas-kinetic theory*, in Proceedings of the 14th ISTTT, Jerusalem, Israel.

[39]Hunt, P.B., Robertson, D.I., Bretherton, R.D., Winton, R.I., 1981: *SCOOT - a traffic responsive method of coordinating signals*, TRRL Laboratory Report 1014.

[40]Husch, D., Albeck, J., 2006, SYNCHRO 7 User Guide, Trafficware.

[41]Int Panis, L., Beckx, C., Broekx, S., De Vlieger, I., Schrooten, L., Degraeuwe, B., Pelkmans, L., 2011, *PM, NO_x and CO_2 emission reductions from speed management policies in Europe*, Transport Policy 18, 32-37.

[42]Jiang, Y., Li, S., Shamo, D.E., 2006, *A platoon-based traffic signal timing algorithm for major–minor intersection types*, Transportation Research Part B 40, 543-562.

[43]Lebacque, J.P., Lesort, J.B., Giorgi, F., 1998, *Introducing buses into first order macroscopic traffic flow models*, Transportation Research Record, 1644, 70-79.

[44]Lighthill, M.J., Whitham, G.B., 1955, *On kinematic waves II. A theory of traffic flow on long crowded roads*, in Proceedings of Royal Society A 229, 281-345.

[45]Little, J.D.C., 1966, *The synchronization of traffic signals by mixed-integer-linear-programming*, Operations Research 14, 568-594.

[46]Liu, Y., Yu, J., Chang, G.L., 2009, *A Dynamic Model for Signal Optimization with Enhanced Traffic Flow Formulations*, in Proceedings Transportation Research Board Annual Meeting.

[47]Logghe, S., Immers, L.H., 2003, *Heterogeneous traffic flow modelling with the LWR model using passenger–car equivalents*, in Proceedings of the 10th World congress on ITS, Madrid, Spain.

[48]Logghe, S., Immers, L.H., 2008, *Multi-class kinematic wave theory of traffic flow*, Transportation Research Part B 42, 523-541.

[49]Maher, M., 2011, *A comparison of the use of the cell transmission and platoon dispersion models in TRANSYT 13*, Transportation Planning and Technology 34-1, 71-85.

[50]Malakapalli, M.P., Messer, C.J., 1993, *Enhancements to the PASSER II-90 Delay Estimation Procedures*, Transportation Research Record 1421, 94-103.

[51]Messer, C.J., Hogg, G.L., Chaudhary, N.A., Chang, E.C.P., 1987, *Optimization of Left Turn Phase Sequence in Signalized Networks using MAXBAND-86*, Technical Report FHWA/RD 1, 87-109.

[52]Newell, G.F., 1993, *A simplified theory of kinematic waves in highway traffic, part I: general theory*, Transportation Research Part B 27, 281-287.

[53]Newell, G.F., 1993, *A simplified theory of kinematic waves in highway traffic, part II: queuing at freeway bottlenecks*, Transportation Research Part B 27, pp. 288-304.

[54]Newell, G.F., 1993, *A simplified theory of kinematic waves in highway traffic, part III: multi-destination flows*, Transportation Research Part B 27, 305-313.

[55]Ntziachristos, L., Gkatzoflias, D., Kouridis, C., Samaras, Z., 2009, *COPERT: A European Road Transport Emission Inventory Model*, Environmental Science and Engineering Part 2, 491-504.

[56]Papola, N, Fusco, G., 1998, *Maximal bandwidth problems: a new algorithm based on the properties of periodicity of the system*, Transportation Research Part B 32, 277-288.

[57]Papola, N., Fusco, G., 2000, *A new analytical model for traffic signal synchronization*, in Proceedings of the 2nd ICTTS Conference, Beijing, China.

[58]Park, B., Messer, C.J., Urbanik II, T., 1999, *Traffic Signal Optimization for Oversaturated Conditions: Genetic Algorithm Approach*, Transportation Research Record 1683, 133-142.

[59]Park, B., Rouphail, N.M., Sacks, J., 2007, *Assessment of Stochastic Signal Optimization Method Using Microsimulation*, Transportation Research Record 1748, 40-45.

[60]Pelkmans, L., Debal, P., Hood, T., Hauser, G., Delgado, M.R., 2004, *Development of a simulation tool to calculate fuel consumption and emissions of vehicles operating in dynamic conditions*, SAE Fuels and Lubricants 1, 1873.

[61]PTV AG, 2010, VISUM 11.5 User Manual, PTV AG.

[62]PTV AG, 2010, VISSIM 5.30 User Manual, PTV AG.

[63]Richards, P.I., 1956, *Shockwaves on the highway*, Operations Research 4, 42-51.

[64]Robertson, D.I., 1969, *TRANSYT method for area traffic control*, Traffic Engineering & Control 10, 276-281.

[65]Transportation Operations Group, 2006, *PASSER V Guide*, Texas Transportation Institute.

[66]TRL Software, 2010, *TRANSYT 14 User Guide*, TRL Software.

[67]Underwood, R.T., 1961, *Speed, Volume, and Density Relationships: Quality and Theory of Traffic Flow*, Yale Bureau of Highway Traffic, 141-188.

[68]Webster, F.V., 1958, *Traffic Signal Settings*, Road Research Technical Paper 39.

[69]Wong, S.C., Yang, H., 1997, *Reserve capacity of a signal-controlled road network*, Transportation Research Part B 31, 397-402.

[70]Ying, Q.Y., Lu, H., Shi, J., 2007, *An algorithm for local continuous optimization of traffic signals*, European Journal of Operational Research 181, 1189-1197.

[71]Yperman, I., Logghe, S., Immers, B., 2005, *The Link Transmission Model: an efficient implementation of the kinematic wave theory in traffic networks*, in Proceedings of 10th EWGT Meeting and 16th Mini-EURO Conference, Poznan, Poland.

[72]Yperman, I., 2007, *The Link Transmission Model for Dynamic Network Loading*, Ph.D. dissertation, Katholieke Universiteit Leuven, Leuven, Belgium.

[73]Zhang, Y., Chen, X., Zhang, X., Song, G., Hao, Y., Yu, L., 2009, *Assessing effects of Traffic Signal Control Strategies on Vehicle Emissions*, Journal of Transportation Systems Engineering and Information Technology 9, 150-155.

[74]Zhang, Y., Lv, J., Ying, Q., 2010, *Traffic assignment considering air quality*, Transportation Research Part D 15, 497-502.

9 Appendix

Function names are self-explanatory and related procedures will not be described.

9.1 Function BuildNetwork

```
function BuildNetwork
    for each j∈U
        j_N = AddNode()                                  'the centre node of the intersection
for each c∈CA(j)
    a_c = AddArc()                                       'the arc related to the conflict area
    capacity(a_c ) = 0                                   'initialization
next c
        for each h∈H_j
            if |L⁻_h| > 0 then h⁻ = AddArc()
            if |L⁺_h| > 0 then E_h,1 = AddNode(), h⁺_1 = AddArc()

            Λ⁺_h = SortByLaneLengthIncreasing(L⁺_h)      'e.g. use a heap
            cumulativeWidth = 0
            for each l∈Λ⁺_h
                cumulativeWidth = cumulativeWidth + max{minimumWidth,W_l}
            next l
            p = 1                                        'progressive number of intermediate nodes along
                                                         'the leg
            previousLength = minimumLength
            for each l∈Λ⁺_h
                'even if lanes are ordered by length we use “less than” to include lanes shorter
                'than the minimum length
                if L_l ≤ previousLength + minimumLength then    'add minimum length to avoid
                                                                'too short intermediate arcs
                    a_l = AddArc()
                    capacity(a_l ) = capacity(h⁺_1 ) * max{minimumWidth,W_l} / cumulativeWidth
                else
                    p = p + 1
                    E_h,p = AddNode()
                    a_l = AddArc()
                    capacity(a_l ) = capacity(h⁺_1 ) * max{minimumWidth,W_l} / cumulativeWidth
                    h⁺_p = AddArc()                      'further arc of the link
                    capacity(h⁺_p ) = capacity(a_l )     'whose capacity is equal to the
                                                         'sum of its downstream lanes
                    previousLength = L_l
                end if
                for each t∈T_l
                    for each c∈CA_t
                        AddArc(a_t,c )                   'arc from previous point to a_c
```

```
                    if capacity(a_c ) < capacity(a_l ) then capacity(a_c ) = capacity(a_l )
                next c
                AddArc(b_t )                                    'arc from previous point to the
                                                                'destination link
            next t
        next l
    next h
next j
end function
```

9.2 Function PropagationOnConflictAreas

```
function PropagationOnConflictAreas
    for each j∈U
        for each c∈CA(j)
            'update the inflow
            for x = 1 to 2                                      'the BS of a conflict area only has 2 links
                a = BS(tail (c ) ) (x)                          'the x-th arc of the BS of the tail of c
                flow_c,x = flow_c,x + flow(a)                   'cumulates the flow from each link
            next x
            'calculate the sending flow
            for x = 1 to 2                                      'the FS of a conflict area only has 2 links
                b = FS(head (c ) ) (x)                          'the x-th arc of the FS of the head of c
                s_cb = prio_c,x * flow_c,x / (prio_c,1 * flow_c,1 + prio_c,2 * flow_c,2)   'splits the flow between
                                                                'links proportionally to
                                                                'priority
            next x

            […ordinary GLTM node model…]                        'calculate outflow (flow(b))

            'update flows on the conflict area
            for x = 1 to 2
                b = FS(head (c ) ) (x)
                flow_c,x = flow_c,x - flow(b)                   'subtract exit flow for x-th manoeuvre
            next x
        next c
    next j
end function
```