
Monograph:
van Vuren, T. (1990) The Interaction Between Signal Control and Traffic Assignment: An Introduction. Working Paper. Institute of Transport, University of Leeds , Leeds, UK.

Working Paper 313

Working Paper 313

September 1990

THE INTERACTION BETWEEN SIGNAL CONTROL AND TRAFFIC ASSIGNMENT: AN INTRODUCTION

Tom van Vuren

FUNDAMENTAL REQUIREMENTS OF FULL-SCALE DYNAMIC ROUTE GUIDANCE SYSTEMS

Route Guidance and Signal Control: An Introduction

Working Paper 5

September 1990

Tom van Vuren

This work was sponsored by the Science and Engineering Research Council under a Rolling Programme.

Institute for Transport Studies, University of Leeds, Leeds, LS2 9JT. England. Tel: Leeds 335325

Transportation Research Group, University of Southampton, Southampton, SO9 5NH, England. Tel: Southampton 592192.

Abstract

Traffic assignment and signal control are generally considered separately. Proper recognition of the interaction that exist between assignment and signal control introduces a number of theoretical and computational complexities; on the other hand considerable network improvements can be achieved.

This report introduces the problem of interaction between signal control and traffic assignment, reviews related issues and previous work. It discusses signal control as a network design problem, and it makes a case for the iterative assignment control procedure for its approximate solution.

CONTENTS

1. An introduction to traffic signal control

1.1 Introduction

In our current congested cities, the number of traffic signals is ever increasing in an attempt to curb the huge traffic flows. The parallel developments in computer technology enable more and more advanced ways of controlling the signals, not just in isolation, but network-wide in e.g. UTC (Urban Traffic Control) systems.

Views on optimum signal design differ widely between countries and continents, related to, e.g.:

(a) philosophies of design
(b) legal requirements
(c) operational requirements
(d) engineering practice, and
(e) general structure of urban areas.

An interesting comparison between signal design in the UK and West-Germany is given by Heydecker and Silcock (1989) and Fellendorf (1989). The grid structure of many American cities also puts different requirements and constraints on signal design, which makes methods developed in the United States less appropriate for the UK. The number of publications worldwide in this area is enormous; it is therefore out of the scope of this research to present a complete overview of developments in signal control; I will concentrate on major developments in the UK. Further, I will limit myself to stage-based control. The 4 approaches to signal control discussed will be:

Webster's method;
delay minimisation;
TRANSYT; and
SCOOT.

This discussion follows a chronological order over the last three decades. Also, it shows the increased complexity and dependence on advanced computer technology in signal control. Throughout this note the following notation will be used:

a_{ij} = 1, if movement i runs during stage j
0, otherwise (stage matrix)
C = cycle time
d_i = delay at traffic signals for movement i
f_i = flow on movement i
g_j = green time for stage j
s_i = saturation flow for movement i
λ_i = green time proportion for movement i = $\Sigma_j\ a_{ij}g_j/C$
x_i = $f_i/\lambda_i s_i$ = degree of saturation for movement i

1.2 Signal control strategies in the UK

In this note I will concentrate on strategies to control green splits. Heydecker et al. (1990) provide an introduction to optimum signal offsets in an assignment context.

The original paper on traffic signal control in the UK must be that by Webster (1958). The paper deals solely with isolated signalised intersections (a fair assumption in those days!). Webster's calculations were based on mathematical queuing models, combined with a sense of practical applicability.
Webster's algorithm for optimum green time splits is based on the premise that overall delays at a junction are minimal when the maximum rates of approach saturation per stage are equal, and is therefore also known as equisaturation. If we take into account possible green time constraints this objective can be expressed as follows:

$$\text{Min}_j\ \text{Max}_i \quad a_{ij}x_i \tag{2.1}$$

The efficacy of this policy lies in the fact that delays increase more than linearly with increasing saturation rates. Problems with Webster's method arise when stage matrices become complicated (so that movements may run in more than one stage) and when the total saturation rate of the junction increases close to one, or even exceeds one.

Whereas Webster's method is more a rule-of-thumb which determines approximately optimum signal green splits, Allsop (1971) suggested a method which calculates green times so that total delays at the junction are explicitly minimised. In formula:

$$\text{Min} \quad \sum_i \quad f_i \, d_i \tag{2.2}$$

for which signal settings can be calculated via

$$Eq_j \quad \sum_i \quad f_i \frac{\partial d_i}{\partial \lambda_i} \tag{2.3}$$

(subject to green time constraints). Eq_j stands here for "equate for all stages j". Only development of computers enabled calculation of more complicated formulae, such as these. For a given flow pattern, and with Webster's expression for delays (see Chapter 3) this problem is one of convex minimisation and the resulting optimum green times are unique. Allsop (1971) presents a comparison between Webster's method and delay minimisation, which shows the latter's supremacy.

Around the same time TRANSYT was introduced by TRRL (Robertson, 1969). The aim of TRANSYT is to minimise a weighted sum of stops and delays in a network with fixed-time UTC system by changing signal cycle times, offsets and green splits. Because of the complexity of this problem TRANSYT operates in an iterative fashion between a traffic simulation model (to estimate delays and stops) and a heuristic optimisation stage. The resulting green splits are therefore not unique, and may be only sub-optimal. The advantage over the two previous methods is that interactions between junctions can be taken into account during the optimisation process, but the computational effort needed is heavily increased.

The continual updating of fixed-time signal plans as developed in TRANSYT is labour-intensive and thus expensive. In addition, fixed-time plans cannot anticipate or adjust to incidents, so that network conditions must be constantly monitored from a central control room. This illustrates the advantages of a responsive traffic control system, or vehicle-actuated system, which analyses data obtained from vehicle detectors to calculate optimum signal settings for current conditions, or expected conditions in the near-future. The best-known responsive method in the UK is SCOOT (Split, Cycle and Offset Optimisation Technique), developed by TRRL during the second half of the 1970's (Hunt et al., 1981). Based on detector information small adjustments to the current signal settings are made. It is interesting to note that SCOOT calculates green splits that balance the degrees of saturation per approach (like Webster's original method), treating junctions independently, although

weighting factors can be applied.

The short overview has shown the increased sophistication in traffic signal control over time. However, each of these methods either assumes a fixed flow pattern arriving at the junctions, or in the case of SCOOT, at least accepts the arriving flows, and optimises the junction performance with respect to those flows, as will most other vehicle-actuated control strategies do. In Chapter 4 it will be shown that the assumption of fixed flows is unrealistic and worse, that optimisation for the current flow pattern will not necessarily lead to decreased delays in the network if drivers' route choice is taken into account.

2 An introduction to traffic assignment

Traffic assignment is the process of finding routes through a road network and loading demand trips onto them. To achieve this, the total number of vehicles that wish to travel between each origin i and destination j in the area must be given in the T_{ij} element of the OD matrix T. Further we must have a representation of the available road network in the form of nodes (representing junctions) and a set of links (representing roads) connecting them. Generally travel impedance or cost functions are related to the links, which could reflect travel time, distance, cost of travel, etc.
The task in the assignment stage is then to determine routes (or paths) through this network, which consist of strings of consecutive links, connecting origin-destination pairs.

To solve the traffic assignment problem, the rule by which drivers choose routes between their origin and destination of travel must be defined. It is not unreasonable to assume that every driver wishes to minimise his personal travel cost in doing so; it is generally accepted that time minimisation is a good explanation for drivers' route choice (Bovy, 1981; Wootton et al., 1981), and unless otherwise stated I will follow this assumption throughout this research. Hence I will consider "time" and "cost" to be interchangeable. Shortest path finding is an essential element in the assignment stage in which trees of consecutive links are constructed. Path costs are determined by a summation over all constituting link costs in other words, they are linear additive. See Van Vuren and Jansen (1988) for an introduction to shortest path algorithms.

In the situation that each driver is on a minimum cost route, we have reached stability, or a User Equilibrium (UE). This condition was first stated by Wardrop (1952) as follows:

> "(For each OD-pair) the journey times on all the routes actually used are equal, and less than those which would be experienced by a single vehicle on any unused route".

which can also be stated as:

> "no driver can improve his travel costs by unilaterally changing routes"

Wardrop's user equilibrium rests heavily on 2 assumptions:

(a) drivers have complete network knowledge;
(b) all drivers have identical behaviour.

Both assumptions can (partly) be relaxed via 2 generalisations:

(a) by distinguishing between the actual and **perceived** travel costs of a link. If we take the perceived travel cost to be a random variable with mean actual travel cost, distributed among the population of drivers, we can define a Stochastic User Equilibrium (SUE) in which "no driver can improve his **perceived** travel cost by unilaterally changing route". This will be discussed further in Section 2.3.
(b) by establishing a number of groups of drivers, each with their own definition of travel cost. For each of these **user classes** an assignment can be carried out in interaction with the other classes, but with class-dependent cost definition, so that in a Multiple User Class (MUC) equilibrium Wardrop's equilibrium conditions also hold.

2.1 Mathematical formulation of the user equilibrium

Let t_a be the cost function for link a, so that $t_a = t_a(f_a)$; f_a being the flow on link a (the assumption that the cost of using link a only depends on the flow on link a itself, i.e. the case of separable cost functions, is not necessary, but used here for simplicity). Let q_{ijk} be the flow on the k-th path between OD-pair i-j and v_{ijk} be the associated path costs. Clearly

$$c_{ijk} = \Sigma_a \, t_a \, \delta_{aijk} \qquad (2.4)$$

where $\delta_{aijk} = 1$ if link a is part of route k connecting OD-pair ij and $\delta_{aijk} = 0$ otherwise (the link-path incidence matrix). Along the same lines

$$f_a = \Sigma_{ijk}\, q_{ijk}\, \delta_{aijk} \tag{2.5}$$

so that the link flow constitutes a sum of all path flows using that link.
If we now call u_{ij} the minimum path cost for OD-pair i-j and T_{ij} the total demand flow on relation i-j we can state Wardrop's equilibrium conditions as follows:

$$q_{ijk}\,(v_{ijk} - u_{ij}) = 0 \qquad \forall ijk \tag{2.6}$$

$$v_{ijk} - u_{ij} \geq 0 \qquad \forall ijk \tag{2.7}$$

$$\Sigma_k\; q_{ijk} = T_{ij} \qquad \forall ij \tag{2.8}$$

$$q_{ijk} \geq 0 \qquad \forall ijk \tag{2.9}$$

Although these conditions are sufficient for a user equilibrium, they do not actually solve the traffic assignment problem. Beckman et al. (1956) showed that this can be achieved by solving the following minimisation:

$$\text{Min } Z(\mathbf{f}) = \Sigma_a \int_0^{f_a} t_a(x)\, dx \tag{2.10}$$

subject to conditions (2.8) and (2.9) above. In fact, conditions (2.6) to (2.9) are known as the Kuhn-Tucker conditions, necessary for a constrained minimum.

This mathematical program is in general non-linear with linear constraints. For the equilibrium link flow and cost pattern to be unique a sufficient condition is that the objective function (2.10) be convex, so that

$$\frac{\partial^2 Z}{\partial f^2} > 0 \tag{2.11}$$

or in other words (with separable cost functions):

$$\frac{\partial t_a(f_a)}{\partial f_a} > 0 \tag{2.12}$$

Uniqueness of the equilibrium **link flows** and costs imposes constraints on the link cost functions as shown in (2.11) and (2.12), which means for the case of separable cost functions continuously differentiable and increasing. This does not, however, imply that the equilibrium is unique with respect to **path flows**. Remember that Wardrop's user equilibrium is first and foremost concerned with travel costs!

2.2 Solution algorithms for user equilibrium

The advantage of stating the UE assignment problem as a constrained convex minimisation is that a number of efficient solution algorithms are available to solve it. The most appropriate one must be the **convex combinations method**, first described by Frank and Wolfe (1956), and therefore better known in transport circles as the Frank-Wolfe algorithm.

The method consists of 2 basic steps:

(a) determination of the maximum "drop"; where the drop is the product of the rate of descent in a feasible direction at the current solution and the length of the feasible region in that direction

(b) determination of an optimum step size, which minimises Z in that direction.

The beauty of the Frank-Wolfe algorithm, and the meaning of descent directions in traffic assignment is best shown by application in a network environment.

If **h** is a vector of feasible flows (...., h_a,...) at iteration n the optimum move direction is found by:

$$\text{Min } Z^n(\mathbf{h}) = \Sigma_a \quad \frac{\partial Z(\mathbf{f}^n)}{\partial f_a} . h_a \tag{2.13}$$

which is the product of the rate of descent and the maximum allowable step size in that direction. However,

$$\frac{\partial Z(\mathbf{f}^n)}{\partial f_a} = t_a^n \tag{2.14}$$

so that (2.13) can be written as

$$\text{Min } \Sigma_a \, t_a^n \, h_a \tag{2.15}$$

(subject to non-negativity and conservation of flow constraints).

The solution to this, minimisation of the total network cost with fixed travel times $t_a^n = t_a (f_a^n)$ is a straightforward shortest path loading.

The optimal move size λ can then be determined as the solution in λ to:

$$\frac{\partial}{\partial \lambda} Z (\mathbf{f}^n + \lambda (\mathbf{h}^n - \mathbf{f}^n)$$

$$= \Sigma_a (h_a^n - f_a^n) \, . \, t_a (f_a^n + \lambda (h_a^n - f_a^n)) = 0 \tag{2.16}$$

(unless a feasible boundary is met).

So, in each iteration n of the Frank-Wolfe algorithm a set of shortest paths, based on the current flow pattern $\mathbf{f}$ is determined, and translated into link auxiliary flows $\mathbf{h}^n$; this is followed by an optimum step size determination in that direction. The improved solution is then determined via

$$\mathbf{f}^{n+1} = (1 - \lambda_n) \, \mathbf{f}^n + \lambda_n \, \mathbf{h}^n \tag{2.17}$$

If condition (2.11) for the link cost functions is satisfied, this algorithm is guaranteed to find the minimum value for objective function Z, where the UE conditions (2.6) to (2.9) hold.

A variation on the Frank-Wolfe algorithm is the Method of Successive Averages (MSA). In this method the step size λ is not determined so as to minimise the objective function along the descent direction, but determined a priori. As long as the search direction is a descent direction and the sequence of step lengths satisfies the following conditions

$$\sum_{n=1}^{\infty} \lambda_n = \infty \tag{2.18}$$

$$\sum_{n=1}^{\infty} \lambda^2_n < \infty \tag{2.19}$$

the iterative procedure will converge to the minimum of the objective function,and thus to a UE (Powell and Sheffi, 1982). A general sequence which is used for step size determination, and satisfies (2.18) and (2.19) is

$$\lambda_n = 1/n \tag{2.20}$$

Thus, the MSA procedure does not require a linear search for optimum step length determination; on the other hand, it loses in overall efficiency, exactly because it does not move in optimal steps. Therefore, the MSA has only limited value in UE traffic assignment; its greatest importance lies in situations when an objective function cannot be evaluated analytically, such as in a stochastic user equilibrium assignment.

A third solution algorithm to the user equilibrium was introduced by Dafermos and Sparrow (1969). This method determines a descent direction by finding for each OD-pair the two most unbalanced paths (i.e. with largest cost difference). The step size chosen in this descent direction is thus, that the path costs in the end should be equal. By iterative application this method is guaranteed to converge to a UE. Its main disadvantage is the path enumeration it requires, which makes its use prohibitive for larger sized networks.

3 The role of cost functions

The importance of cost functions in the process of traffic assignment and conditions for existence of a unique equilibrium were already discussed in Chapter 2. In the case of separable costs the relation between flows on a link and associated costs of using that link must satisfy (2.12); which basically means that the cost of travel along a link is continuously differentiable and increasing with respect to flow.

Cost functions play an important role in signal optimisation, too, although the dependent variable in that case will be the green split over the various stages. In signal optimisation the fixed cost of travel along a link are of no importance and we are mainly concerned with flow-dependent delays as a function of green times. In the case of delay minimisation the derivative of the cost function with respect to the green splits must be calculated; in TRANSYT after each iteration in the green time optimisation heuristic the resulting delays must be evaluated in the traffic model. The cost function employed is evidently of great influence on both resulting greem

times and delays, plus on computational effort required.

A large number of cost functions has been proposed worldwide over the years, and it would take too much time and space to try and describe these exhaustively. Important reference works are by Allsop (1972), Hutchinson (1972) and Branston (1976).

3.1 Four cost functions

A cost function which is used very often, particularly in the United States, is the one proposed by the US Bureau of Public Roads (1964), and generally known as the BPR-function:

$$t_a = t_a^0 \, (1 + \alpha(f / c)^{\beta}) \qquad (2.31)$$

with c being capacity and t_a^0 the free flow link travel time.
Originally values of 0.15 for α and 4 for ß were suggested, without clear support. Note that c in this case stands for "practical capacity" (at which travel times along the link are 15% higher than free flow travel times), but the BPR function can equally be used with c denoting the real link capacity. For signal-controlled links this would mean that c can be substituted by λs. The functional form of (2.31) has no theoretical backing, but it is extremely convenient in traffic assignment because of its ease of integration and differentiation.
Note that cost function (2.31) can be expressed as a combination of (fixed) free flow travel time and flow-dependent junction delay: $t_a = t_a^0 + d_a$.

If we assume Poisson arrivals at the junction and a constant departure rate we can express the average junction delay to arriving vehicles via the Pollaczek-Khintchine equation (omitting subscripts a from now on):

$$d = \frac{x^2}{2f\,(1-x)} \qquad (2.32)$$

where $x = f / c$ (saturation rate).

Webster (1958) used this expression for estimating delays at traffic signals, replacing c by λs. The complete cost function at traffic signals contains another term (usually called Webster's first term), which expresses the delay for uniform arrivals at the stop line with red and green signal positions. The complete expression for Webster's

cost function for signalised intersections is thus:

$$d = \frac{C(1-\lambda)^2}{2(1-f/s)} + \frac{x^2}{2f(1-x)} \qquad (2.33)$$

Actually, Webster added a third term for an optimum fit to observations, but this term lacks theoretical backing and is usually omitted. Similar expressions were derived by Miller (1963) and Newell (1965); neither, though, has the same simplicity as Webster's function.

Davidson (1966) also applied the P-K formula for delay calculation, but (like the BPR-formula) he transformed it, so as to express the ratio of delay to the normal service time (i.e. the free flow travel time t_a^0). The amended cost function then takes the form:

$$t_a = t_a^0 \left(1 + \frac{x}{1-x}\right) \qquad (2.34)$$

which again consists of a free flow component and a delay component. In addition Davidson introduced a delay parameter I, also called level of service, which depends on e.g. the quality of the road and the amount of interference by parking etc. I takes values between 0 and 1, with the lower values of I expressing higher levels of service; appropriate values for I should be based on observations in situ. The resulting cost function looks like this:

$$t_a = t_a^0 \left(1 + I\ \frac{x}{1-x}\right) \qquad (2.35)$$

and a graphical representation (with varying values for I) is given in Figure 1.

The Pollaczek-Khintchine equation, plus cost functions based on it, assume a steady state in the network. In such conditions an unrealistic property of these curves that infinite queues and delays are calculated at and above capacity (so that the curve is asymptotic the vertical at f/c = 1). In real-life, however, such enormous queues do not develop, even though junctions may well be **temporarily** overloaded.

If we only consider a limited overloaded time period T, the value of the average queuing time will be T/2 (f/c - 1). This delay curve we call the deterministic queuing curve, which is only defined above capacity.

Both curves are shown in Figure 2. Each of the curves can be expected to be a good predictor under different circumstances: the P-K formula if the approach is undersaturated ($x \leq 0.8$), and the deterministic queuing curve if the time T of oversaturation is only of moderate length and the saturation ratio appreciably larger than 1 ($x \geq 1.1$). Both curves are also completely incompatible around the point $x = 1$ where they ideally would meet.

An ingenious new cost function was developed by P D Whiting at TRRL, combining properties of both curves. The base of this curve is formed by a coordinate transformation, which tilts the asymptote of the P-K curve at capacity to become equal to the deterministic queuing curve above capacity. The form of this **sheared delay curve** is as follows:

$$d = 1/2 \left(\sqrt{A^2+B} - A \right) \qquad (2.36)$$

where $A = \frac{c-f}{2c}$

$B = \frac{fT}{c^2}$

Figure 3 shows this sheared delay curve, compared with the P-K curve and the deterministic queuing curve. It can be seen that this cost function approximates both other functions very closely in the regions mentioned before. In the region inbetween ($0.8 < x < 1.1$) the function behaves sensibly; it must be remembered though, that this curve is not based on any realistic queuing analysis.

In the same way as before, λs can be substituted for c at traffic signals, whilst again Webster's first term should be added to take account of uniform delays. The resulting form of the sheared delay function then becomes:

$$d = \frac{C(1-\lambda)^2}{2(1-f/s)} + \frac{T}{4\lambda} \left\{ \left(f/s - \lambda \right) + \sqrt{(f/s - \lambda)^2 + 4f/s^2T} \right\} \qquad (2.37)$$

Actual behaviour on the road does not seem to warrant any of the cost functions discussed. They have been derived, either through theoretical queuing analysis, or because of their convenient functional form. Their behaviour in transport modelling, however, will be completely different, certainly in congested situations when the degree of saturation ratio of certain approaches may well exceed 100%. Both traffic assignment results and signal optimisation results depend heavily on the delay

assumptions made. Any results of such exercises should be viewed in the light of the cost functions used, and assessed on the reality of the assumed behaviour.

4 The interaction between signal control and assignment

4.1 Introduction

As was stated in Chapter 1. signal control strategies generally optimise signals for the current flow pattern, which is assumed not to change as a result of the altered green times. Along the same lines, green times are usually assumed fixed in traffic assignment: we apply separable cost functions $t_a(f_a)$ with the influence of the green splits λ_a incorporated in the parameters of the cost function (generally in the link capacity).

Allsop (1974) was the first to establish that the two processes are not independent. He expressed the link cost functions with explicit reference to the green times: $t_a(f_a, \lambda)$, and he suggested, through this dependence, that the traffic signals can, in principle, be used to influence the traffic flow pattern, and to improve network conditions. A similar argument was set up by Dickson (1981).

Because of the complicated dependence of the equilibrium flows $\mathbf{f}^*$ on green times λ, Allsop suggested solving the problem of optimum signal control in interaction with route choice using an iterative procedure, in which the green split optimisation and traffic assignment problems are solved alternately. The advantage of this procedure, which we shall call the iterative assignment control procedure, is that in this way the problem is decomposed into two well-researched and well-behaved sub-problems.

In the following years a number of computer tests were carried out by Allsop and Charlesworth (1977), in which the iterative assignment control procedure was applied using TRANSYT on a small six-junction network. These tests showed that for the combined assignment / control problem several mutually consistent points may exist, where the flows are in equilibrium and signals optimal for those flows, with quite different flow and green time patterns, but virtually equal total travel times in the network.

Dickson (1981) expressed these mutually consistent points as follows:

$$\text{Min}_{\lambda} \Sigma f_a^* t_a(f_a^*, \lambda) \quad (2.38)$$

where f_a^* is an equilibrium flow pattern. He also showed via a simple example how the iterative assignment control procedure does not necessarily find green time / equilibrium flow combinations that minimise total travel time in the network:

$$\text{Min}_{\lambda,f} \Sigma f_a^* t_a (f_a^*, \lambda) \quad (2.39)$$

He remarked that for network optimal signal settings λ and resulting flows $\mathbf{f}^*$ signal settings should be adjusted in a way which simultaneously takes into account the resulting changes in the equilibrium flows. No indication of how this could be done was given, however.

Finally, Smith (1979a) showed with a simple two-link example how Webster's signal control policy in interaction with route choice does not solve the optimum control problem. Applying Webster's two-term delay formula he also showed how an iterative application of Webster's policy and traffic re-assignment can actually push the flow pattern to an infeasible boundary, where some flows exceed capacity. He proposed an alternative control policy, P_0, which should avoid such problems.

4.2 The policy P_0

The local control policy P_0 was introduced for a simple two-stage junction by Smith (1979a) as follows:

$$s_1 d_1 = s_2 d_2 \quad (2.40)$$

This was later (Smith, 1981a) generalised to:

$$\text{Eq}_j \Sigma_i a_{ij} s_i d_i \quad (2.41)$$

Smith (1981b) shows that this policy P_0 possesses capacity-maximising properties, so that it maximises the total demand flow in the OD-matrix that can be catered for by the network; this property was expressed as follows. "For each $k < K$ the network will certainly provide in a stable way for the demand kT". K is the largest number with this property; P_0 maximises this value K for a network, using only local traffic information. Thus, P_0 should guarantee a feasible solution to the

combined assignment / signal control problem under the highest demand conditions that can be met by the supply in a network.

This is achieved by attracting traffic to higher capacity roads, as the control policy favours roads with high saturation flows by allocating proportionally more green time to these. This is in accordance with an observation by Heydecker (1981) that signal control policies should discourage drivers from using roads with high marginal costs; see also Section 2.4 on system optimal assignment. Smith (1981b) calls P_0 a distributed control system, in which a local control policy is designed to take action which adds up to achieve an overall objective, namely to maximise network capacity, which in heavily congested situations may be a good approximation to minimising total network cost at equilibrium.

4.3 A theoretical discussion

If signal settings are allowed to vary with changing flows, the resulting link cost functions are not separable any more, and furthermore, they will depend on the control policy applied. The resulting non-separable cost functions

$$\mathbf{t}(\mathbf{f}, \lambda) \tag{2.42}$$

depend on flows and green times throughout the network. Dafermos (1971) was the first to investigate such cost functions and established that if the Jacobian $\partial \mathbf{t}/\partial \mathbf{f}$ of these cost functions is symmetric, the resulting problem reduces to one of convex minimisation, where the Kuhn-Tucker conditions (2.6) to (2.9) hold. Smith (1979b) and later Dafermos (1980) derived a weaker condition for a unique and existing equilibrium with non-separable cost functions via a variational inequality approach. This condition is that the cost function $\mathbf{t}(\mathbf{f}, \lambda)$ must be monotone, which can be expressed mathematically as:

$$(\mathbf{t}(\mathbf{f}, \lambda) - \mathbf{t}(\mathbf{h}, \lambda)) \cdot (\mathbf{f} - \mathbf{h}) \geq 0 \tag{2.43}$$

Monotonicity of the cost function can be checked via the Jacobian J (the matrix of its first order derivatives), which must be positive semi-definite, or in other words: the determinant of the sum of the Jacobian and its transposed matrix should be greater than or equal to 0.

$$\left\| \frac{J+J^T}{2} \right\| \geq 0 \qquad (2.44)$$

Heydecker (1983) investigated a simple two-way junction with Webster's control policy, delay minimisation and Po, and investigated monotonicity of Webster's two-term delay formula in combination with these policies. To control for monotonicity he employed a weaker condition than (2.44):

> the determinants of all principal submatrices of the Jacobian matrix should be everywhere non-negative (2.45)

Thus, if condition (2.45) fails, so does condition (2.44). Heydecker discovered that neither of the three policies (Webster's, delay minimisation and P_0) in interaction with Webster's delay formula satisfied monotonicity condition (2.45), so that no unique equilibrium point for the combined signal control / assignment problem necessarily exists.

The variational inequality statement for a traffic equilibrium is as follows (Smith, 1979b; Dafermos, 1980):

$$-t(\mathbf{f}^*, \lambda) \cdot (\mathbf{f} - \mathbf{f}^*) \leq 0 \qquad (2.46)$$

or in words (Smith, 1979b)

$$-t(\mathbf{f}^*, \lambda) \text{ is normal, at } \mathbf{f}^*, \text{ to } D \qquad (2.47)$$

where D is the set of demand-feasible link flows. Along the same lines we can express conditions on signal control policies: at equilibrium, where the policy gives optimum settings for the flow pattern:

$$c(\mathbf{f}, \lambda^*) \text{ is normal, at } \lambda^*, \text{ to } E \qquad (2.48)$$

where E is the set of allowable green times. Smith refers to (2.48) as a directional constraint, which control policies must satisfy to ensure existence of an equilibrium solution to the assignment problem in interaction with signal control. In Smith (1981a) and Smith (1981b) he shows that Webster's control policy and delay minimisation do not satisfy directional constraint (2.48) if the P-K formula is employed for delay calculations, and therefore do not guarantee a solution to the

assignment problem. P_0, however, has been designed explicitly with this constraint in mind, and guarantees a solution, though not necessarily unique. Conditions on the delay formula, which render the solution to the assignment problem in interaction with Po unique, are derived in Smith (1985); these conditions are that the delays on a link depend only on the spare capacity (λs - f).

4.4 Conclusion

We have seen that the existence of a solution to the traffic assignment problem in the case of signal controlled junctions depends on the control policy and the cost function employed. Conventional control policies, which optimise the signals for the current flow pattern do not satisfy the necessary conditions for this; a new policy P_0 possesses the necessary qualities to ensure an equilibrium solution under maximum demand, and under certain conditions on the link cost functions this equilibrium is unique.

This is not to say that equilibrium solutions to the combined assignment / control problem never exist with Webster's control policy or with delay minimisation, nor that the resulting equilibria with any policy is optimal, in the sense that total network cost is minimised. This problem of finding signal settings which minimise total travel time in the network, with drivers following a UE routing pattern, is a form of the network design problem (NDP), which will be discussed in the next Section.

5 Signal control as a network design problem

The network design problem (NDP) is one of the most intriguing problems in transport research, but also one of the most complicated. The task is to determine optimum network improvements or additions (i.e. investments), so that some measure of total network cost will be minimised. Usually a budget constraint applies, and drivers are expected to follow a UE flow pattern, so that their route choice will adapt to the changed conditions. Interesting overview papers are written by Abdulaal and LeBlanc (1979), Magnanti and Wong (1984) and Heydecker (1986).

In the case of signal design the investment variables are formed by the green times per stage, whilst total cycle time forms a budget constraint; the resulting NDP is like the one stated by Dickson (1981).

$$\text{Min} \quad \Sigma_{f,\lambda} \, f^{*}. \, t(f^{*} , \lambda) \tag{2.49}$$

$$\text{Sub} \quad -t (f^{*} , \lambda) . (f - f^{*}) \leq 0 \qquad \forall f \in D \tag{2.50}$$

(where D is the set of demand-feasible flows).

As the equilibrium flows do not vary linearly with the control parameters λ, the constraints (2.50) are non-linear, leading to non-convexity in the main problem. Also, the problem is non-differentiable, as the set of equilibrium routes may change with the investment variables λ. Finally, because of the computational complexity of calculating a UE flow pattern (2.50), direct search methods are prohibitively expensive.

Various researchers have proposed different approaches to find solutions to the NDP. I will here give an overview of the most interesting developments, concentrated on the optimum signal design problem.

Tan et al. (1979) use the observation that an equilibrium flow pattern can be expressed as a set of constraints, to solve the main problem (2.49) under an increased number of constraints, some of which are non-linear. They call this a Hybrid Optimization Formulation and suggest an Augmented Lagrangian Method for its solution, but the path enumeration this requires makes this algorithm inappropriate for networks of even reasonable size.

Marcotte (1983) suggests a constraint relaxation technique, but the resulting sub-problem is non-convex. Therefore solution by this algorithm is also extremely time-consuming, particularly where the network size is considerable. Because of the non-convexity the resulting solution is not unique.

Sheffi and Powell (1983) acknowledge the non-convexity of the optimal signal control problem. Their solution algorithm is a feasible descent procedure which, because of this non-convexity, is bound to find only a local optimum. To determine a descent direction, a gradient needs to be determined for every link with respect to all stages in the network. Because of non-differentiability this is done numerically, but this requires in each iteration a number of equilibrium assignments equal to the overall number of stages in the network, which will be prohibitive for real-sized networks.

The solution methods that Abdulaal and LeBlanc (1979) suggest suffer from the same drawbacks.

Heydecker and Khoo (1990) suggest approximating the non-linear constraints (2.50) by linear relationships, which have to be fitted numerically to a small number of UE flow patterns. Although this reduces the computational complexity of the NDP considerably, the resulting problem is still of rather large proportions, certainly for life-size networks. Reported results are encouraging, though.

A bilevel linear programming formulation for the NDP is suggested by LeBlanc, Boyce and others (LeBlanc and Boyce, 1986; Ben-Ayed et al., 1988). As the name indicates, here the NDP is split up in an upper level optimisation and a lower level optimisation; at the upper level the total system costs are minimised by changing green times, whilst at the lower level the user equilibrium is pursued by minimising cumulative user travel costs, as in (2.10). To simplify solution, a piecewise linearisation of the link cost functions is proposed. Unfortunately, no convergent solution algorithm for the bilevel linear program is known.

Frequently the iterative assignment control procedure is suggested to approximately solve the NDP (Tan et al., 1979; Sheffi and Powell, 1983), using delay minimisation in the control stage. The possibility of the non-existence of a feasible solution (as described in Chapter 4) is usually quoted against this approach. Friesz and Harker (1985) explain how in the signal optimisation process this procedure assumes that $\partial f/\partial \lambda = 0$, and only if this is roughly true will the iterative assignment control procedure approximate the solution to the NDP. We can only expect this to happen when congestion is very low.

Fisk (1984) explains from game theory that the iterative assignment control procedure is actually a Nash non-cooperative game, as the signal controller does not take into account the expected behaviour of the drivers. She describes the NDP as a Stackelberg game, in which the leader (i.e. the traffic manager, who sets the signals) knows how the follower (i.e. the driver) will react to any decision he makes. Subsequently, the game is stated as a maxmin problem and a penalty approach is suggested for its approximate solution. However, possible difficulties with convexity will be encountered when realistic link cost functions are used.

Finally, some authors have suggested omitting the user equilibrium constraints (2)

(Dantzig et al., 1979; Marcotte, 1981). The resulting problem is convex, and at its solution point the flow pattern will be a system optimum; computation is straight forward. As the original NDP has the additional constraint of user equilibrium flows, this program will provide a lower bound to the optimum signal control problem under equilibrium conditions. The appropriateness of this approximation relies heavily on the closeness of the user equilibrium and system optimum flow pattern. Though some authors (Sheffi and Powell, 1983; and others) argue that these patterns should be very similar under low and high congestion (though not necessarily in the intermediate area), this depends heavily on the assumptions for the link cost functions. Dantzig et al. (1979) suggest solving the unconstrained problem after which an equilibrium re-assignment should be carried out with green times fixed. These two solutions would give an upper and a lower bound to the actual NDP.

From this overview it may be clear that as yet no efficient solution algorithms to the NDP exist. The methods described here suffer from one or more of the following shortcomings:

1) The solution (if any!) is only approximated, as heuristics must be applied to overcome the computational complications associated with the NDP.

2) They can only be applied to small networks, with a limited number of variables and constraints.

3) The assumptions made are too restrictive. In order to enforce a better behaviour on the NDP, strong assumptions are needed, particularly with respect to the delay formula and network structure. This severely undermines the real-life applicability of these methods.

6 Overall conclusions

Optimum signal control in an environment where drivers follow a user equilibrium is still an unsolved problem. The approach via the network design problem suffers from a lack of solution algorithms, and the general ill-conditioning of the problem. In addition this approach requires extensive data, such as e.g. an OD matrix and a complete network description.

On the other hand, application of local control policies is straightforward, and requires only local information. In interaction with traffic assignment such policies can be used to approximate a solution to the network design problem. This iterative assignment control procedure has been despised by many researchers but, in my view, possesses many positive features. In the first place it is easy to solve. In addition it allows the use of realistic network descriptions and complex cost functions.
Further, it enables us to investigate the long-term behaviour of vehicle actuated control, which may well follow an iterative course of signal and route changes. This behaviour could result in a deteriorating system, as examples by various authors have illustrated. This shows the need for an insight in the iterative behaviour of signals and drivers' route choice and for development of advantageous control policies, as already stated by Maher and Akcelik (1975).

The non-cooperative behaviour of drivers is a major obstacle in all these calculations. Because drivers try to minimise their personal travel cost, Braess' paradox can occur (Braess, 1968), in which case network improvements (such as improved signal timings) may result in deteriorated network conditions after drivers have adapted their route choice. Route guidance may be a way to (partially) overcome this, as it will supply us with some control over the driver's routing decisions. In such conditions the application of sensible control strategies may be of even greater importance for achieving optimal system benefits.

7. REFERENCE LIST

Abdulaal M and LeBlanc LJ (1979) "Continuous equilibrium network design models" *Transportation Research, Vol 13B, pp 19-32*

Allsop RE (1971) "Delay minimizing settings for fixed-time traffic signals at a single road junction" *Journal Inst Maths Applics, Vol 8, pp 164-185*

Allsop RE (1972) "Delay at fixed time traffic signals I: Theoretical analysis" *Transportation Science, Vol 6, N°3, pp 260-285*

Allsop RE (1974) "Some possibilities for using traffic control to influence trip distribution and route choice" *In Buckley DJ (ed): Proceedings of the 6th International Symposium on Transportation and Traffic Theory, Sydney, AH and AW Reed Pty Ltd, Artarmon, New South Wales, pp*

Allsop RE and Charlesworth JA (1977) "Traffic in a signal-controlled road network: an example of different signal timings inducing different routes" *Traffic Engineering and Control, Vol 18, N°5, pp 262-264*

Beckman MJ, McGuire, CB and Winsten CB (1956) "Studies in the economics of transportation *Yale University Press, New Haven, Conn*

Ben-Ayed O, Boyce DE and Blair CE (1988) " A general bilevel linear programming formulation of the network design problem" *Transportation Research, Vol 22B, N°4, pp 311-318*

Bovy PHL (1981) " The shortest path route choice principle - a test of its predictive quality (Het kortste-tijd-routekeuzeprincipe: een toetsing van de voorspellende kwaliteit)" *Verkeerskunde, Vol 32, pp 291-296 (in Dutch)*

Braess D (1968) " Ueber ein Paradoxen aus der Verkehrsplanung" *Unternehmensforschung, Vol 12, pp 258-268*

Branston D (1976) "Link capacity functions: a review" *Transportation Research, Vol 10, N°4, pp 223-236*

Bureau of Public Roads (1964) "Traffic assignment manual" *US Dept of Commerce, Urban Planning Division, Washington DC*

Dafermos SC (1971) "An extended traffic assignment model with applications to two-way traffic" *Transportation Science, Vol 5, N°4, pp 367-389*

Dafermos SC (1980) "Traffic equilibrium and variational inequalities" *Transportation Science, Vol 14, pp 42-54*

Dafermos SC and Sparrow FT (1969) "The traffic assignment problem for a general network" *J Res Nat Bureau of Standards, Vol 73B, N°2, pp 91-118*

Dantzig GB, Harvey RP, Landsdown ZF, Robinson DW and Maier SF (1979) "Formulating and solving the network design problem by decomposition" *Transportation Research, Vol 13B, N°1, pp 5-17*

Davidson KB (1966) "A flow-travel time relationship for use in transportation planning" *Proc Aust Road Res Board, pp 183-194*

Dickson TJ (1981) "A note on traffic assignment and signal timings in a signal-controlled road network" *Transportation Research, Vol 15B, N°4, pp 267-271*

Fellendorf M (1989) "A comparison of British and German signal control methodology" *Paper presented at the 21st UTSG Conference, Edinburgh (unpublished)*

Fisk CS (1984) "Optimal signal controls on congested networks" *In Volmuller J and Hamerslag R (eds): "Proceedings on the 9th International Symposium on Transportation and Traffic Theory", Delft, VNU Science Press, Utrecht, pp 197-216*

Frank M and Wolfe P (1956) "An algorithm for quadratic programming" *Naval Research Logistics Quarterly, Vol 3, pp 95-110*

Friesz TL and Harker PT (1985) "Properties of the iterative optimization-equilibrium algorithm" *Civil Engineering Systems, Vol 2, pp 142-154*

Heydecker BG (1981) "Some consequences of signal control policies for traffic assignment" *Paper presented at the 13th Annual UTSG Conference, Leeds*

(unpublished)

Heydecker BG (1983) "Some consequences of detailed junction modelling in road traffic assignment" *Transportation Science, Vol 17, Nº3, pp 263-281*

Heydecker BG (1986) "The equilibrium network design problem: a critical review" *In Odoni A and Szegö G (eds): Flow control on congested networks, NATO ASI Series F*

Heydecker BG and Khoo TK (1990) "The equilibrium network design problem" *Paper presented at the 22nd UTSG Conference, Hatfield (unpublished)*

Heydecker BG and Silcock JP (1989) "Anglo-German comparison of signal design methods: the design exercise" *Paper presented at the 21st UTSG Conference, Edinburgh (unpublished)*

Heydecker BG, Van Vliet D and Van Vuren T (1989) "Optimum signal offsets for traffic assignment networks" *In Yagar S and Rowe SE (eds): "Traffic Control Methods", Engineering Foundation, New York, pp 295-305*

Hunt PB, Robertson DI, Bretherton RD and Winton RI (1981) "SCOOT - a traffic responsive method of coordinating signals" *TRRL Report LR 1014, Crowthorne*

Hutchinson TP (1972) "Delay at a fixed-time traffic signal II: Numerical comparisons of some theoretical expressions" *Transportation Science, Vol 6, Nº3, pp 286-305*

LeBlanc LJ and Boyce DE (1986) " A bilevel programming algorithm for exact solution of the network design problem with user-optimal flows" *Transportation Research, Vol 20B, Nº3, pp 259-265*

Magnanti T and Wong R (1984) "Network design and transportation planning: models and algorithms" *Transportation Science, Vol 18, pp 1-55*

Maher M J and Akçelik R (1975) "The re-distributional effects of an area traffic control policy" *Traffic Engineering and Control, Vol 16, Nº9, pp 383-385*

Marcotte P (1981) "An analysis of heuristics for the continuous network design problem" *in Hurdle VF et al (eds): "Proceedings of the 8th International Symposium on Transportation and Traffic Theory", Toronto, University of Toronto Press, pp 27-34*

Marcotte P (1983) "Network optimization with continuous control parameters" *Transportation Science, Vol 17, pp 181-197*

Miller AJ (1963) "Settings for fixed-cycle traffic signals" *Operations Research Quarterly, Vol 14, Nº4, pp 373-386*

Newel GF (1965) "Approximation methods for queues with application to the fixed-cycle traffic light" *SIAM Review, Vol 7, pp 223-239*

Powell WB and Sheffi Y (1982) "The convergence of equilibrium algorithms with pre-determined step sizes" *Transportation Science, Vol 16, Nº1, pp 45-55*

Robertson DI (1969) "TRANSYT method for area traffic control" *Traffic Engineering and Control, Vol 11, Nº10, pp 276-281*

Sheffi Y and Powell WB (1983) "Optimal settings over transportation networks" *Journal of Transportation Engineering, Vol 109, Nº6, pp 824-839*

Smith MJ (1979a) "Traffic control and route choice: a simple example" *Transportation Research, Vol 13B, Nº4, pp 289-294*

Smith MJ (1979b) "The existence, uniqueness and stability of traffic equilibria" *Transportation Research, Vol 13B, Nº4, pp 295-304*

Smith MJ (1981a) "The existence of an equilibrium solution to the traffic assignment problem when there are junction interactions" *Transportation Research, Vol 15B, Nº6, pp 443-452*

Smith MJ (1981b) "Properties of a traffic control policy which ensure the existence of a traffic equilibrium consistent with the policy" *Transportation Research, Vol 15B, Nº6, pp 453-462*

Smith MJ (1985) "Traffic signals in assignment" *Transportation Research, Vol 19B, N°2, pp 155-160*

Tan H and Gershwin S and Athans M (1979) "Hybrid optimization in urban traffic networks" *Final Report, US-DOT-TSC-RSPA-79-7*

Van Vuren T and Jansen GRM (1988) "Recent developments in path finding algorithms: an overview" *Transportation Planning and Technology, Vol 12, pp 57-71*

Wardrop JG (1952) "Some theoretical aspects of road traffic research" *Proc Inst Civ Eng, Part II, pp 325-378*

Webster FV (1958) "Traffic signal settings" *Road Research Technical Paper, N°39, HMSO, London*

Wootton HJ, Ness MP and Burton RS (1981) "Improved direction signs and the benefits for road users" *Traffic Engineering and Control, Vol 22, pp 264-268*

Figure 1. Davidson's delay curve for different values of I
Source: Davidson (1966)

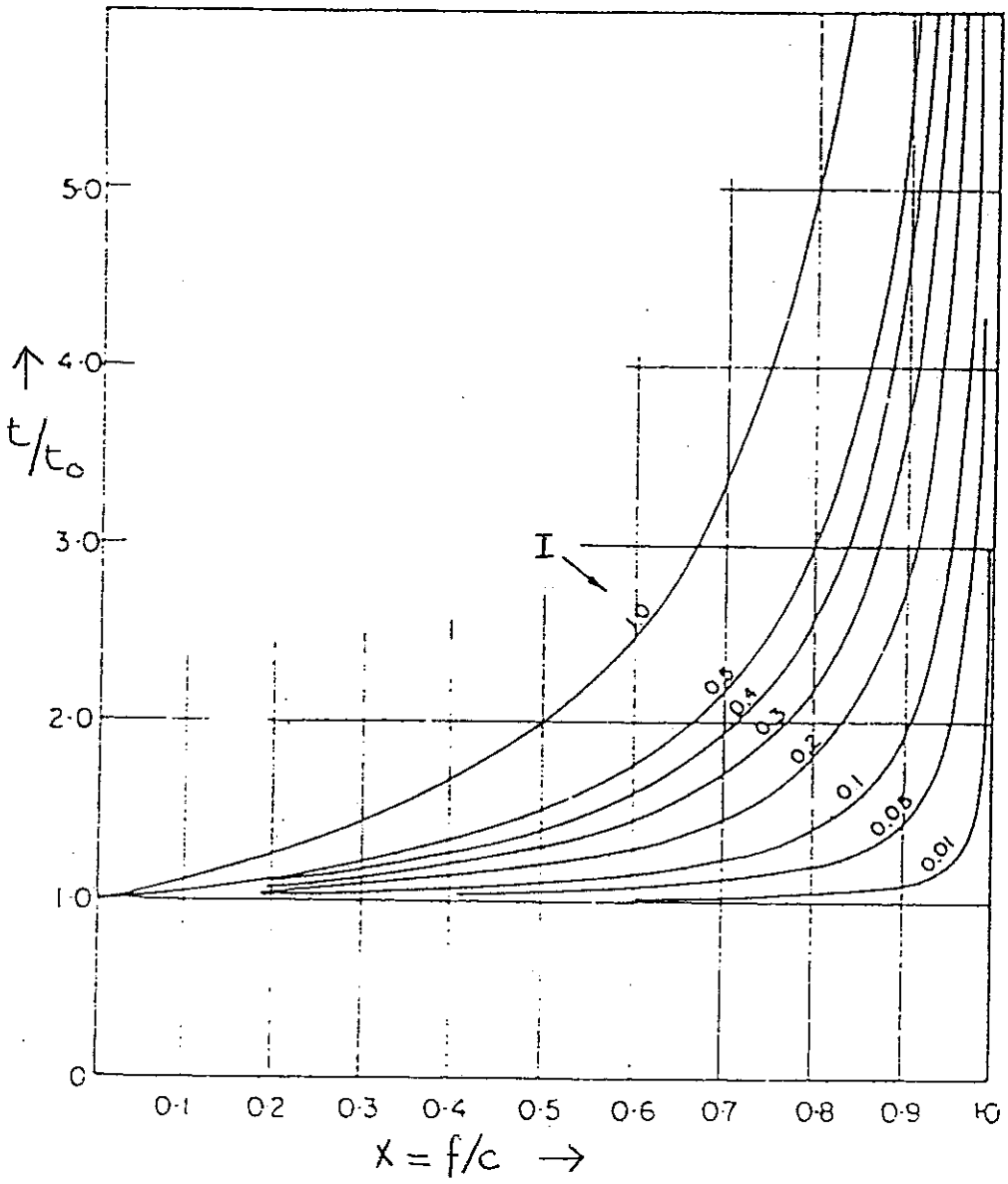

Figure 2. Pollaczek-Khintchine delay curve and deterministic queueing curve

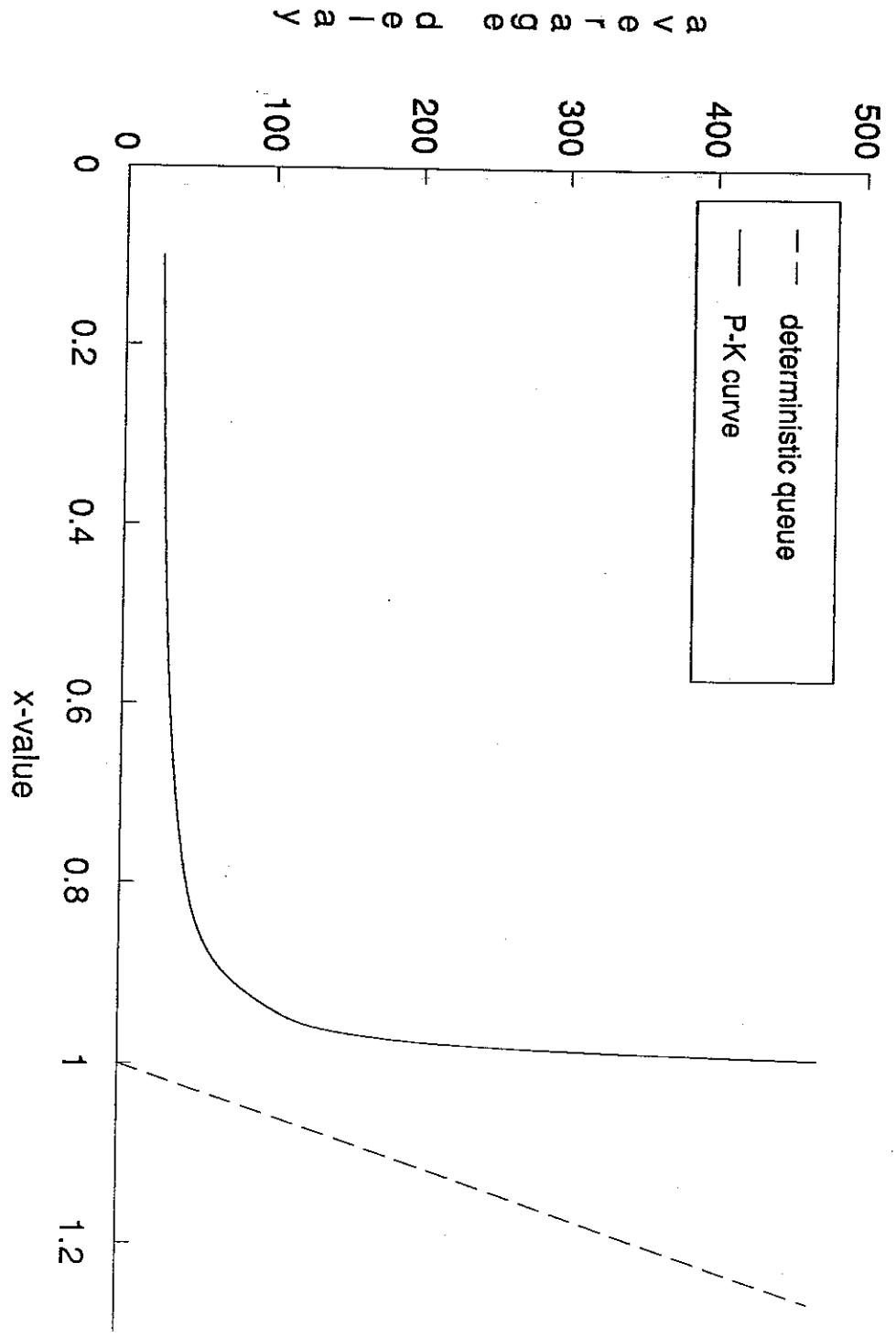

Figure 3. Pollaczek-Khintchine delay curve, deterministic queueing curve and sheared delay curve.