GROUP-BASED OPTIMIZATION OF SIGNAL TIMINGS FOR TRAFFIC EQUILIBRIUM NETWORK

S.C. WONG
Associate Professor
Department of Civil Engineering
The University of Hong Kong
Pokfulam Road, Hong Kong
P.R. China
Fax: +852-2559-5337
E-mail: hhecwsc@hkucc.hku.hk

C. YANG
Research Fellow
Division of Geotechnical and
Transportation Engineering
School of Civil & Structural Engineering
Nanyang Technological University
Nanyang Avenue, Singapore 639798
Fax: +65-792-1650
Email: ccyang@ntu.edu.sg

C.O. TONG
Associate Professor
Department of Civil Engineering
The University of Hong Kong
Pokfulam Road, Hong Kong
P.R. China
Fax: +852-2559-5337
E-mail: cotong@hku.hk

C.K. WONG
Graduate Student
Department of Civil Engineering
The University of Hong Kong
Pokfulam Road, Hong Kong
P.R. China
Fax: +852-2559-5337
E-mail: a9595583@graduate.hku.hk

Abstract: In this paper, we study the optimization of signal settings for equilibrium network based on the TRANSYT traffic model, a well-known procedure for the evaluation of queues and delays in a signal-controlled network. The group-based notation is used to specify the signal control variables and constraints. A sensitivity analysis is carried out to determine the derivatives of performance index, the total travel time in the network, with respect to the signal parameters. These derivatives are then used in an integer programming method to derive the discrete search direction for optimizing the equilibrium signal settings. An optimization procedure, combining the integer programming method and hill-climbing method, is developed to solve the problem. Numerical examples are used to demonstrate the effectiveness of the proposed method.

Key Words: combined signal and assignment problem, group-based optimization method, traffic signals, network equilibrium

1. INTRODUCTION

Substantial research have been carried out for the optimization of signal plans for signal-controlled networks (Allsop, 1968a,b; Gartner, 1974; Hillier, 1965, 1966; Robertson, 1969; Traffic Research Corporation, 1966; Vincent, 1980; Wong, 1995, 1996, 1997; Wong *et al.*, 2000). All these methods assumed that the link flows in the network are fixed. However, many people have already pointed out the fact that equilibrium flow pattern of a network is strongly related to the traffic signal settings. By changing the control strategies, the traffic will arrange itself in an user optimal manner (Wardrop, 1952), for which users traveling from any origin to any destination react to the new control strategy by choosing paths such that their individual costs are minimized. This redistribution effect on equilibrium traffic flow pattern will affect the performance of the network. Allsop (1974) was amongst the first who suggested that the signal control can be explored to affect the distribution and assignment of traffic on an equilibrium network, and provided a rigorous mathematical framework for the problem. This problem has been known as equilibrium traffic signal settings in the literature.

Recently, attempts have been made to solve the equilibrium traffic signal settings. Yang and Yagar (1995) studied the problem of traffic assignment and signal control in saturated road networks with capacity side constraints. Wong and Yang (1997) extended the concept of reserve capacity, originally applied to individual junctions (Allsop, 1972), to a signal-controlled network. These methods, however, do not consider the signal coordination effects in the network. A very early study has already shown that substantial benefits can be achieved by coordinating adjacent signals (Holroyd and Hillier, 1971). For modeling these coordination effects, the TRANSYT traffic model was developed and is now widely used for the estimation of queues and delays in a signal-controlled network (Robertson, 1969; Vincent, 1980). This traffic model, however, does not consider the re-routing behavior of road users. To extend the traffic model for equilibrium traffic signal settings, a path-based assignment algorithm was recently formulated for the determination of the user equilibrium traffic pattern using TRANSYT traffic model for the evaluation of queues and delays in the network (Wong *et al.*, 2001). Signal optimization was not considered in this paper. In general, signal optimization for equilibrium traffic network is broadly classified into iterative optimization and assignment approach and global optimization approach (Yang and Yagar, 1995). Based on the work by (Wong *et al.*, 2001), an iterative optimization and assignment scheme was developed to determine mutually consistent signal settings (Wong and Yang, 1999). However, an optimized signal plan is not guaranteed to obtain with this method (see Allsop (1974)).

In this paper, we develop a global optimization method for the determination of optimized signal settings for the equilibrium network based on the TRANSYT traffic model. The group-based notation is used to specify the signal control variables and constraints (Gallivan and Heydecker, 1988; Heydecker and Dudgeon, 1987; Silcock, 1997; Wong, 1996). A sensitivity analysis is conducted to determine the derivatives of performance index, the total travel time in the network, with respect to the signal parameters. Good agreement of results with the values determined by numerical differentiation is obtained. The derivatives are then used in an integer programming method to obtain the discrete search direction for optimizing the signal settings. An optimization procedure, combining the integer programming method and hill-climbing method, is developed to solve the problem.

In Section 2, the group-based notation is introduced. The TRANSYT traffic model is briefly discussed in Section 3 and the path-based assignment algorithm is summarized in Section 4. The sensitivity analysis and the group-based optimization procedure are given in Sections 5 and 6 respectively. Numerical examples are shown in Section 7 to demonstrate the effectiveness of the proposed method.

2. GROUP-BASED VARIABLES AND CONSTRAINTS

The group-based control variables for a signal-controlled network are given as follows (for more details, interested readers are referred to Wong (1996)). The junctions are operated at a common cycle time. The period during which a particular signal group at a junction has right of way is specified by two control variables: the start and duration of green for the signal group. The start of green is measured from a master clock for all junctions. All the control variables are expressed in time units. The offsets of junctions are very common variables in most linked signal calculations, but it is one of the advantages of group-based method that the offset variables have been implicitly included in the group-based control variables. These group-based variables are usually subject to the following constraints:

- For the case of unspecified cycle time, the common cycle time is also considered as a control variable and is confined to a certain practicable range.
- A minimum acceptable duration of green indication is usually specified for a vehicular or pedestrian stream. For a pedestrian stream, this minimum duration depends on the width of crossing and the walking speed of pedestrians.
- For any two incompatible signal groups, a clearance time is required between the end of green of a signal group and the start of green of another group so that all the vehicles from the former group have left the conflict points before the vehicles from the latter group arrive.
- When a traffic stream operates near to its capacity, delay to vehicles is substantial. Sometimes, one may need to specify the maximum acceptable degree of saturation for a traffic stream so as to make sure the stream is always operating below a certain acceptable congestion level.
- For reasons of practicability, there may be some constraints to be imposed on the relative timing of starts and ends of green for different signal groups.

3. THE TRANSYT TRAFFIC MODEL

The TRANSYT traffic model (Robertson, 1969; Vincent *et al.*, 1980) simulates the movement of traffic through a network and takes into account of the effect of platoon dispersion. It is a widely used procedure to determine the queues and delays in a signal-controlled network with explicit consideration of the signal coordination effects. However, the traffic model does not consider the re-routing of traffic in the network in response to the traffic conditions. In this paper, the TRANSYT traffic model is employed for the evaluation of delays in the network, which forms the basic module of the assignment problem discussed in the paper. In our model, the cruise time on a link is fixed, but the delays at the end of the link depends on the flows of many other links in the network. The travel time on a link, therefore, consists of two components: the cruise time and the delay at the end of the link. It can be expressed as

$$t_a(\mathbf{v}) = t_a^c + d_a(\mathbf{v}), \quad \forall a \in A \tag{1}$$

where $d_a(\mathbf{v})$ is the delay at the end of a link a, which is determined by the TRANSYT traffic model, and $\mathbf{v} = (v_a, a \in A)$ is the set of link flows in the network. This travel time function is generally asymmetric with respect to the link flows, i.e.

$$\frac{\partial t_a(\mathbf{v})}{\partial v_{a'}} \neq \frac{\partial t_{a'}(\mathbf{v})}{\partial v_a} \tag{2}$$

In the traffic model, the delay consists of two components, i.e.

$$d_a(\mathbf{v}) = d_a^u(\mathbf{v}) + d_a^r(v_a) \tag{3}$$

where $d_a^u(\mathbf{v})$ and $d_a^r(v_a)$ are the uniform and random-and-oversaturation delays respectively. The uniform delay represents the delay incurred with an identical pattern of traffic arriving during every cycle, and the random-and-oversaturation delay takes into account respectively the variations in traffic arrivals from cycle to cycle and the steady increase in queues on oversaturation links.

Before discussing these delays, the following traffic patterns as functions of time during a cycle are defined to describe how the vehicles arrive at and depart from a link: (i) IN pattern: the pattern of traffic that would arrive at the stop line at the end of the link if the traffic were not impeded by the signals at the stop line; (ii) OUT pattern: the pattern of traffic leaving a link; (iii) GO pattern: the pattern of traffic that would leave the stop line if there was enough traffic to saturate the green. These definitions were employed in Vincent *et al.* (1980). To facilitate discussions in the subsequent sections, we also define a traffic pattern in this paper: (iv) EN pattern: the pattern of traffic entering a link. The uniform components of delays were obtained through simulation of two cycles of the IN, OUT, EN and GO patterns to obtain the queue formation patterns of all links, which were then used to calculate the uniform delays. For the random-and-oversaturation delays, sheared delay formulae were employed to estimate their values.

In this paper, the performance index of the network is defined as the total travel time incurred by all users in the network as:

$$T = \sum_{a \in A} v_a \left(t_a^c + d_a^u(\mathbf{v}) + d_a^r(v_a) \right) \tag{4}$$

where the uniform and random-and-oversaturation delays are determined in the following subsections. Let i_a, o_a, g_a and e_a be respectively the IN, OUT, GO and EN patterns on a link a. The GO pattern depends on the link characteristics such as saturation flow and the signal settings, which is assumed to be unchanged for a given signal plan. During the simulation process, the OUT pattern from the link is determined from the IN and GO patterns

$$o_a = G(i_a, g_a) \tag{5}$$

In a network, the EN pattern is affected by all the OUT patterns from the upstream links. Therefore, we can write

$$e_a = H(o_b, b \in B_a) \tag{6}$$

where B_a is the set of upstream links of link a. The traffic entering into a link will travel along the link in accordance with a linear recursive platoon dispersion function to determine the IN pattern as

$$i_a = J(e_a) \tag{7}$$

This linear recursive platoon dispersion function is additive, i.e.

$$J(A+B) = J(A) + J(B) \tag{8}$$

This property is very useful in the sensitivity analysis that will be discussed in Section 5. For a network with a simple tree structure, the traffic model only needs to iterate once using equations (5-7). However, for a general network structure, the traffic model has to iterate until the traffic patterns stabilize. Then the uniform delays can be calculated from the stabilized (or converged) traffic patterns as

$$D_a^u = \Gamma(i_a^*, g_a) \tag{9}$$

where D_a^u is the uniform rate of delay on link a, and i_a^* is the stabilized IN pattern on link a. The average uniform delay can then be obtained as

$$d_a^u = \frac{D_a^u}{v_a} \tag{10}$$

The random-and-oversaturation delay is evaluated based on a sheared delay formula used in Wong (1995) taking into account of the initial random-and-oversaturation queue length for a prescribed interval. The formula was derived by applying the coordinate transformation method (Kimber and Hollis, 1979). The random-and-oversaturation rate of delay for a link is estimated by the following:

$$D_a^r = \frac{1}{2}\left\{\sqrt{U_a^2 + V_a} - U_a\right\} \tag{11a}$$

where

$$U_a = \frac{(1-\rho_a)(\mu_a\tau)^2 + 2(2C\rho_a - L_a^0)\mu_a\tau + 8CL_a^0}{2(\mu_a\tau - 2C)} \tag{11b}$$

$$V_a = \frac{2C(2L_a^0 + \rho_a\mu_a\tau)^2}{\mu_a\tau - 2C} \tag{11c}$$

D_a^r is the random rate of delay on link a, μ_a is the capacity of the link a which depends on the saturation flow and signal settings, L_a^0 is the initial queue length at the start of the duration of analysis for link a, $\rho_a = v_a / \mu_a$ is the traffic intensity (or degree of saturation) for link a, τ is the duration of the analysis, and C is the constant in the Pollaczek-Khintchine formula (Kendall, 1951) for which a value between 0.5 and 0.6 was suggested by Branston (1978). For application in area traffic control, further research is required to calibrate the value of C so that a more accurate prediction of random-and-oversaturation delay can be obtained. The average random delay can be estimated as

$$d_a^r = D_a^r / v_a \tag{12}$$

4. NETWORK EQUILIBRIUM FOR TRANSYT TRAFFIC MODEL

Let $K \subset N \times N$ be the set of OD pairs. Denote q_k as the demand between OD pair k, $k \in K$. These demands, when assigned onto the network, give rise to link flows $v_a, a \in A$. Let P_k be the set of paths between OD pair k, and $P = \bigcup_{k \in K} P_k$ be the set of all paths in the network. Denote f_{kp} as the path flow on path p between OD pair k. We have the flow conversation equations as

$$\sum_{p \in P_k} f_{kp} = q_k, \quad \forall k \in K \tag{13}$$

The corresponding link flows are given by

$$v_a = \sum_{k \in K} \sum_{p \in P_k} \delta_{kpa} f_{kp}, \quad \forall a \in A \tag{14}$$

where

$$\delta_{kpa} = \begin{cases} 1 & \text{if link } a \text{ belongs to path } p \text{ between OD pair } k \\ 0 & \text{otherwise} \end{cases} \tag{15}$$

is the link-path incidence variable. The travel time along a path can be expressed as

$$h_{kp} = \sum_{a \in A} \delta_{kpa} t_a, \quad \forall p \in P_k, k \in K \tag{16}$$

where h_{kp} is the travel time on path p between OD pair k. Let u_k be the least travel time between OD pair k. We have

$$u_k = \min_{p \in P_k} h_{kp} \tag{17}$$

A network equilibrium that satisfies Wardrop's user optimal principle for path flows is achieved when

$$\text{if } f_{kp} > 0 \quad \text{then } h_{kp} = u_k, \quad \forall p \in P_k, k \in K \tag{18a}$$

$$h_{kp} \geq u_k, \quad \forall p \in P_k, k \in K \tag{18b}$$

$$f_{kp} \geq 0, \quad \forall p \in P_k, k \in K \tag{18c}$$

In a recent paper (Wong *et al.*, 2001), the network equilibrium problem was formulated as the following minimization program:

$$\underset{\mathbf{f},\mathbf{u}}{\text{minimize}}\ Z(\mathbf{f},\mathbf{u}) = \sum_{k \in K} \sum_{p \in P_k} \left(h_{kp} - u_k\right) f_{kp} \tag{19a}$$

subject to

$$\sum_{p \in P_k} f_{kp} - q_k = 0, \quad \forall k \in K \tag{19b}$$

$$h_{kp} - u_k \geq 0, \quad \forall p \in P_k, k \in K \tag{19c}$$

$$u_k \geq 0, \quad \forall k \in K \tag{19d}$$

$$f_{kp} \geq 0, \quad \forall p \in P_k, k \in K \tag{19e}$$

where $\mathbf{f} = (f_{kp}, p \in P_k, k \in K)$ and $\mathbf{u} = (u_k, k \in K)$. This minimization program was solved by a Frank-Wolfe algorithm. Using the assignment results, the performance index can now be written in an abstract form as $T(\Phi, \mathbf{f}(\Phi))$.

5. SENSITIVITY ANALYSIS

In this section, the change in performance index of the network with respect to the change in signal parameter, taking into account the signal progression and route changing characteristics, is determined by a sensitivity analysis. Denote ε as a signal parameter in the network. The derivative of the performance index with respect to this parameter can be expressed as

$$\nabla_\varepsilon T(\Phi,\mathbf{f}(\Phi)) = \nabla_\varepsilon T(\Phi,\mathbf{f}^*) + \left\{\nabla_\mathbf{f} T(\Phi^*,\mathbf{f})\right\}^t \left\{\nabla_\varepsilon \mathbf{f}(\Phi)\right\} \tag{20}$$

where

$$\nabla_\mathbf{f} T(\Phi^*,\mathbf{f}) = \mathrm{Col}\left(\frac{\partial T}{\partial f_{kp}}, \forall p \in P_k, k \in K\right) \tag{21}$$

$$\nabla_\varepsilon \mathbf{f}(\Phi) = \mathrm{Col}\left(\frac{\partial f_{kp}}{\partial \varepsilon}, \forall p \in P_k, k \in K\right) \tag{22}$$

and the superscript "t" denotes the transpose of a matrix or vector. The quantities on the right-hand side of equation (25) are determined in the following subsections.

5.1 Determination of $\nabla_\varepsilon T(\Phi,\mathbf{f}^*)$

The derivatives of performance index with respect to the group-based control variables for TRANSYT traffic model, $\nabla_\varepsilon T(\Phi,\mathbf{f}^*)$, have been obtained in Wong (1995), in which approximate expressions for upstream, downstream and further downstream links taken into account platoon dispersion in the traffic model were derived. These derivatives have been used in the group-based optimization of signal timings for a signal-controlled network with given link flows (Wong, 1996). The impact of re-routing on the performance index with respect to change in signal parameters is modeled by the second term in equation (20).

5.2 Determination of $\nabla_\mathbf{f} T(\Phi^*,\mathbf{f})$

Combining equations (4, 9, 11), the performance index can be written as

$$T = T^c + T^u + T^r \tag{23}$$

where

$$T^c = \sum_{a\in A} v_a t_a^c \tag{24}$$

$$T^u = \sum_{a\in A} D_a^u(\mathbf{v}) \tag{25}$$

$$T^r = \sum_{a\in A} D_a^r(v_a) \tag{26}$$

The derivatives can also be written as a sum of three components as

$$\frac{\partial T}{\partial f_{kp}} = \frac{\partial T^c}{\partial f_{kp}} + \frac{\partial T^u}{\partial f_{kp}} + \frac{\partial T^r}{\partial f_{kp}} \tag{27}$$

The first term can be determined by

$$\frac{\partial T^c}{\partial f_{kp}} = \sum_{a\in A} \delta_{kpa} t_a^c \tag{28}$$

which is the cruise travel time along the path concerned. The third term can be expressed as

$$\frac{\partial T^r}{\partial f_{kp}} = \sum_{a \in A} \delta_{kpa} \frac{\partial D_a^r(v_a)}{\partial v_a} \tag{29}$$

which is the sum of derivatives of the random rate of delay on a link with respect to its link flow along the path concerned. The analytical expression for the derivative on a link, $\partial D_a^r / \partial v_a$, can be found in Wong (1995).

Due to the interaction among different links in the network, simple expressions as the first and third terms are not available for the second term in equation (27). A post-simulation sensitivity analysis of the traffic pattern in the network is used. The analysis is a first order approximation to the derivatives based on the stabilized traffic patterns obtained from the TRANSYT traffic model. Consider a path in the network composed by a sequence of links, $a_1, a_2, \ldots, a_w$. Usually, the first link connected to an origin is also a boundary link in the TRANSYT network. The IN pattern on the first link is assumed to increase by a small flow Δf. Since this link is on the boundary of the network, the increased flow is assumed to be uniformly distributed. Therefore, the change in IN pattern becomes $\Delta i_{a1} = \Delta f$ over the whole cycle. The derivative of the uniform delay on the first link with respect to the path flow concerned can then be determined by the following

$$\frac{\partial D_{a1}^u}{\partial f_{kp}} = \frac{\Gamma(i_{a1}^* + \Delta i_{a1}, g_{a1}) - D_{a1}^{u*}}{\Delta f} \tag{30}$$

where i_{a1}^* is the stabilized IN pattern on link a_1, and D_{a1}^{u*} is the uniform delay on the first link a_1 evaluated at the stabilized traffic patterns of the TRANSYT traffic model. The change in EN traffic pattern the second link a_2 can be approximated by

$$\Delta e_{a2} = G(i_{a1}^* + \Delta i_{a1}, g_{a1}) - o_{a1}^* \tag{31}$$

where o_{a1}^* is the stabilized OUT pattern on link a_1. Now applying the platoon dispersion, the change in IN pattern on link a_2 becomes

$$\Delta i_{a2} = J(\Delta e_{a2}) \tag{32}$$

since the platoon dispersion function is linearly addictive by equation (8). The derivative for the second link can then be evaluated by

$$\frac{\partial D_{a2}^u}{\partial f_{kp}} = \frac{\Gamma(i_{a2}^* + \Delta i_{a2}, g_{a2}) - D_{a2}^{u*}}{\Delta f} \tag{33}$$

The process continues for all the subsequent links in a similar fashion. The second term in equation (27) can be determined as

$$\frac{\partial T^u}{\partial f_{kp}} = \sum_{a=a_1,a_2,\ldots,a_w} \frac{\partial D_a^u}{\partial f_{kp}} \tag{34}$$

5.3 Determination of $\nabla_\varepsilon \mathbf{f}(\Phi)$

The derivatives, $\nabla_\varepsilon \mathbf{f}(\Phi)$, are determined following the work from Tobin and Friesz (1988). From the equilibrium solution obtained in Section 4, we first determine the non-degenerate extreme point of positive path flow solutions and delete the non-binding constraints using the approach in Tobin and Friesz. The system of equations will then be reduced to:

$$\hat{\mathbf{h}} - \hat{\Lambda}^t \mathbf{u} = 0 \tag{35}$$

$$\hat{\Lambda}\hat{\mathbf{f}}^* - \mathbf{q} = 0 \tag{36}$$

where $\hat{\mathbf{f}}^*$ is the non-degenerate extreme point of equilibrium path flow solution, $\mathbf{u}$ is the vector of minimum OD travel times, $\mathbf{q}$ is the vector of OD demands, $\hat{\mathbf{h}}$ is the vector of path travel times, and $\hat{\Lambda}$ is the OD-path incidence matrix. Performing a perturbation analysis with respect to a signal parameter ε, we have

$$\begin{Bmatrix} -\nabla_\varepsilon \hat{\mathbf{h}} \\ 0 \end{Bmatrix} = \begin{bmatrix} \nabla_{\mathbf{f}} \hat{\mathbf{h}} & -\hat{\Lambda}^t \\ \hat{\Lambda} & 0 \end{bmatrix} \begin{Bmatrix} \nabla_\varepsilon \hat{\mathbf{f}} \\ \nabla_\varepsilon \mathbf{u} \end{Bmatrix} \tag{37}$$

We can then show that

$$\nabla_\varepsilon \hat{\mathbf{f}} = -\nabla_{\mathbf{f}} \hat{\mathbf{h}}^{-1} \left[\mathbf{I} - \hat{\Lambda}^t \left[\hat{\Lambda} \nabla_{\mathbf{f}} \hat{\mathbf{h}}^{-1} \hat{\Lambda}^t \right]^{-1} \hat{\Lambda} \nabla_{\mathbf{f}} \hat{\mathbf{h}}^{-1} \right] \nabla_\varepsilon \hat{\mathbf{h}} \tag{38}$$

where $\nabla_{\mathbf{f}} \hat{\mathbf{h}}$ can be obtained from a post-simulation sensitivity analysis developed in Wong *et al.* (2001). The elements in the vector $\nabla_\varepsilon \hat{\mathbf{h}}$ are determined by

$$\frac{\partial \hat{h}_{kp}}{\partial \varepsilon} = \sum_{a \in A} \delta_{kpa} \left(\frac{1}{v_a} \frac{\partial D_a^u}{\partial \varepsilon} + \frac{1}{v_a} \frac{\partial D_a^r}{\partial \varepsilon} \right) \tag{39}$$

where $\hat{h}_{kp}$ is the travel time on path p between OD pair k, and $\partial D_a^u / \partial \varepsilon$ and $\partial D_a^r / \partial \varepsilon$ are determined by the formulas for upstream, downstream and further downstream links in Wong (1995).

6. GROUP-BASED OPTIMIZATION METHOD

6.1 Integer Programming Method

The application of integer programming method to optimize the network signal settings for the case of fixed link flows was studied in Wong (1996). Given the derivatives obtained from the sensitivity analysis in Section 5, the same method can also be applied. At a particular feasible solution point, the discrete search direction together with the maximum move size are determined, in which the whole problem is divided into a number of much smaller subprograms, one for each junction. Having obtained the discrete search direction and the corresponding maximum move size, the performance index is then explored along the search direction until the minimum is found. The procedure is repeated until no improvement in the performance index can be obtained. Note that during the search process, a complete traffic assignment based on the path-based algorithm in Section 4 is conducted at every point of searching to evaluate the performance index at equilibrium condition.

6.2 Hill-Climbing Method

When adjustment of offsets only is considered, the hill-climbing technique in TRANSYT (Vincent *et al.*, 1980) is used. The performance index is evaluated for a forward increment of the junction offset. If the performance index decreases as a result of this forward increment, the search is continued in this direction. However, if the performance index increases as a result of the forward increment, the search is carried out backwards. The maximum move size α_{max} is taken as infinity in this case because although the performance index is periodic in each offset, it is useful for the search process to step through successive cycles. The offsets at all junctions are changed in turn as happens in TRANSYT, until all the junctions are handled.

6.3 Optimization Procedure

In order to avoid being trapped into some poor local minimums, large and small hill-climbing steps are used in the optimization procedure (Vincent *et al.*, 1980). This provides a mechanism for the algorithm to jump from the neighborhoods of poor local minimums to those of better ones. The reallocation of green times and cycle time optimization are carried out by means of an integer programming method. An optimization procedure similar to that suggested in TRANSYT version 8 and in Wong (1996) is proposed, which is listed in Table 1.

Table 1. The Optimization Procedure for the Group-Based Method

Step	Procedure	Description of optimization steps
1	H15	Use large and small step sizes for offsets in hill-climbing method to find a better initial solution for the problem
2	H40	
3	N 0 or N 2	Reallocation of green times and cycle time selection by integer programming method
4	H15	Use large and small step sizes for offsets to provide a chance to jump to other better local minimums
5	H40	
6	H 1	Use the finest step size for offsets to search for local minimum
7	N 0 or N 2	Reallocation of green times and cycle time selection by integer programming method
8	H 1	Final check for optimality of solution

Note:

N 0 - Integer programming method for reallocation of green times (Section 6.1)

N 2 - Integer programming method for reallocation of green times and cycle time selection (Section 6.1)

H *x*- Hill-climbing method for offsets only (Section 6.2)
where l = length of hill-climbing step in seconds

$$= \begin{cases} [xc/100] & \text{if } x > 1 \\ 1 & \text{if } x = 1 \end{cases}$$

In the table, the first two steps use large and small step sizes for offsets to find a better initial solution for subsequent analysis. In Step 3, the integer programming method is employed to locate the neighborhood of a minimum point for further optimization. For the case of unspecified cycle time, Step 3 also includes the cycle time variable in optimization using the parallel searching technique. This enables a better cycle time to be determined during optimization. In Steps 4 and 5, again both large and small hill-climbing step sizes are used to provide a chance for the algorithm to jump from the neighborhood of a poor local minimum to that of a better minimum. To search for the local minimum, Step 6 uses the finest hill-climbing step for offsets (1 second) and Step 7 tries to reallocate the green times and/or cycle time selection again. These two steps enable the algorithm to locate the stationary point corresponding to the local minimum. Finally, the finest hill-climbing step is employed in Step 8 to check for the optimality of the solution.

7. COMPUTATIONAL RESULTS

Consider a network shown in Figure 1 consisting of 9 signal-controlled intersections, 66 links and 6 origin/destination nodes. The lengths and speeds of all links are 500 meters and 50 km/hr, respectively. The clearance times between all incompatible signal groups are 5 seconds. Each signal group is subject to a minimum green of 5 seconds. The saturation flows for all exclusive turning links are 1,400 veh/hr, and those for other links are 1,600 veh/hr. The OD matrix is given as

$$\begin{bmatrix} 0 & 91 & 125 & 223 & 9 & 8 \\ 145 & 0 & 234 & 37 & 277 & 281 \\ 273 & 178 & 0 & 29 & 124 & 132 \\ 279 & 189 & 87 & 0 & 148 & 162 \\ 163 & 78 & 131 & 247 & 0 & 69 \\ 225 & 145 & 52 & 173 & 132 & 0 \end{bmatrix} \text{veh/hr}$$

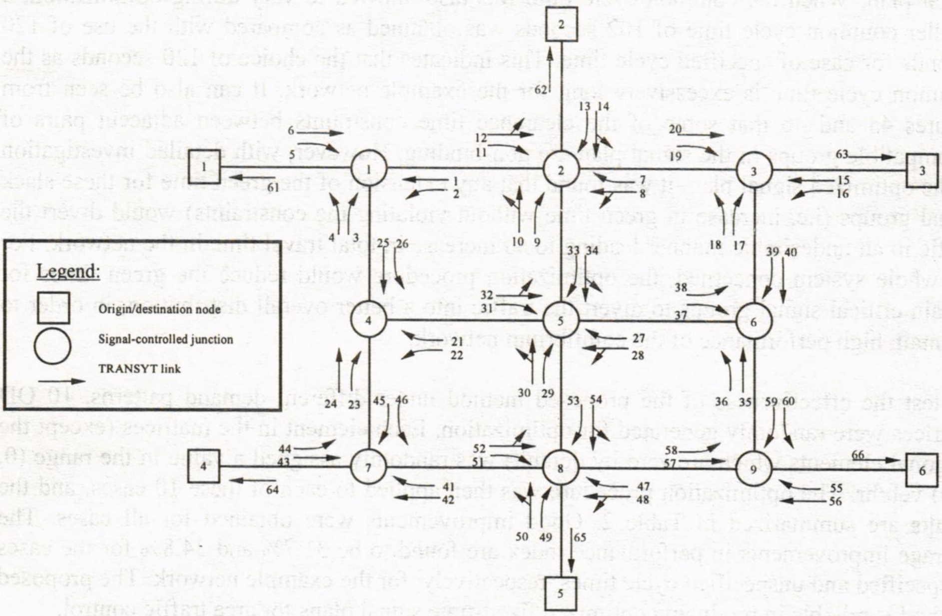

Figure 1. The example network.

The derivatives of the performance index with respect to group-based control variables were determined by the sensitivity analysis in Section 5. To demonstrate the effectiveness of the sensitivity analysis in estimating the derivatives, the results are compared with those calculated by numerical differentiation in which a signal parameter was perturbed by a small amount and the path-based assignment algorithm in Section 4 was used to evaluate the new performance index at equilibrium of traffic. The numerical derivative was determined as the difference of this value from the original performance index, divided by the perturbed amount of the signal parameter. The procedure was repeated for all control variables in the network. The comparison of results is shown in Figure 2 in which good agreement of results from sensitivity analysis and numerical differentiation is found.

The typical convergence characteristics of the method is presented in Figure 3 for the cases of specified and unspecified cycle times. For the case of specified cycle time, a common cycle time of 120 seconds was used for the analysis. For unspecified cycle time, the common cycle time was subject to a maximum of 120 seconds. After optimization, an optimal common cycle time of 102 seconds was determined for the example network. The improvements in performance index are found to be 44.3% and 45.4% for the cases of specified and unspecified cycle times, respectively, for the example network. The results from the case of unspecified cycle time are slightly better than those for specified cycle time. This is because the cycle time variable can also be optimized for the case of unspecified cycle time, while the case of specified cycle time does not have this flexibility. It can also be seen that most improvements in performance index are come from the two integer programming steps N0 or N2.

The optimal signal plans for both cases of specified and unspecified cycle times are compared with the original plan in Figure 4. The central node 5 is used for the demonstration. It can be seen that the advantages of group-based flexibility in automatically generating new phases (stages) and deleting non-effective ones have been fully explored to derive an optimized signal plan. When the common cycle time was also allowed to vary during optimization, a smaller common cycle time of 102 seconds was obtained as compared with the use of 120 seconds for case of specified cycle time. This indicates that the choice of 120 seconds as the common cycle time is excessively long for the example network. It can also be seen from Figures 4a and 4b that some of the clearance time constraints between adjacent pairs of incompatible groups in the signal plan are non-binding. However, with detailed investigation of the optimized signal plan, it was found that any extension of the green time for these slack signal groups (i.e. increase in green time without violating the constraints) would divert the traffic in an undesirable manner leading to an increase in total travel time in the network. For the whole system concerned, the optimization procedure would reduce the green times for certain critical signal groups to divert the traffic into a better overall distribution, in order to maintain high performance of the equilibrium network.

To test the effectiveness of the proposed method under different demand patterns, 10 OD matrices were randomly generated for optimization. Each element in the matrices (except the diagonal elements which are zero by default) was randomly assigned a value in the range (0, 300) veh/hr. The optimization procedure was then applied to each of these 10 cases, and the results are summarized in Table 2. Good improvements were obtained for all cases. The average improvements in performance index are found to be 31.7% and 34.8% for the cases of specified and unspecified cycle times, respectively, for the example network. The proposed method is reliable in producing optimized fixed-time signal plans for area traffic control.

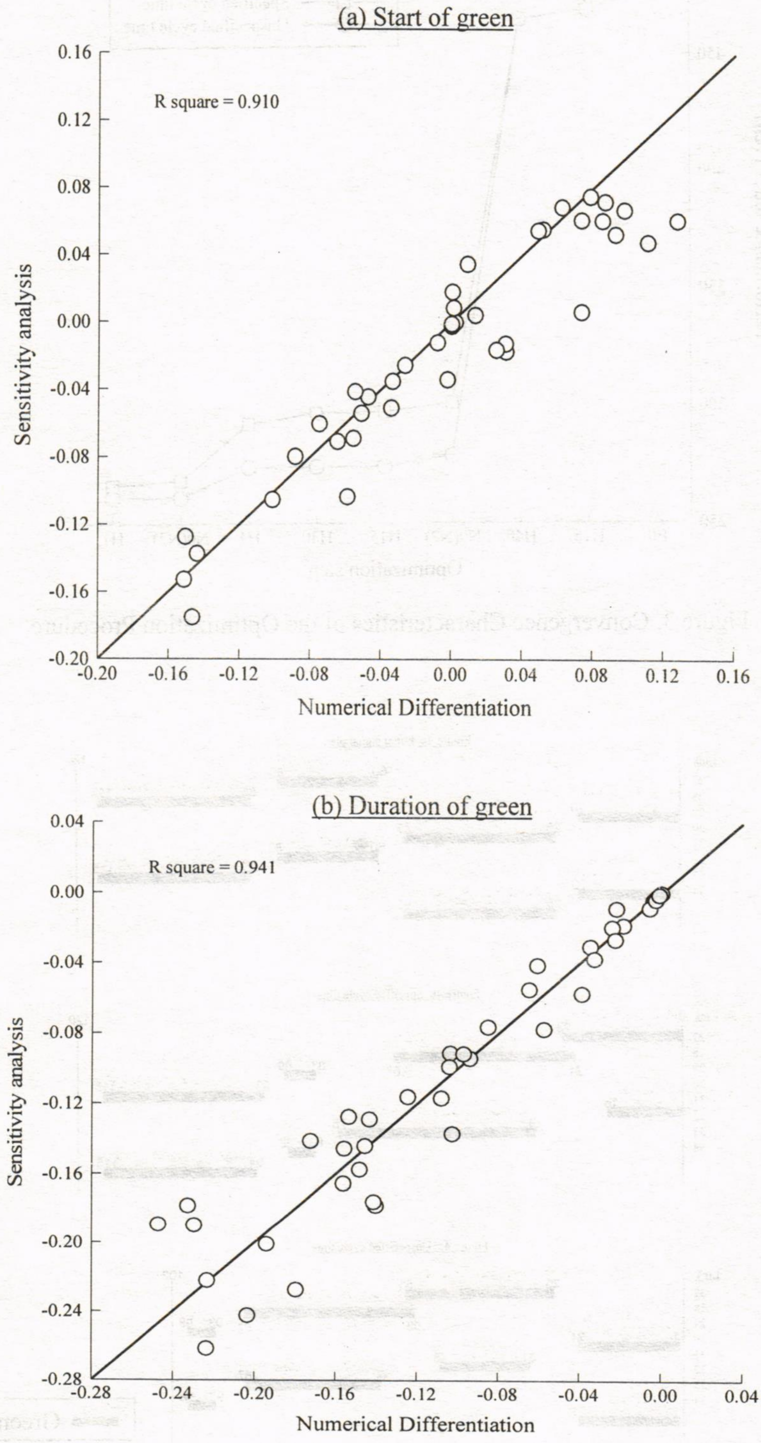

Figure 2. Comparison of the Derivatives of Performance Index (veh/sec)

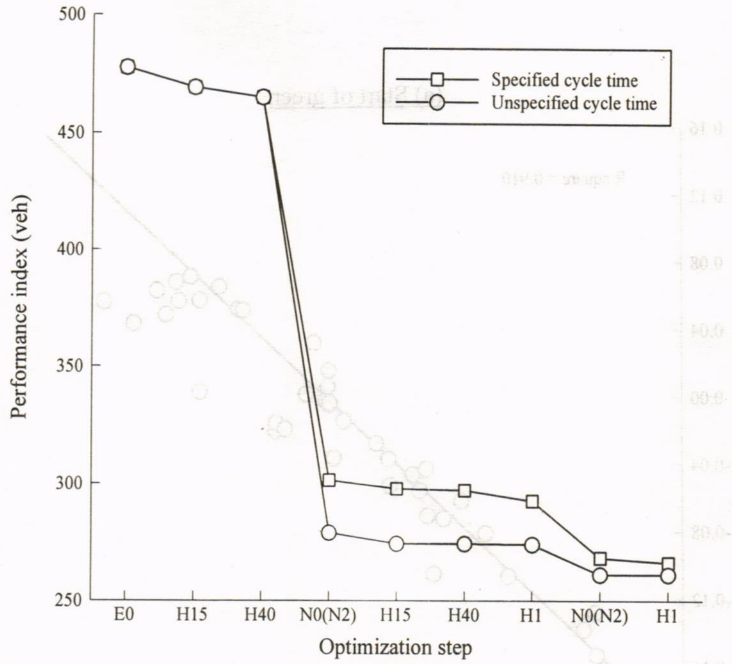

Figure 3. Convergence Characteristics of the Optimization Procedure

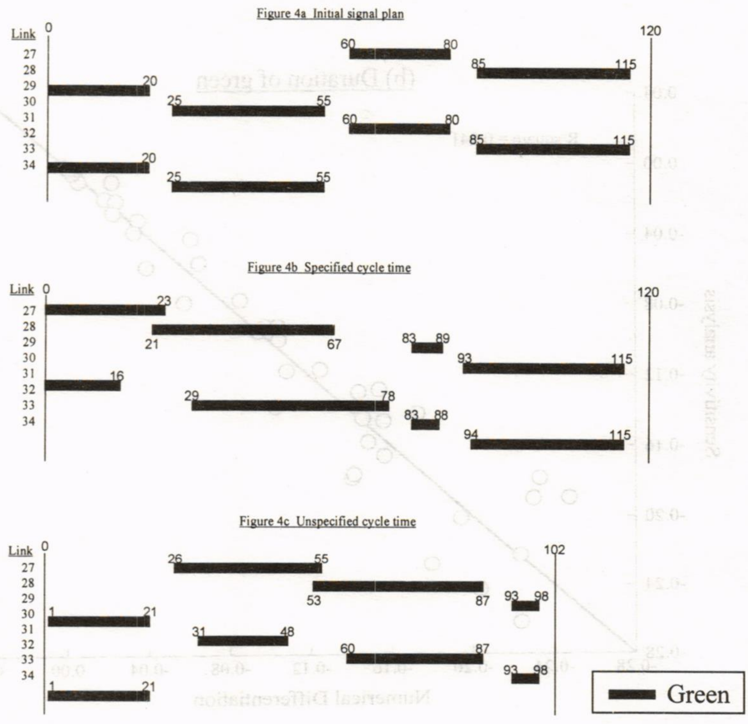

Figure 4. Signal Plans for Junction 5 in the Example Network

Table 2. Optimization Results for Randomly Generated OD Matrices

Randomly generated OD matrix	Performance index (PI)					
	Initial PI (veh)	Specified cycle time (cycle time = 120 sec)		Unspecified cycle time		
		Optimized PI (veh)	Improvement (%)	Cycle time (sec)	Optimized PI (veh)	Improvement (%)
1	478	266	44.3%	102	261	45.4%
2	296	192	35.2%	98	181	38.8%
3	397	270	32.0%	96	247	37.6%
4	275	201	26.9%	106	195	29.0%
5	319	215	32.7%	100	191	40.1%
6	347	287	17.4%	108	274	21.1%
7	611	361	40.9%	120	362	40.7%
8	363	267	26.3%	106	247	31.8%
9	369	242	34.5%	110	242	34.5%
10	300	218	27.2%	110	213	28.9%
Average	375	252	31.7%	-	241	34.8%

8. CONCLUSIONS

In this paper, the optimization of signal settings for equilibrium network based on the TRANSYT traffic model has been studied. The TRANSYT traffic model was used to evaluate the performance index (the total travel time) in a signal-controlled network, taking into account platoon dispersion and signal coordination. The group-based notation was used to specify the signal control variables and constraints, which maximizes the flexibility of signal timing specification for the problem. A sensitivity analysis has been carried out to determine the derivatives of performance index with respect to the signal parameters. These derivatives were then used in an integer programming method to derive the discrete search direction for optimizing the equilibrium signal settings. An optimization procedure, combining the integer programming method and hill-climbing method, has been developed to solve the problem. Numerical examples were used to demonstrate the effectiveness of the proposed method. Encouraging results were obtained.

ACKNOWLEDGEMENTS

This research was supported by a research grant HKU 7015/97E from the Hong Kong Research Grant Council. Transport Research Laboratory is acknowledged for supplying the source code of the TRANSYT program for this research.

REFERENCES

Allsop, R.E. (1968a) Selection of offsets to minimize delay to traffic in a network controlled by fixed-time signals, **Transportation Science, Vol. 2**, 1-13.

Allsop, R.E. (1968b) Choice of offsets in linking traffic signals, **Traffic Engineering and Control, Vol. 10**, 73-75.

Allsop, R.E. (1972) Estimating the traffic capacity of a signalized road junction, **Transportation Research, Vol. 6**, 245-255.

Allsop, R.E. (1974) Some possibilities for using traffic control to influence trip distribution and route choice. **Proceedings of the Sixth International Symposium on Transportation and Traffic Theory**, Elsevier, Amsterdam, 345-374.

Branston, D. (1978) A comparison of observed and estimated queue lengths at oversaturated traffic signals, **Traffic Engineering and Control, Vol. 19**, 322-327.

Gartner, N. (1974) Area traffic control and network equilibrium. In M. Florian, (eds.), **Traffic Equilibrium Methods**, 274-297. Springer-Verlag, Berlin.

Gallivan, S. and Heydecker, B.G. (1988) Optimising the control performance of traffic signals at a single junction, **Transportation Research, Vol. 22B**, 357-370.

Heydecker, B.G. (1992) Sequencing of traffic signals. In J.D. Griffiths (eds.) **Mathematics in Transport and Planning and Control**, 57-67, Clarendon Press, Oxford.

Heydecker, B.G. and Dudgeon, I.W. (1987) Calculation of signal settings to minimise delay at a junction. **Proceedings of 10th International Symposium on Transportation and Traffic Theory**, MIT, July, 159-178, Elsevier, New York.

Hillier, J.A. (1965) Glasgow's experiment in area traffic control. **Traffic Engineering and Control, Vol. 7**, 502-509.

Hillier, J.A. (1966) Appendix to Glasgow's experiment in area traffic control. **Traffic Engineering and Control, Vol. 7**, 569-571.

Holroyd, J. and Hillier, J.A. (1971) The Glasgow experiment: PLIDENT and after. RRL Report, LR 384, Department of the Environment, Road Research Laboratory, Crowthorne.

Kendall, D.G. (1951) Some problems in the theory of queues. **Journal of Royal Statistical Society B, Vol. 13**, 151-185.

Kimber, R.M. and Hollis, E.M. (1979) Traffic queues and delays at road junctions. TRRL Report, LR909, Transport and Road Research Laboratory, Crowthorne.

Robertson, D.I. (1969) TRANSYT: a traffic network study tool. Transport and Road Research Laboratory Report, LR253, Crowthorne, Berkshire, U.K.

Silcock, J.P. (1997) Designing signal-controlled junctions for group-based operation, **Transportation Research, Vol. 31A**, 157-173.

Tobin, R.L. and Friesz, T.L. (1988) Sensitivity analysis for equilibrium network flows, **Transportation Science, Vol. 22**, 242-250.

Traffic Research Corporation (1966) SIGOP: traffic signal optimisation program. Report, PB 173738, New York.

Vincent, R.A., Mitchell, A.I. and Robertson, D.I. (1980) User guide to TRANSYT version 8. Transport and Road Research Laboratory Report, LR888, Crowthorne, Berkshire, U.K.

Wardrop, J.G. (1952) Some theoretical aspects of road traffic research, **Proceedings of the Institute of Civil Engineers, Part II**, 325-378.

Wong, S.C. (1995) Derivatives of performance index for the traffic model from TRANSYT, **Transportation Research, Vol. 29B**, 303-327.

Wong, S.C. (1996) Group-based optimisation of signal timings using the TRANSYT traffic model, **Transportation Research, Vol. 30B**, 217-244.

Wong, S.C. (1997) Group-based optimisation of signal timings using parallel computing, **Transportation Research, Vol. 5C**, 123-139.

Wong, S.C., Wong, W.T., Leung, C.M. and Tong, C.O. (2000) Group-based optimization of a time-dependent TRANSYT traffic model for area traffic control, **Transportation Research B**, in press.

Wong, S.C. and Yang, C. (1999) An iterative group-based signal optimization scheme for traffic equilibrium networks, **Journal of Advanced Transportation, Vol. 33, No. 2**, 201-217.

Wong, S.C., Yang, C and Lo, H.K. (2001) A path-based traffic assignment algorithm using the TRANSYT traffic model, **Transportation Research, Vol. 35B**, 163-181.

Wong, S.C. and Yang, H. (1997) Reserve capacity for signal-controlled network, **Transportation Research, Vol. 31B**, 397-402.

Yang, H. and Yagar, S. (1995) Traffic assignment and signal control in saturated road networks, **Transportation Research, Vol. 29A**, 125-139.