Pergamon

Transpn. Res.-B. Vol. 30, No. 2, pp. 99–114, 1996

0191-2615/96 $15.00 + 0.00

0191-2615(95)00025-9

A DECOMPOSITION APPROACH FOR SIGNAL OPTIMISATION IN ROAD NETWORKS

B. G. HEYDECKER
Centre for Transport Studies, University College London, London WC1E 6BT, England

(Received 31 December 1993; in revised form 23 March 1995)

Abstract—The optimisation of signal timings plays an important role in the management of urban traffic, and in the full usage of existing and planned road networks. In recent years, considerable advances have been made in techniques for the optimisation of signal timings at isolated junctions operating under fixed-time control. This paper shows how these techniques can be applied in the optimisation of signal timings in coordinated networks by using a decomposition approach. The signal timings at a junction in a network can be specified fully by the sequence of stages, interstage structures, cycle time, stage durations and offset. Of these variables, the third, fourth and last are endogenous to network optimisation methods, the first and second being exogenous. Techniques have been developed recently to optimise all but the last variable (which is not there defined) at individual junctions, and these have been found to give considerable improvements in operational performance. The computational requirements of these methods is such that their direct extension to networks is not yet a practical proposition. This paper shows how the differences inherent between individual junction and network optimisation methods can be reconciled within a decomposition approach so that the latter can benefit from some of the advantages of the former. A simple example is used to illustrate the substantial benefits that can arise from this approach.

1. INTRODUCTION

The optimisation of signal timings in road networks is important both for their operation and for design purposes. In the latter case, the relative performance of alternative designs can only be evaluated after signal timings have been optimised for each one. Because of this, signal optimisation methods have a dual role as operational techniques and as design tools.

A number of methods have been developed for the optimisation of signal timings in road networks, including TRANSYT (Vincent *et al.*, 1980; Crabtree, 1988), MAXBAND (Little *et al.*, 1981), MITROP (Gartner *et al.*, 1975a,b) and SIGMA (Bielefeldt, 1987). However, considerations of computational complexity have restricted the level of detail used in each of the traffic and control models that are incorporated in these methods. Furthermore, the optimisation techniques adopted do not yield information on sensitivity of the optimised performance to changes in parameter values. Information of this kind is especially useful when considering improvements to elements of a design.

By contrast, recently developed methods for the optimisation of signal timings at individual junctions operating under fixed-time control offer greater detail and flexibility in their traffic and control models, and more powerful and informative optimisation techniques. Examples of these methods are those of Improta and Cantarella (1984), Möller (1987), Heydecker and Dudgeon (1987), Silcock and Sang (1990), Rouphail and Radwan (1990), Alçelik (1991) and Silcock (1992). Amongst the advantages of these methods are more detailed optimisation resulting in improved performance, reduced requirements for preprocessing of data, and information on sensitivity of the optimised solution to changes in design parameters. An extensive review of methods for individual junctions and discussion of their characteristics has been given by Allsop (1992).

As they stand, these individual junction methods do have some limited applicability within networks of coordinated signals: their analysis of cycle time and capacity are equally appropriate to isolated and to coordinated junctions. However, the analysis of

delay depends on arrivals of vehicles being serially uncorrelated which is a reasonable assumption only at isolated junctions: the platooning of traffic that occurs in dense urban road networks violates this assumption. Furthermore, these methods do not model offset variables which are known to play an important role in the coordinated optimisation of signals in dense urban networks. Whilst there is no objection in principle to the application of similarly detailed models of junction control simultaneously in whole networks, suitable analytical models of traffic behaviour are not available and the computational complexity of optimisation would present practical difficulties.

This paper describes a decomposition approach to the design of road networks and signal control plans for them which uses methods for individual junctions and networks in a complementary combination. The individual junction methods are applied first, and those elements of the signal timings which cannot be varied within the network optimisation process are extracted from the resulting solutions. A coordinated network method is then applied to optimise all of the remaining elements of the signal timings. Some elements are common to both optimisation methods and will be varied by each of them. In order to apply these methods consistently, the traffic models used in the two optimisation processes must be made mutually consistent.

The present approach is sub-optimal in that it does not take into account interactions between adjacent junctions when optimising those elements which are to be fixed during the network optimisation. On the other hand, it makes available in an appropriate form advanced individual junction techniques for use in the context of network optimisation so that they can be used to assist in the design process. Results from the application of this method to a simple example network show that worthwhile benefits can be achieved by using it.

2. SIGNAL OPTIMISATION TECHNIQUES

2.1. Introduction

The design of signal-controlled networks involves the two distinct processes of the specification of physical details and the calculation of signal timings: these are necessarily undertaken sequentially. Before the operational performance that can be achieved using a physical design can be estimated, appropriate signal timings must be calculated. Any sensitivity analysis that is performed during the calculation of the signal timings can then be used to identify ways in which the physical design can be improved. Thus signal optimisation programs can also perform in the wider role of tools to assist in physical design.

A set of signal timings at a junction can be described in two distinct ways, according to which is more convenient. In the first description, intervals can be specified throughout which a set of signals each display green indications concurrently while all others display red ones: these are called *stages*. The intervals between stages, during which signal displays change between red and green, are called *interstages*. Examples of these quantities are depicted in Fig. 1, (Heydecker & Dudgeon, 1987) which shows that at junctions with non-standard layout, the optimal sequence and interstage structures can be intricate to an extent that manual procedures would be unlikely to identify them.

The signal timings at a junction can be described in terms of the sequence of stages, their durations, the times during each of the interstages at which individual signals change, the time taken for a complete set of signal indications to repeat (known as the *cycle time*), and for a junction in a network, the time relative to a master clock at which the first stage of the sequence starts, which is known as the *offset*.

As an alternative, signal timings at a junction can be described in terms of the times relative to some master clock at which each of the signals changes to green, the durations of the subsequent intervals throughout which they remain green, and the cycle time. In practice, sets of signals are switched simultaneously or with fixed lags between them: these sets are called *groups*. An economy can thus be achieved by describing the start times and durations of groups rather than those of individual signals. Several mutually compatible groups may receive green concurrently, and their order of occurrence can be changed

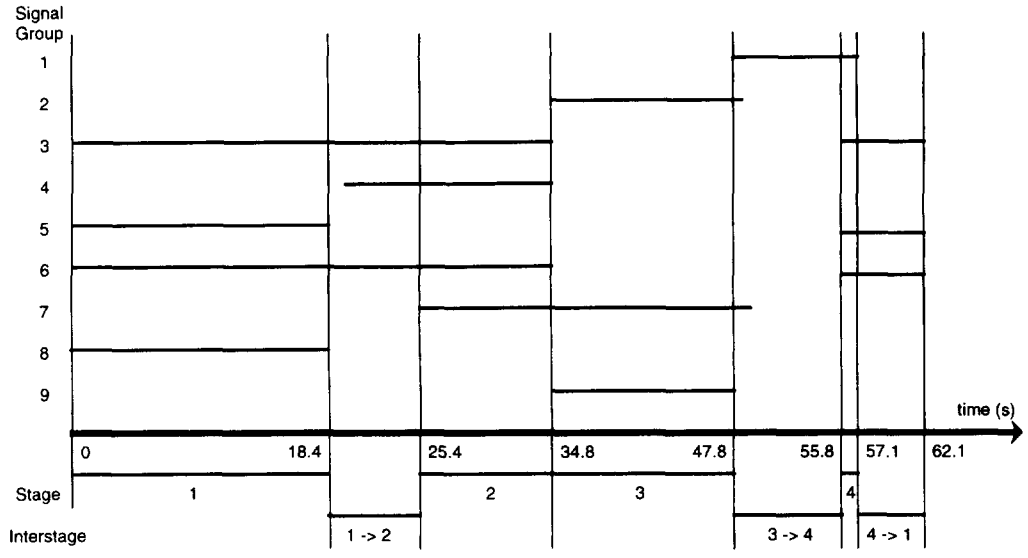

Fig. 1. Stages and interstages in signal control (Source: Heydecker & Dudgeon, 1987).

automatically during optimisation. This group-based description normally requires more variables and constraints than does the stage-based one, but it offers greater flexibility in optimisation and is more closely related to the operation of the junction thus reducing the amount of pre-processing of data that is required.

The stage-based description is often used as the basis of optimisation methods for signal timings. In these methods, the variables for each junction corresponds to the stage durations, cycle time and offset. Elements that are specified exogenously include the sequence of stages and the structure of the interstages. A number of constraints apply which ensure that each green interval exceeds some minimum acceptable value, that adequate capacity is provided, and that the cycle time lies in a suitable range. In this method, stages cannot normally be eliminated from or introduced into the sequence by any automatic process because of difficulties that this would cause with the associated interstages.

The group-based description can also be used as the basis of signal-timing optimisation methods. The elements that are specified exogenously for these methods are related directly to safety and other traffic engineering considerations. Thus the minimum acceptable duration of green for each group, the capacity requirement for each stream of traffic and the minimum time required between green indications for each pair of mutually incompatible groups (known as *group intergreen times*), are related directly to elements of the design and operation of the junction that can be determined readily beforehand. The signal timings resulting from a group-based optimisation procedure can be described in terms of a sequence of stages, stage durations, interstage structures, cycle time, and offset, even though some of these were not used explicitly as variables within the procedure that was used to calculate their values. Because of this, the stage durations, interstage structures, and to some extent the sequence of stages, can be optimised by group-based methods in a way which is not possible using stage-based ones: this is especially important at complicated junctions, where it can give rise to substantial improvements in performance over that attained by stage-based methods (Heydecker & Dudgeon, 1987).

In the remainder of this section, the methods adopted for the optimisation of signal timings in networks and at individual junction are discussed in turn. Particular reference is made to differences between the modelling and optimisation techniques used at the two levels.

2.2. Network methods

Methods which are used to optimise signal timings simultaneously throughout networks are subject to the requirement of simplicity for reasons of data coding, computer

storage, and the computational demands of optimisation. Because of this, existing practical methods are all based upon stage-based formulations in view of the fewer variables and constraints required. A number of problem-specific solution methods have been developed for the resulting optimisation.

A direct consequence of the adoption of this stage-based approach is that the stage sequence and intergreen structures for each signal-controlled junction have to be specified exogenously. This can constitute a considerable amount of information and requires a correspondingly great effort in data preparation. Furthermore, the choice of stage order is known (Vincent *et al.*, 1980, p. 11; Crabtree, 1988, p. 9) to influence substantially the optimised performance in some networks. Some formulations such as MAXBAND (Little *et al.*, 1981) and that of Rouphail and Radwan (1990) can inform the choice between certain prescribed stage configurations where these relate to the treatment of opposed turning traffic. However, the choice of sequence will generally be greater than that arising from this in isolation: this is often the case when explicit consideration is given to pedestrians, separate public transport movements or cyclists, or when the junction layout is unusual.

The TRANSYT program does provide a facility called STAGOR to investigate the consequences of reordering the stages at each junction which takes into account the effects of coordination between adjacent junctions (Crabtree). However, the procedure used can neither eliminate nor introduce stages and yet gives rise to a considerable number of sequences to be considered, being the product of the number of possibilities at each junction in the network. Furthermore, the interstage structures cannot be optimised but rather the start and end times for each signal group are fixed relative to the start and end times of the stages regardless of which other groups start or end green during the same interstage.

In network optimisation methods, the cycle times used at all junctions which are optimised together are either identical or related by integer multiplicative factors. Because of this, the flow of traffic on each link of the network follows a cyclic pattern, thus allowing a single cycle to be analysed and taken as representative of all operation. The TRANSYT program (Vincent *et al.*, 1980, Crabtree, 1988) provides a facility called CYOP to evaluate approximately the optimised performance of the network at each of a range of cycle times. This facility can be used to identify an appropriate cycle time at which to undertake a full optimisation of the remaining signal timing variables.

The elements of traffic that are modelled individually correspond to links of the network. All traffic on a link is controlled at the same stop-line and receives identical signal indications. Despite advice that separate links be coded for each set of turning movements that form a single queue, the usual criterion for coding is sets of adjacent movements that receive identical signal indications.

Various optimisation strategies have been adopted for signal-controlled road networks. The MITROP (Gartner *et al.*, 1975a,b) and MAXBAND (Little *et al.*, 1981) methods use relatively simple models of traffic behaviour which lead to analytical expressions for the delay on each link: this makes possible use of a simultaneous optimisation technique for the endogenous variables. On the other hand, TRANSYT (Vincent *et al.*, 1980) uses a mesoscopic simulation of traffic, including the dispersion of platoons of traffic as they travel along links and the passage of platoons from link to link as they pass through junctions. The price of this relatively detailed model of traffic is that no simple analytical optimisation methods are appropriate: instead this is undertaken as a direct search of a discrete space. This search is undertaken considering in turn changes to individual offsets and transfers of time between consecutive stages. Each of these methods uses an additional term in their estimates of delay to account for the effects of congestion caused by stochastic deviations from their idealised analytic or simulation models of traffic behaviour: this component is known as *random delay*.

2.3. Individual junction methods

The level of detail available for models of traffic behaviour and signal control at individual junctions is greater than for those in networks. In particular, sets of lanes which

are used by traffic that forms a single queue are normally modelled separately as streams. This corresponds to a refinement of the link-based models that is usually used for networks. Furthermore, use of a group-based description of signal timings at individual junctions gives rise to flexible and powerful optimisation formulations which have acceptable size and complexity.

Estimates of reserve capacity and range of acceptable cycle times can be made at junctions individually whether or not they are part of a network of coordinated signals. Under the assumption of serially uncorrelated arrivals of vehicles, the mean delay incurred by them can be estimated using simple formulae such as that of Webster (1958). This makes possible the use of efficient analytical optimisation techniques for individual junctions which are not part of coordinated networks.

Webster's simplified (two-term) formula for the mean delay incurred by vehicles at a signal-controlled road junction can be expressed as

$$d = \frac{9}{20}\left[\frac{sr^2}{(s-q)c} + \frac{x^2}{q(1-x)}\right]$$

where

s is the saturation departure rate of vehicles
q is the mean arrival rate of vehicles
r is the effective red time
c is the cycle time, and
$x = \frac{qc}{s(c-r)}$ is the degree of saturation.

The first term

$$d_u = \frac{9}{20}\frac{sr^2}{(s-q)c}$$

represents the mean delay that would be incurred by vehicles if they arrived uniformly throughout the signal cycle. As with network models, a random component of delay, here given by

$$d_r = \frac{9}{20}\frac{x^2}{q(1-x)}$$

is added to account for the effects of congestion caused by stochastic deviations from the idealised model of uniform vehicular arrivals. Whilst Webster's two-term formula is not the only one available to estimate mean delays incurred at isolated signal-controlled road junctions, it has the desirable properties of convexity in the signal-control variables $(c\text{-}r)/c$ and $1/c$, and of computational convenience.

Improta and Cantarella (1984) developed a mixed-integer programming method to maximise the reserve capacity at an individual junction with respect to group timings, cycle time and the order of occurrence of mutually incompatible groups. Möller (1987) developed graph-theoretic techniques to optimise reserve capacity at individual or closely adjacent junctions between which the requirement for good coordination can be expressed directly in terms of constraints on the signal-group timings. Heydecker and Dudgeon (1987) developed convex programming techniques to minimise the total rate of delay at isolated junctions with respect to group timings and cycle time. This method provides the basis for the SIGSIGN computer program (Silcock & Sang, 1990) which includes detailed representations within the traffic model of semi-compatible movements such as opposed turners, thus allowing for fully automatic optimisation of their treatment (Silcock, 1992). Rouphail and Radwan (1990), and Akçelik (1991) describe other methods for detailed modelling and analysis of this kind.

The use of gradient methods for the solution of optimisation problems arising at individual junctions means that estimates of the sensitivity of the optimised performance

Table 1. Nature of variables in individual junction and coordinated optimisation

Variable	Individual junction	Coordinated control
Sequence	Total number $\sum_{j=1}^{J} n_j$	Total number $\prod_{j=1}^{J} n_j$
	Optimised implicitly by repeated trial	Exogenous
Interstage structure	Optimised implicitly	Exogenous
Stage durations	Optimised implicitly	Optimised endogenously
Cycle time	Optimised endogenously	Common value Optimised by repeated trial
Offset	Not defined	Optimised endogenously

to changes in parameter values can be provided. This post-optimality sensitivity analysis can be valuable in identifying beneficial modifications to the physical design of junctions.

In Möller's, and Heydecker and Dudgeon's methods, the order of occurrence of the groups in each mutually incompatible pair is specified exogenously. The method of Tully (1976) can be used to generate all possibilities which can be numerous: these can usually be reduced to a few distinct cases for optimisation (Heydecker, 1992), each of which can correspond to several sequences of stages. In many cases the choice between these orderings has a profound effect on optimised performance at the junction. In cases where the group intergreen times vary between pairs of groups, the order of occurrence of mutually incompatible groups can be especially important.

Although these group-based optimisation methods do not use stage durations directly as variables, the solution signal timings can be described using stage-based variables. Within each of the cases considered, a partial optimisation of the stage sequence occurs automatically during optimisation of the group timings, including the possible elimination or introduction of stages. Each distinct ordering of incompatible groups that is available for a junction can be optimised in this way in turn. Thus these methods can be used, with some repeated application, to optimise the sequences of stages, stage durations and interstage structures fully. The relationships between the variables in the network and the isolated junction formulations are summarised in Table 1.

3. A DECOMPOSITION APPROACH

3.1 Introduction

The approaches adopted for network optimisation and individual junction optimisation differ in a number of respects. The variables used by each are different: in the group-based individual junction methods, offset is not defined, and the stage sequence and interstage structures are implicit variables whereas they are fixed in the network approaches. Furthermore, the level of detail of the traffic model used in the individual junction approaches is greater, often representing separately several streams within a single link of the network approaches.

The decomposition approach presented here seeks to overcome the incompatibilities inherent in these differences in order to take advantage of the additional facilities and capabilities offered by the individual junction approaches when optimising timings in networks. The present approach represents a development of earlier complementary use of signal optimisation procedures on arterial roads by Cohen (1983), Cohen and

Mekemson (1985), and Skabardonis and May (1985). These approaches used arterial optimisation programs, including MAXBAND, to identify appropriate treatment of opposed turning traffic before proceeding to optimise offset and green times using TRANSYT. The approach presented here extends these earlier ones by using group-based optimisation techniques at the individual junction level. This allows advantage to be taken of their greater flexibility than stage-based techniques in the optimisation of sequences and interstage structures. This approach is general in its applicability and appropriate for use in networks of any topology, including arterials, grids, and more irregular ones. In this section, details of the approach are given, and some of its characteristics are noted and discussed.

3.2. The proposed approach

The idea underlying this decomposition approach is straightforward: to design each signal-controlled junction in a network individually and then to adopt those elements of the resulting designs that cannot be optimised simultaneously in the whole network. However, some practical difficulties arise in the implementation of this idea because of the use of a common cycle time throughout the network, the interactions between cycle time and other elements which are not variable within network optimisation procedures, the different levels of detail in modelling, and the different measures of performance used at the individual junction and the network levels. The SIGMA program (Bielefeldt, 1987) uses a procedure similar to this in that junction details are designed first and network-specific items designed later, varying some elements of the junction designs. However, in that method a stage-based formulation is used throughout, and a coarser rather than finer traffic model is used in the optimisation of individual junctions. The procedure presented here is a follows.

First, each junction is analysed individually and signal timings are optimised using a group-based method. For each ordering of the incompatible groups, three optimisations are performed: to find the minimum cycle time at which all flow and timing constraints are satisfied, to find the reserve capacity of the junction, and to find the minimum total rate of delay at the junction under the assumption of uncorrelated arrivals. This yields for each ordering of the incompatible groups three distinct optimised measures of performance: these are normally correlated, and can be used to rank the orderings. Furthermore, the intersection over all the junctions of the admissible ranges of cycle times corresponds to the range of admissible cycle times for the whole network.

Second, the sequence of stages and the interstage structures are extracted from the optimised signal timings for each junction that provide the best performance. These are then introduced into the network model as fixed data. In some cases, a choice will exist for this information according to which of the three objectives of cycle-time, capacity or delay has been optimised. On the one hand, the cycle-time minimising and capacity maximising timings have the advantage that they are equally appropriate at the individual junction and the network levels. However, they are normally underspecified in that these objectives are determined by a subset of groups that control the critical movements and this results in some degree of indeterminacy in the solutions. On the other hand, the delay minimising timings are determined fully by the optimisation, but the estimates of delay on which they are based are not appropriate in coordinated networks.

The third step is to optimise the common cycle time in the network with respect to the network measure of performance. This can be achieved by scanning a range of possible cycle times, optimising approximately the other signal timings at each value (Gartner *et al.*, 1976; Vincent *et al.*, 1980) and selecting the one for which the best performance is achieved. The resulting common cycle time will usually differ from that associated with the stage sequences and interstage structures at the individual junctions. Because of this, optimisation at the individual junctions is repeated with the common cycle time imposed as an equality constraint in order to check that the stage sequence and interstage structures remain appropriate. If a junction is sufficiently lightly loaded that it has a critical cycle time that is less than half of the maximum acceptable one for the network, then it is

a candidate for double cycling. Whether or not this would be beneficial can be determined experimentally.

As the final step, once the stage sequence, interstage structures and common cycle time have been established, a full optimisation of the stage durations and offsets at the network level is undertaken. If the performance of each junction is then found to be satisfactory, the optimisation is complete. Otherwise, further changes are made, making use of the sensitivity information provided by the individual junction methods where appropriate, and associated parts of the optimisation procedure are repeated.

3.3. Advantages of the proposed approach

The principal advantage of this decomposition approach to network optimisation is that it makes available for use in networks of coordinated signals those features which are provided by individual junction optimisation methods. Thus stage sequences and interstage structures can be optimised systematically. Furthermore, any junctions which are found to be critical in the network can be examined in detail and redesigned making full use of the flexibility and sensitivity analysis provided by individual junction methods.

A second advantage arises in consideration of the computational effort involved in evaluating the full range of possible sequences of stages. If each of these were evaluated separately at the network level, then the number of cases to be considered would be equal to the product of the number of different sequences at each of the junctions. Thus the number of sequence combinations n^p for which the whole network would be optimised is given by

$$n^p = \prod_{j=1}^{J} n_j$$

where

J is the number of junctions, and

n_j is the number of sequences at junction j ($1 \leqslant j \leqslant J$).

If, however, the present approach is adopted, then the number of cases that are evaluated — in this case at their respective individual junctions — is equal to the sum of the number of different sequences. Thus the number of sequences n^s for which individual junctions are to be optimised is given by

$$n^s = \sum_{j=1}^{J} n_j$$

The decomposition approach therefore offers a considerable computational advantage in networks that include complicated junctions because the number of optimisations to be performed is substantially reduced, and each optimisation is for a single junction rather than for the whole network.

3.4. Difficulties in the proposed approach

The decomposition approach presented here is subject to a number of methodological difficulties. Principal amongst these is that the method is sub-optimal because the effects of coordination between adjacent junctions are not taken into account when the stage sequence and interstage structures are optimised. In some cases, achieving good coordination can be an overriding consideration, and this can make the partial designs generated by the individual junction methods sub-optimal. Although an optimised network could be inspected to see if some changes might be beneficial, this step has not yet been included in the present approach.

A possible difficulty could arise because of the use of a common cycle time for all signal-controlled junctions in the network. Imposing this on all junctions might lead to the sub-optimality of the sequence and interstages that were generated at each of them individually with a free choice of cycle time. In the present procedure, optimality of the sequences and interstage structures is checked by reoptimising each junction individually

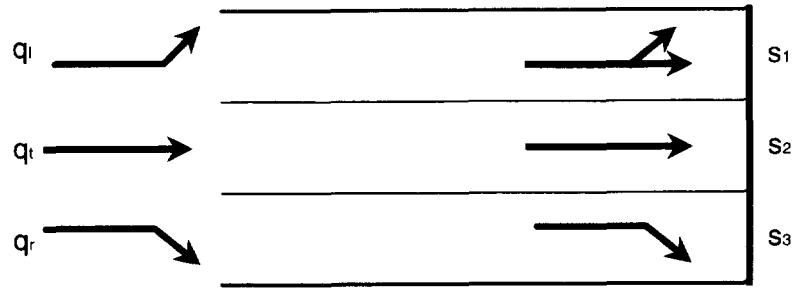

Fig. 2. A three-lane, three movement approach.

with the cycle time constrained to be equal to the common one. If this leads to the selection of new sequences or interstage structures, the common network cycle time can then be re-optimised with the new data and the process repeated. This process does converge because the optimised performance is bounded and improves at each iteration. However, the possible need to revise the sequences and interstage structures several times adds to the complexity of the approach.

A difficulty arises where a single link of a network model is used to represent several streams of traffic. Where this is done, it gives rise to a systematic underestimation of delay because of the economies of scale that apply to queueing systems. The cause of this can be seen as follows. If a stop-line is subdivided for use by two separate streams, then there will be times at which vehicles queue in one stream after the queue in the other has dissipated. The remaining queue will then clear at a time after that at which it would have done if all vehicles could use the whole stop-line, so some extra delay will be incurred. However, modelling both streams as a single link will not reflect this. Thus the amalgamation of streams into a single link is less realistic and causes delays to be underestimated.

To investigate this effect, we consider the three-lane link shown in Fig. 2, where left turners use the left-hand lane, through traffic uses the left and centre lanes, and right turners have exclusive use of the right-hand lane. Let the arrival rates be q_l, q_t, and q_r respectively, and the saturation departure rates during green be s_1, s_2 and s_3 respectively for each of the three lanes. The single link model shown in Fig. 3 has arrival rate $q_l + q_t + q_r$ and saturation departure rate $s_1 + s_2 + s_3$. According to Webster's delay formula for traffic at an isolated junction, the rate of delay on the link will be

$$D_b = \frac{9}{20}\left(\frac{sr^2q}{(s-q)c} + \frac{x^2}{(1-x)}\right)$$

where in this case $s = s_1 + s_2 + s_3$, and $q = q_l + q_t + q_r$.

However, according to the lane usage shown in Fig. 2, left turning and through traffic will form one queue whilst right turning traffic will form another. Thus two streams of traffic are required to model this lane configuration accurately, as is shown in Fig. 4. In this case, the total rate of delay is estimated as the sum of the rates of delay in each of the two streams. Suppose that the flows are such that the degree of saturation x in the two steams are equal, so that

$$\frac{(q_l + q_t)}{(s_1 + s_2)} = \frac{q_r}{s_3} = \frac{q}{s}.$$

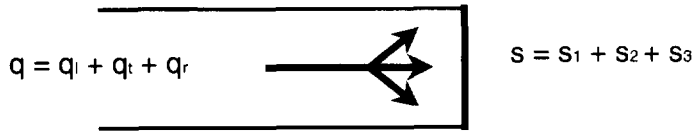

Fig. 3. A single link representation of the approach in Fig. 2.

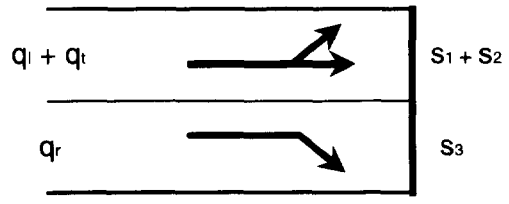

Fig. 4. A two stream representation of the approach in Fig. 2.

The total rate of delay is then given by

$$D_c = \frac{9}{20}\left[\frac{(s_1+s_2)r^2(q_1+q_l)}{(s_1+s_2-q_1-q_l)c} + \frac{x^2}{1-x} + \frac{s_3 r^2 q_r}{(s_3-q_r)c} + \frac{x^2}{1-x}\right]$$

$$= D_b + \frac{9}{20}\left(\frac{x^2}{1-x}\right)$$

so with these flows, the single link model gives rise to an estimate of the rate of delay, D_b, which is less than that which will occur, D_c as estimated by the more detailed two stream model.

Other divisions of flow between the two streams will give results which differ from this to some extent. However, we now show that the delay estimated from a single link model of two separate steams will always under-estimate that incurred. First, we note that the sum of the uniform delay terms is minimised by the allocation of flows to each of the separate streams in proportion to their capacities, as in the case analysed above. Furthermore, the second term of the formula,

$$D_r(x) = \frac{x^2}{1-x}$$

is a convex function of x (the second derivative, $(1-x)^{-3}$, being positive) and hence of q. Thus for any $\alpha \in [0, 1]$,

$$\alpha D_r(x_a) + (1-\alpha)D_r(x_b) \geqslant D_r[\alpha x_a + (1-\alpha)x_b] .$$

Using the specific values

$$\alpha = \frac{(s_1+s_2)}{(s_1+s_2+s_3)} ,$$

$$x_a = \frac{(q_1+q_t)c}{(s_1+s_2)(c-r)} , \text{ and}$$

$$x_b = \frac{q_r c}{s_3(c-r)}$$

in the general inequality gives

$$\frac{(s_1+s_2)}{(s_1+s_2+s_3)} D_r(x_a) + \frac{s_3}{(s_1+s_2+s_3)} D_r(x_b) \geqslant D_r(x)$$

so that

$$D_r(x_a) + D_r(x_b) \geqslant D_r(x)$$

where the inequality is strict except in either of the extreme cases $s_1 + s_2 = 0$ or $s_3 = 0$. This establishes that for any values of q_l, q_t, and q_r, the sum of the random delay terms D_r

in D_c will always exceed the value of the corresponding term of D_b and so $D_c \geqslant D_b$ for any division of traffic between the streams. Similar arguments apply to delay as estimated in networks from other traffic models, including simulation. This will occur because the component of delay estimated from the idealised model of traffic will behave like the first term of Webster's formula whilst any additional component used to represent congestion due to stochastic effects will behave like the second term D_r.

Ideally, a separate link would be used in network models for each stream of traffic at a junction as is done in individual junction models. However, this is not always practical because of the storage and computational implications of modelling at that level of detail. In cases where streams are aggregated to form links in a network model, the resulting estimates of delay incurred by traffic on those links will be systematically low. Conversely, if as is recommended by Akçelik (1984), 2 lanes are modelled separately when in practice a traffic movement uses both of them, then the delay incurred will be overestimated systematically.

The criterion for identification of sets of lanes to form streams is that the traffic in the lanes form a single queue. According to this criterion, two lanes belong to the same stream if the same movement is made by some traffic in each of them. On the other hand, the presence of permissive lane markings for this is insufficient: if, for example, the left turning flow in the example shown in Fig. 2 is so great that no through traffic uses the left lane, then the three lanes will each be used by only one movement and will therefore form three separate streams.

4. EXAMPLE CALCULATIONS

4.1. Introduction

In order to illustrate the use of the approach described in Section 3, it is applied here to a simple example network with four signal-controlled junctions.The network used here is based upon that given in the TRANSYT/8 users' guide (Vincent *et al.*, 1980) and is shown in Fig. 5. The particular optimisation methods used are that of Silcock and Sang (1990) for individual junctions, and TRANSYT/8 for the whole network. For purposes of comparison, the network was also optimised using all the facilities of TRANSYT in the manner recommended by Vincent *et al.* (1980).

The network used here was modified from the original one in two respects. First, all the bus-only links were removed, and second, two additional links, 112 and 1111, were introduced for offside (right) turning traffic at Junction 11 so that each link of the network corresponds to a single stream of traffic. This ensures that the traffic models at the individual junction and the network levels are mutually consistent. The lane configurations adopted for Junctions 1 and 11 are shown in Figs 6 and 7, respectively, together with the arrival rates and saturation departure rates used. Junctions 2 and 7 are relatively simple and were found to offer no scope for optimisation using the individual junction methods.

4.2. Results

First of all, the base network was optimised using the normal TRANSYT method, adopting the stage sequences for Junctions 1 and 11 that are provided with the example, as shown in Figs 6 and 7. The TRANSYT stage reordering procedure STAGOR did not recommend any changes to these sequences. A scan of cycle times in the range 70–180 s for this base network using the CYOP procedure gave the estimates of performance shown in Fig. 8. On the basis of this, full optimisation was undertaken at cycle times of 90 and 100 s. The final optimised performance indexes, corresponding to a weighted sum of stops and delay, are given in Table 2. This performance could certainly be improved, as was indicated by the application of the decomposition approach. However, little indication is available from the TRANSYT results as to how any such improvement could be achieved.

Next, the decomposition approach was applied to the network. No design changes were found to be appropriate at Junctions 2 and 7 because of their relatively straightforward

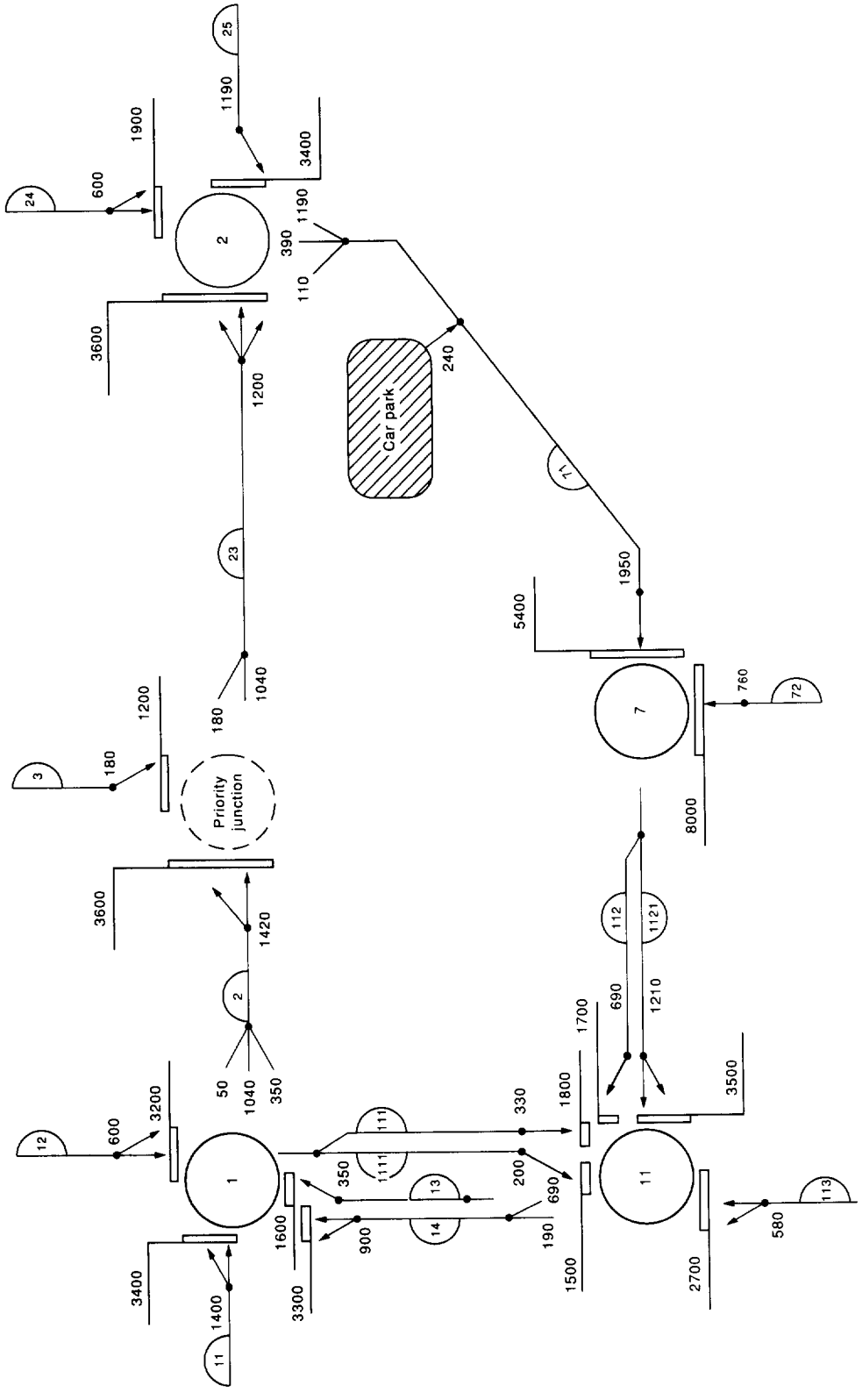

Fig. 5. The example network (after Vincent *et al.*, 1980).

layouts. Thus attention was focused on Junctions 1 and 11. The stage sequences investigated in each case included the TRANSYT sequences and the other sequences shown in Figs 6 and 7, each taken in both forward and reverse directions. A summary of the results of the individual junction optimisations is given in Table 3, where the sequence adopted is shown first for each junction. The sequences adopted at each junction performed better than any other in respect of each of minimum cycle-time, reserve capacity and minimised

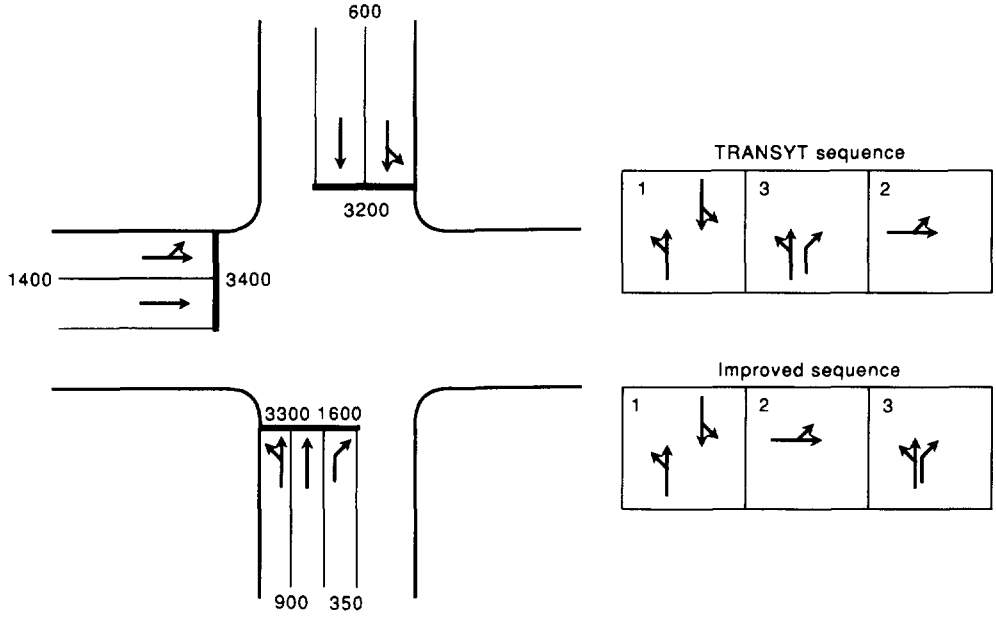

Fig. 6. Lane allocation, arrival and departure rates, and stage sequences at Junction 1.

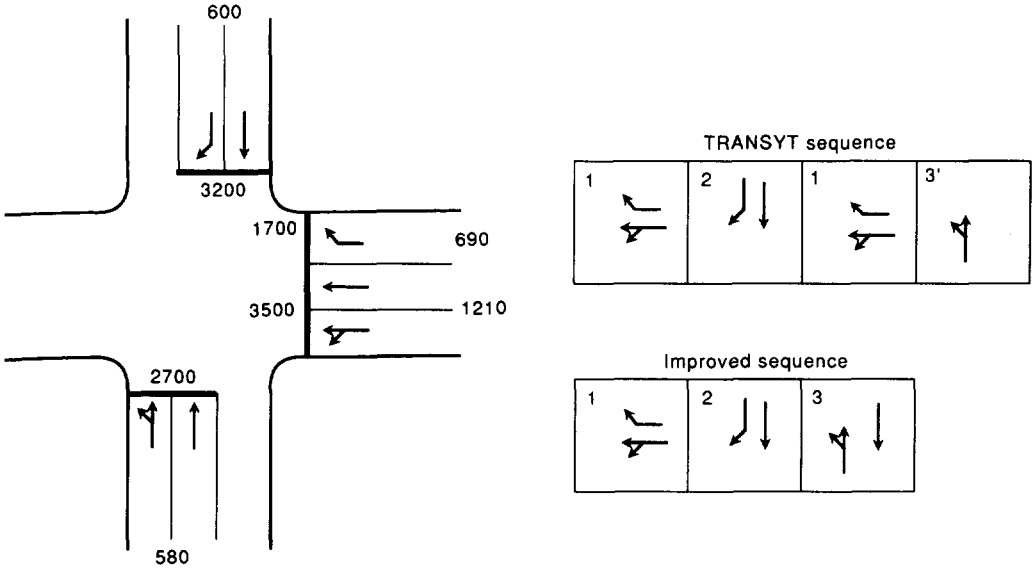

Fig. 7. Lane allocation, arrival and departure rates, and stage sequences at Junction 11.

Table 2. Optimised performance indexes for the example networks

Network	Cycle time (s)	Performance index (£/h)
Base network	90	530
	100	509
Improved sequences	90	429
	80	422

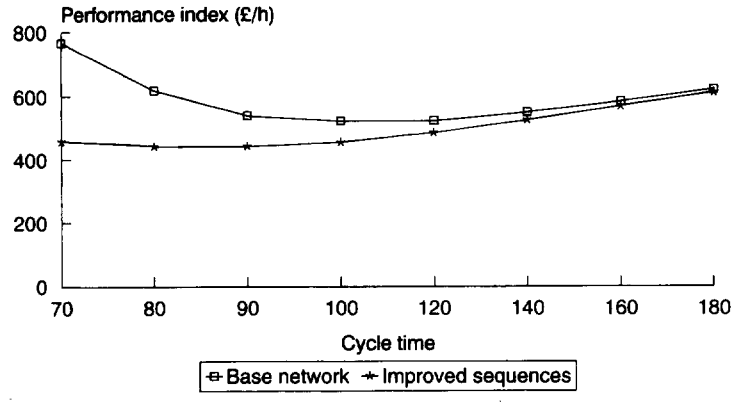

Fig. 8. Performance of the networks at different cycle times estimated by CYOP.

Table 3. Sequences investigated and their performance

Junction	Sequence	Minimum cycle (s)	Reserve capacity (%)	Minimised delay (vehicles)	Cycle time (s)
1	1 2 3 (adopted)	56	6.80	17.15	56
1	1 3 2 (TRANSYT)	65	5.85	19.58	69
11	1 2 3 (adopted)	62	9.41	22.73	69
11	1 3 2	74	7.42	25.03	79
11	1 2 1 3' (TRANSYT)	141	- 2.06 (overloaded)	28.93	120

delay. Advantages were gained from the stage sequences adopted because of asymmetric group intergreen times, as is indicated by the lower minimum cycle times for these sequences than their reverses. In both cases, late starts were used for traffic opposing offside (right) turners.

The stage sequences and associated interstage structures were extracted from the delay minimising signal timings and were re-coded for use in TRANSYT. The delay minimising timings for junction 1 are shown in Fig. 9 by way of example. The protected green indication for the opposed turners on link 13 starts at the beginning of stage 3. This is followed by a permissive green indication in stage 1 during which a period of gap acceptance starts once the opposing queue from link 12 has dissipated: this is about 10 s after the start of that stage. This behaviour is modelled fully in the initial junction optimisation. The layout of the junction allows the effective green period for links 12 and 13 to continue for about 1 s after the end of that for link 14, and this leads to the interstage structure that is illustrated.

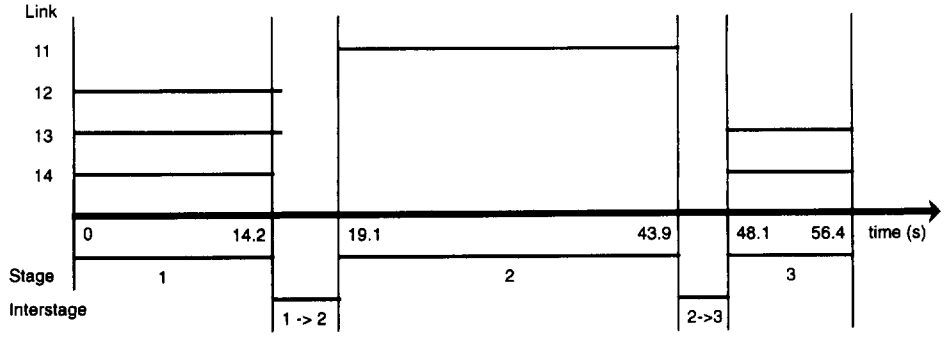

Fig. 9. Delay minimising signal timings for Junction 1.

A scan of common cycle-times for the whole network gave the estimates of performance for the network with improved sequences shown in Fig. 8. This indicates that at each cycle time, the optimised performance of the network with improved sequences is better than that of the base network. Furthermore, it indicates that with the improved sequences, the optimal cycle time is lower, and that the optimised index is less sensitive to variations in cycle time. On the basis of this scan, full optimisation was undertaken at cycle times of 80 and 90 s. Imposing each of these cycle times on the individual optimisation of Junctions 1 and 11 did not lead to any changes in sequence or interstage structures.

The results of the full TRANSYT optimisations at the cycle times of 80 and 90 s are given in Table 2. At the cycle time of 90 s, at which each of the base and the improved networks was optimised, the improvement in performance achieved using the decomposition approach was some 19%. Optimisation of the cycle time in each case led to the use of 100 s with the base sequences and 80 s with the improved ones. The improvement associated with use of the decomposition approach was reduced to 17% with this independent choice of cycle time. This lesser improvement in the latter case arises because the network with improved sequences is less sensitive to deviation from the optimal cycle time.

5. DISCUSSION AND CONCLUSIONS

The decomposition approach to optimisation of networks of traffic signals that is presented in this paper provides a systematic method by which details of the junction and signal control designs can be improved. This approach takes advantage of the power and flexibility of individual junction optimisation procedures whilst using them in a network context. It allows for full sequence optimisation, irrespective of the layout and complexity of the junctions. Furthermore, it is applicable in networks of arbitrary topology.

The approach presented here is sub-optimal in that some important effects are not considered throughout the optimisation of all of the variables, and the objective functions used in the two optimisation steps differ. Notwithstanding this, the method has proved to be effective in improving the operational performance of a small example network. The benefits achieved using this approach include improved network performance, reduced sensitivity of the optimised performance to variations in cycle time and a reduced optimal cycle time, which leads to greater flexibility of control as a greater range of cycle times is made available. This flexibility is especially important in networks throughout which a common cycle time is used because the requirements of any individual junction can influence the performance of all the others.

In addition to these benefits of improved network performance, the application of individual junction techniques within a network has implications for the practice of design. In particular, all the post-optimality information which is available from these methods can be used to aid the improvement of junction designs. The information concerning the sensitivity of minimum cycle time and reserve capacity is appropriate to individual junctions whether or not they are part of coordinated networks. However, the corresponding information concerning the sensitivity of optimised delay is only strictly appropriate to isolated junctions as any platooning effects due to adjacent signal-controlled junctions are not considered.

The differences in the level of detail normally used in individual junction methods and network methods can cause systematic differences in estimates of performance and optimised signal timings. In order to achieve consistency between the traffic models used at the individual junction and the network levels, separate links of the network should be used for each stream of traffic. This will lead to more realistic modelling of traffic behaviour, and consequently to increases in the estimates of delay. In cases where this is not possible, the network traffic models will underestimate delays on links that are used to represent several streams.

The approach of applying optimisation methods for individual fixed-time signal-controlled junctions to junctions in networks can be extended in several respects. The

operation of traffic-responsive signals requires some parameters to be set in advance, including which changes of stage are to be allowed to occur and the structure of the associated interstages. These parameters could be investigated using fixed-time optimisation techniques and, analogously to the method presented in the present paper, the optimised results could then be adopted for use in the different context of traffic responsive operation.

Acknowledgement—The author is grateful to an anonymous referee for his helpful comments on an earlier draft of this paper.

REFERENCES

Akçelik R. (1984) SIDRA—2 does it lane by lane. *Proc. 12th ARRB Conf.* **12**, 137–149.
Akçelik R. (1991) SIDRA 4.0 software status. *Traff. Engng Control* **32**, 585–589.
Allsop R. E. (1992) Evolving application of mathematical optimisation in design and operation of individual signal-controlled junctions. *Mathematics in Transport Planning and Control* (Griffiths, J. D. Ed.), pp. 1–25. Oxford University Press, Oxford.
Bielefeldt C. (1987) *SIGMA—A New Program for Optimising Signal Timings*. Heusch Boesefeldt GmbH, Aachen.
Cohen S. L. (1983) Concurrent use of MAXBAND and TRANSYT signal timing programs for arterial signal optimization. *Transpn Res. Rec.* **906**, 81–84.
Cohen S. L. and Mekemson J. R. (1985) Optimization of left-turn phase sequence on signalized arterials. *Transpn Res. Rec.* **1021**, 53–58.
Crabtree M. R. (1988) TRANSYT/9 Users' manual. Transport and Road Research Laboratory Report AG8. TRL, Crowthorne.
Gartner N. H., Little J. D. C. and Gabbay H. (1975a) Optimisation of traffic signal settings by mixed-integer linear programming. Part I: the network coordination problem. *Transpn Sci.* **9**, 321–343.
Gartner N. H., Little J. D. C. and Gabbay H. (1975b) Optimisation of traffic signal settings by mixed-integer linear programming. Part II: the network synchronization problem. *Transpn Sci.* **9**, 344–363.
Gartner N. H., Little J. D. C. and Gabbay H. (1976) Simultaneous optimisation of offsets, splits and cycle time. *Transpn Res. Rec.* **596**, 6–15.
Heydecker B. G. (1992) Sequencing of traffic signals. *Mathematics in Transport Planning and Control* (Griffiths J. D. Ed.), pp. 45–56. Oxford University Press, Oxford.
Heydecker B. G. and Dudgeon I. W. (1987) Calculation of signal settings to minimise delay at a junction. *Transportation and Traffic Theory* (Gartner N. H. and Wilson N. H. M. Eds), pp. 159–178. Elsevier, Amsterdam.
Improta G. and Cantarella G. E. (1984) Control system design for an individual signalized junction. *Transpn Res.* **18B**, 147–167.
Little J. D. C., Kelman M. D. and Gartner N. H. (1981) MAXBAND: a program for setting signals on arterials and triangular networks. *Transpn Res. Rec.* **795**, 40–46.
Möller K. (1987) Calculation of optimum fixed-time signal programs. *Transportation and Traffic Theory* (Gartner N. H. and Wilson N. H. M., Eds), pp. 179–198. Elsevier, Amsterdam.
Rouphail N. M. and Radwan A. E. (1990) Simultaneous optimization of signal settings and left-turn treatments. *Transpn Res. Rec.* **2187**, 1–10.
Silcock J. P. (1992) Phase-based optimisation of isolated signal-controlled junctions: sensitivity analysis and a treatment of double green phases. *Mathematics in Transport Planning and Control* (Griffiths J. D. Ed.), pp. 45–55. Oxford University Press, Oxford.
Silcock J. P. and Sang A. J. (1990) SIGSIGN: a phase-based optimisation program for individual signal-controlled junction. *Traff. Engng Control* **31**, 291–298.
Skabardonis A. and May A. D. (1985) Comparative analysis of computer models for arterial signal timings. *Transpn Res. Rec.* **1021**, 45–52.
Tully I. M. S. N. Z. (1976) Synthesis of sequences for traffic signal controllers using techniques of the theory of graphs. Ph.D. thesis, University of Oxford.
Vincent R. A., Mitchell A. I. and Robertson D. I. (1980) User guide to TRANSYT version 8. Transport and Road Research Laboratory Report LR888. TRL, Crowthorne.
Webster F. V. (1958) Traffic signal settings. Road Research Technical Paper, 39. HMSO, London.