Abstract: In this paper a brief state of art on signal setting design is reported. Preliminarly single junction optimisation methods are described both in under-saturation and in over-saturation condition. Then the main approaches adopted for signal setting design in urban networks are illustrated. Several open issues have been addressed as it concerns the decision variables, the optimisation algorithms, the optimisation criteria (mono and multi), the traffic flow modeling. Finally after some theoretical remarks, research perspectives are outlined.


1 Introduction
The continuous challenge for a sustainable and eco-rational transportation system go through demand based strategies or supply based strategies. The former consists in the well-known travel demand management policies ([1]), including policies based on Intelligent Transportation Systems ([2],[3]), and enhanced technologies such as electric and hybrid vehicles ([4]). Supply based strategies aim to increase the supplied transport capacity, introducing new transport modes ([5]), building new road infrastructures by applying strategic approach ([6]), strengthening the transit system ([7],[8]) and/or optimizing the existing transport capacity.

Among all the cited policies, the optimisation of actual transport capacity is the most economical solution, may solve most of the traffic congestion problems and is complementary to any other transport policy.

At urban level the main framework of the supply design can be defined as Road Network Design Problem (RNDP) and this is addressed to the identification of the optimal configuration of the network in terms of link directions (e.g. road network topology) and to the road traffic control (e.g. Signal Setting Design).

The Signal Setting Design (SSD) is usually addressed through optimisation models where the decision variables are the signal timings, while the network layout is usually assumed given.

Existing contributions on the SSD, address the following problems: single junction optimisation ([9],[10]) or network optimisation ([11]). Furthermore SSD can be generally addressed considering the junctions as isolated or interacting within a network. In the former case, the green timings, their scheduling and the cycle length are calculated without considering the influence of upstream on downstream junctions, in the latter case, the interaction between successive junctions must be considered.

Optimisation models for SSD can be solved through feasible direction algorithms, still these algorithms may fail to find the optimal solution when the objective function is not convex, and/or has several local optimal points.

Moreover, multi-criteria optimisation, which is receiving an increasing attention, can hardly be solved through traditional feasible direction algorithms.

On the basis of these points of weakness, in the last years several metaheuristics derived from biological metaphors, such as evolutionary algorithms, have been developed first to deal with discrete optimisation, and then extended to cope with continuous optimisation possibly with several optimal points. These methods include genetic algorithms (GAs), particle swarm optimisation (PSO) ant colony (ACO) bee colony (BCO) bacteria foraging algorithms (BFO).

In accordance with the literature, among them, particular attention was given to the GAs. They have successfully been applied to mono-criteria network SSD; however, few papers proposed their application to multi-criteria single junction SSD and to multi-criteria network SSD.

Over the years a number of traffic flow models have been proposed and described in the literature. Traffic flow theories seek to describe in a precise mathematical way the interactions among vehicles, drivers and the infrastructure.

One objective of traffic flow theory is to derive theoretical relationships between the various traffic variables so that the engineer can determine the
characteristics of traffic streams and hence predict the consequences of alternative designs.

2 Single junction signal setting design

The traffic signal setting for an isolated junction can be addressed through delay minimisation or capacity factor maximisation. In particular, in under-saturation conditions, two main problems with a different set of variables are defined: green timing problem and green timing and scheduling problem.

In the former problem green timings and Cycle length are calculated while the stage matrix is fixed. The optimisation can be obtained through specific “local policy”:

- Equisaturation: Green timing is dealt with through a simple close form method still with some limiting assumptions ([12]);
- Junction total delay minimisation: The problem is formulated in a convex mathematical program and solved by an ad hoc procedure. The whole algorithm can be implemented with a Fortran program called SIGSET ([13]);
- Junction capacity factor maximisation: Once determined the traffic capacity for isolated signal controlled junctions, the problem can be formulated as a linear programming (LP) model that can be effectively solved by adopting a computer program, known as SIGCAP ([14]).

A commercial program, usually adopted for designing signal timings, suitable both for delay minimisation and capacity factor maximisation, is OSCADY ([15]).

With respect to the green timing and scheduling problem, in under-saturation conditions, the signal setting optimisation for isolated junctions can be solved through the capacity factor maximisation or through the delay minimisation adopting discrete and non-linear optimisation techniques ([16])

The mathematical programming techniques that can be adopted to evaluate the signal settings optimisation problem for isolated junctions can be grouped into two classes:

1. Stage based approach: The composition and the sequence of the stages are assumed to be fixed initially and need to be specified in advance while the green times for the stages optimising a given performance index are calculated. With respect to the graph theory, a signal controlled junction can be represented by a compatibility graph which imposes that no conflicting movements are permitted to move together. Considering a clique as a sub-graph that represents a stage, the analysis of the sequences of compatibility cliques is conducted. The calculation of the optimal clique sequence allows to enhance the overall efficiency of a junction within a signal cycle ([17]; [18]). The computation of the green time for each clique is evaluated by a performance index.

2. Phase based (group based) approach: green timing and scheduling can be obtained with respect to the incompatibilities among the streams. The stage structure is variable during optimisation and so these methods can be called phase based.

[19] and [20] formulate a group based approach for the problem of signal setting optimisation as a Binary Mixed Integer Linear Program (BMILP) which is solved by discrete programming techniques (i.e. a branch and bound technique). Control variables are calculated simultaneously, the number and the composition of stages are an implicit result of the computation procedure.

[21] and [22], starting from Zuzarte’s procedure ([18]) propose a related approach to minimise delay for isolated signal controlled junctions in which the cycle structure is specified by one of the stage sequences (clique sequences). Hence this stage sequence can vary during optimisation process providing much flexibility for the calculation of signal variables. The limit of this method is in adopting Zuzarte’s procedure that produces a large number of stage sequences and requires an iterative gradient search method (in which the descent direction at each iteration can be determined) to solve the problem. To overcome this limit [23] introduces a procedure to group all alternatives possibilities in a smaller number of equivalent classes with a successor function that represents each of this latter. The green timings and the scheduling are represented by two variables for each group (e.g. start and duration of each group) and the computations can be faster. [24] develops a more detailed mathematical model for the group based method and implements the framework in a software called SIGSIGN ([25]) in which double green signals with varying saturation flows designs are available and the signal timings are designed by minimising critical cycle length, maximising capacity factor or minimising total delay (computed through Webster two terms expression).

According to these methods [15] develop Oscady PRO (Phase based Rapid Optimisation),
a computer program for optimising phase-based signal timings and calculating capacities, queue lengths and delays (both queuing and geometric) for isolated traffic signal controlled junctions.

The over-saturation conditions for an isolated junction can be addressed through total delay minimisation during the entire oversaturated period and not per cycle only, looking for best phase switching strategies.

The problem of optimising the control of oversaturated junctions can be solved by using the semi-graphical methods ([26]) in which the optimum control involves values of the control variables that lie along edges of the control region, which in this case is defined by the permissible ranges of the green phase splits. The Pontryagin maximum principle is applied to derive analytical solutions of the optimal trajectories.

[27] and [28] propose a method called “bang-bang control” that attempts to find an optimal switch over point during the oversaturated period to interchange the timing of the approaches.

[29] develop a discrete dynamic model and performance index approach to optimise signal parameters during the entire period of oversaturated conditions. [30] describes the optimal control as a function of current queue lengths so that it can be applied in real time.

The optimal lane use into the stage based design framework has been introduced in [31]. In this paper the authors propose a mixed-integer linear programming model for an integrated design of lane uses and signal phases which allows to increase the capacity of road junctions. Traffic and pedestrian movements are considered in a unified framework. In [32] and [33], three optimisation problems are considered: the capacity maximisation; the cycle length minimisation and the delay minimisation. The first two cases are addressed by formulating Binary-Mix-Integer-Linear- Programs (BMILPs) that are solved by standard branch and bound routines. The delay minimisation, instead, is formulated as a Binary-Mixed-Non-Linear-Program (BMINLP) that is solved by applying heuristic cutting plane algorithm.

[34] extend the lane based design method for the reserve capacity optimisation by including new integer variables such as the number of approaches and exit lanes in various arms and introducing a unified optimisation framework with the lane markings, the lane flows and signal timings. The reserve capacity maximisation is addressed by maximising the common flow multiplier that is formulated as a Binary-Mixed-Integer-Linear-Programming (BMILP) which is solved by a branch and bound routine.

3 Traffic flow model

Network signal setting design could be addressed by considering the junctions composing the network as isolated or interacting. These different assumptions lead to different set of optimisation variables and thus to alternative optimisation frameworks. When the junctions in the network are considered as isolated the signal settings is obtained by considering uniform arrival profiles. On the contrary, when the interaction among the junctions is considered, the traffic flow modeling determines how the arrivals at the junctions are distributed over the time and strongly affects the optimisation outputs.

Traffic flow models can be classified according to the:
- Scale of the independent variables (continuous, discrete);
- Level of detail (microscopic, mesoscopic, macroscopic);

3.1 Scale of independent variables

In this group two time scales, namely continuous and discrete are considered. A continuous model describes how the traffic system’s state changes continuously over time in response to continuous stimuli. Discrete models assume that state changes occur discontinuously over time at discrete time instants.

3.2 Level of detail

Traffic models may be classified according to the level of detail with which they represent the traffic systems: Microscopic models describe both the space-time behavior of the systems’ entities (i.e. vehicles and drivers) as well as their interactions at a high level of detail (individually); in Mesoscopic models traffic is represented by (small) groups of traffic entities, the activities and interactions of which are described at a low detail level; Macroscopic flow models describe traffic at a high level of aggregation as a flow without distinguishing its constituent parts. For instance, the traffic stream is represented in an aggregate manner using characteristics as flow-rate, density, and velocity. Macroscopic flow models can be classified according the number of partial differential equations that frequently underlie the model on the one hand, and their order on the other hand.

Despite the high level of detail reached by using microscopic models, an extreme computation’s ability is required. Hence, when simulation is aimed to real time traffic predictions (relevant to the dynamic description of traffic), macroscopic models are preferred due to the possibility of using mean state variables’ values (e.g. mean speed, flow rate, etc.) and considering traffic as hydraulic flow.
In the case of a signalised network, two main issues are to be addressed: (i) the modelling of the dispersion between interacting junctions, which is strictly related to the distance travelled on the connecting links and (ii) the spillback (i.e. the link blockage) and the merging and diverging modelling (i.e. the lane blockage).

All these phenomena may be observed in macroscopic traffic flow models. At macroscopic level (as described above) the vehicles are not looked as separate but as aggregate entities. Generally two main classes of models can be identified:

- Space discrete models
- Space continuous models

The space discrete models describe the propagation of flows through a link by relationships between whole link variables such as link travel time, link inflows, outflows or link volume (i.e. the number of vehicles on the link) at each point in time. These models do not require any space discretization and for this reason are also named as link based. Whole link models ([35]; [36]) are widely used in mathematical programming models for dynamic traffic assignment (DTA) because of their simplicity. However they show several limits: Firstly, as well as the link length increases, they are not able to represent reliable hypocritical congestion (and so the spillback effects) due to the fact that the propagation of flow states along the link is not considered; They further cannot be applied at network level because in this case it may be difficult to define which aspects of results are due to behaviour within individual links and which is due to network effects.

The continuous space models derive from the analogy between vehicular flow and flow of continuous media (e.g. fluids or gasses), yielding flow models with a limited number of partial differential equations that allow to describe the dynamics of variables like the following:

- Density (k): Typical variable from physics adopted by traffic science to express the number of vehicles per kilometer of road.
- Flow rate (q): Represents the number of vehicles that passes a certain cross-section per time unit.
- Mean speed (u): Defined as the quotient of the flow rate and the density.

By explicating the propagation of the flow states in every space-time point of the link, these models require a dense space discretization and for this reason are also named as point based models.

The most elementary continuous traffic flow model is the first order model developed concurrently by [37] and [38], based around the assumption that the number of vehicles is conserved between any two points if there are no entrances (sources) or exits (sinks). This produces a continuous model known as the Lighthill-Whitham-Richards (LWR). This particular model suffer from several limitations. The model does not contain any inertial effects, which implies that vehicles adjust their speeds instantaneously, nor does it contain any diffusive terms, which would model the ability of drivers to look ahead and adjust to changes in traffic conditions, such as shocks, before they arrive at the vehicle itself. In order to address this limitations [39] develop a second order continuous model governing traffic flow.

[40] demonstrates that the Payne model, as well as several other second-order models available in the literature, produces false behaviour for some traffic conditions. Specifically, it is noted that traffic arriving at the end of a densely-packed queue would result in vehicles travelling backwards in space, which is physically unreasonable. This is due to the isotropic nature of the models, as the behaviour of vehicles is influenced by vehicles behind them due to diffusive effects.

As the differential equations used in LWR model are difficult to solve, especially in situations of high density variations like bottlenecking (in this cases the LWR calls for a shock wave), different approximate techniques have been purposed to solve that equations. [41], introduces a simplified theory of kinematic waves in which, by using cumulative inflow/outflow curves, the state of flow at an extreme, according to the traffic conditions of another one, can be predicted without considering traffic conditions at intermediate sections. This theory provides a relation between traffic flow q and density k, captured in the triangular shaped fundamental diagram. The author purpose in this way a space discrete model (link based) which provides link travel times complying with the simplified kinematic wave theory.

Consistently with simplified first order kinematic wave theory after Newell, [42] present the Link Transmission Model (LTM) in which link volumes and link travel times are derived from cumulative vehicle numbers. Another way to solve the LWR space continuous problem is introduced by [43] through the “Cell Transmission model”, developed as a discrete analogue of the LWR’ differential equations in the form of difference equations which
are easy to solve and also take care of high density
density changes.

In assuming a uniform speed for all the vehicles in a
road, Daganzo’s CTM and Yperman’s LTM cannot
fully predict realistic traffic flow behaviour as the
platoons keep the same density when moving from the
upstream stop-line section to the downstream
section, and all vehicles travel at the same free flow
speed. For this reason some authors ([44]; [45])
implement a traffic flow model that takes into
account a platoon dispersion volume function which
describes a more realistic volume from the upstream
section to the downstream one when the density is
low.
The just mentioned CTM ([43]) as well as its
modified version MCTM ([46]) or its piecewise
linearized version (i.e. the Switching Mode Model
SMM proposed by [47]) are shown to be
computationally efficient and analytically simple,
yet able to capture many important traffic
phenomena, such as queue build-up and dissipation,
backward propagation of congestion waves, etc.
However, at same time, deterministic in/out-flow
profiles in such model is taken into account and no
randomness in demand and supply sides of the
traffic network is provided. To overcome this
modelling limitations, [48] propose the Stochastic
Cell Transmission Model (SCTM), in which the
demand is modelled as stochastic exogenous input
whereas the supply uncertainties (i.e. weather
conditions, traffic incidents, traffic control etc.) are
governed by the random parameters of the triangular
fundamental flow-density diagram, e.g. free-flow
speed, jam-density, backward wave speed, etc. Five
probabilistic events are identified. In order to
capture such traffic dynamics under demand and
supply uncertainties proper traffic control strategies
have to be adopted.

3 Network signal setting design
The signal setting optimisation in urban network, is
carried out by the coordination of signals and the
synchronisation of signals. In case of signals
coordination, the procedure is based on the
optimisation of the offsets (based on delay
minimisation), once known the green timings, the
cycle length and the scheduling for each junction
(based on delay minimisation or maximisation of
capacity factor). For traffic signal synchronisation,
the procedure is based on the optimisation of the
offsets and, at the same time, of the green timings
and of the cycle length at each junction (based on
delay minimisation). Although it’s possible, with the
currently available commercial software to
implement the network delay minimisation by
considering as decision variables the green timings
at each junction and the offsets among the
interacting junctions, literature lacks of methods
which include the stages scheduling. This variable
makes the problem considerably harder and it is
often neglected, assuming the stage sequence as
given.
In terms of network signal setting design different
contributions may be find in literature.

Three open issues can be identified:
1. the decision variables: at network level,
   the signal timings, the offset, the stages
   sequence and the cycle length may be
   considered, even if the stages sequence
   optimisation is not properly stated in
   literature;
2. the optimisation criteria: two approaches
   may be adopted i.e. the mono-criteria and
   multi-criteria approach;
3. The traffic flow models.

The straightforward adopted software are:
TRANSYT© (TRL, UK) and TRANSYT-7F©
(FHWA, USA). Both are addressed to the signal
setting design problem by optimising the signal
timings, the offsets and the cycle length by
combining the traffic flow model and the signal
setting optimiser. The traffic flow model is based on
the platoon dispersion model ([49]) whereas the
optimisation procedure is based on the minimisation
of a performance index (P.I), a single term or
weighted combination of two indicator as a proxy of
multi-criteria optimisation (i.e. delay, number of
stops in TRANSYT and delay, throughput in
TRANSYT-7F). by metaheuristic approach.
With regard to the metaheuristics, in TRANSYT the
Hill Climbing and the simulated annealing are
implemented whereas in TRANSYT-7F the Hill
Climbing and the Genetic Algorithms are
implemented. One of the main problem in
metaheuristics application is related to the search
space definition. In fact the main difference between
simulated annealing/hill climbing and genetic
algorithms is that the first one are based on a single
point search space where the “new” point is selected
by the neighbourhood of the current point and the
second one is based on a multidirectional search
space. However depending on the search space
definition different computational efforts can be
obtained.
In addition, it can be observed that each
metaheuristic is classified with respect to the
number of parameters to be preliminary selected; in
In terms of traffic conditions, two principal classes of signal control strategies can be distinguished: In the first class, strategies are only applicable to networks with under-saturated traffic conditions, whereby all queues at the signalised junctions are served during the next green phase. In the second class, we have strategies applicable to networks with oversaturated traffic conditions, whereby queues may grow in some links with an imminent risk of spillback and eventually even of gridlock in network cycles.

When the traffic conditions in the network are below saturation, the TRANSYT ([49]), which assumes that the average flow demand at an approach remains constant, and SCOOT ([59]; [60]), an evolution of TRANSYT which assumes flows detected in real time, are two well-known and widely used coordinated traffic strategies. These latter lead to effective results, but their performance may deteriorate when severe congestion persists during the peak period.

When congested traffic conditions occur the most widely used strategy is the “Store-and-forward”. The objective of this method is to allocate the queues with respect to the longitudinal road capacity as avoiding the diffusion of the congestion conditions within the network. The used approach address to a signal setting which may vary among successive cycles. The Store and Forward was first suggested by [61] and has been used in various works notably for road traffic control. This modeling approach can be used for the coordinated control of large-scale congested urban networks but, on the other hand, the modeling simplification allows only for split optimisation, while cycle length and offsets must be delivered by other control algorithms. A recently developed signal control strategy of this type is TUC ([62]). More recently, a number of approaches have been proposed employing various computationally expensive numerical solution algorithms, including genetic algorithms ([63]; [64]). In [65] and [66] the traffic flow conditions are modeled using the cell transmission model ([43]), even if in view of the high computational requirements, the real-life implementation of these optimisation-based approaches might face some difficulties in terms of real-time feasibility.

4 Conclusions

In this paper a literature review on Network signal setting design has been addressed. First the most straightforward strategies which may be followed for the optimisation of decision variables have been
described. In particular, we can identify a (i) three step optimisation, in which first the stage matrix (SM) i.e. stage composition and sequence, the green timings (GT) at each single junction are optimised, then the node offsets (Off) are computed in three successive steps; (ii) two step optimisation, in which: the stage matrix is defined at a first step, then the green timings and the node offsets are computed at a second step; the scheduling and green timings were computed at a first step, and the offsets were computed at a second step. In both approaches the stage matrix optimisation is carried out through explicit complete enumeration (Ex.en). No one-step optimisation method for the simultaneous optimisation of green times, their schedule, and node offsets, the so-called scheduled synchronization, is already available to authors’ knowledge. Formally, existing approach-based methods for single junctions may easily be extended to specify one-step methods for NSSD, since the node offsets may easily be obtained from decision variables, say the start and the end of the green of each approach, and if needed the stage composition and sequence as well. Nonetheless the resulting problem may be hard to solve since several equivalent local optima exist; this condition may quite easily dealt with for a single junction, but it is rather unclear how it can effectively be circumvented for a network (with loops).

In terms of traffic flow models the literature offers several contributions which may be considered as an integration or an extension of the two most widely adopted models: the Platoon Dispersion Model (PDM) and the Cell Transmission Model (CTM). Only few authors (Cantarella et al.2015) have investigated on the assumption of mixing the two models in order to explicitly represent both horizontal queuing phenomena (avoiding the vertical queue representation obtained in PDM) as well as dispersion along a link (avoiding that the platoons keep the same density when moving from the upstream stop-line section to the downstream section as it occurs in the discrete version of the LWR model such as the CTM).

Summing up, in the following Table 1, it is shown a comparison among the optimisation strategies with respect to the optimisation criteria (Crt) in terms of mono and multi- criteria approach, the objective functions (Of), by considering the generally adopted one such as the Total Delay (TD) and the Capacity Factor (CF), possibly mixed or switched by generic objective functions (of) and the optimisation (Opt) Tools in terms of Benchmark (Ben), Proposal (*) and Future Perspectives (FP)

Further topics worth of research effort are: NSSD based on reserve capacity maximization at network level, and the specification of solution algorithms following the phase-based methodology. Finally it should be worthy of interest integrating the NSSD with the traffic assignment ([67],[68]) in order to evaluate the strategy effectiveness in terms also of urban sustainable impact ([69],[70]).

Table 1 - Paper contributions, Benchmarks and Future perspective on Network signal setting design.

<table>
<thead>
<tr>
<th>Opt</th>
<th>var</th>
<th>Strategy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>three step</td>
</tr>
<tr>
<td>Crt</td>
<td>SM</td>
<td>Ex.en</td>
</tr>
<tr>
<td></td>
<td>GT</td>
<td>mono</td>
</tr>
<tr>
<td></td>
<td>Off</td>
<td>mono</td>
</tr>
<tr>
<td>Of</td>
<td>SM</td>
<td>ND</td>
</tr>
<tr>
<td></td>
<td>GT</td>
<td>CF</td>
</tr>
<tr>
<td></td>
<td>Off</td>
<td>TD</td>
</tr>
<tr>
<td>Tools</td>
<td>SM</td>
<td>Ex.en</td>
</tr>
<tr>
<td></td>
<td>GT</td>
<td>OSCADY (GT)</td>
</tr>
<tr>
<td></td>
<td>Off</td>
<td>TRANSYT (C)</td>
</tr>
</tbody>
</table>

-Sy: Synchronisation;
-GT&S: Green timing and Scheduling
-C: coordination;
PDM/CTM: platoon dispersion model or cell transmission model; CT&PDM: cell transmission and platoon dispersion model.
References:


[22] S. Gallivan, B. G., Heydecker, Optimising the control performance of traffic signals at a


