Available online at www.sciencedirect.com
SCIENCE DIRECT®
Transportation Research Part A 40 (2006) 227–243
TRANSPORTATION RESEARCH PART A
www.elsevier.com/locate/tra

Transportation network improvement and tolling strategies: The issue of intergeneration equity

W.Y. Szeto [a,*], Hong K. Lo [b]

[a] *Department of Civil, Structural, and Environmental Engineering, Trinity College, University of Dublin, Dublin 2, Ireland*
[b] *Department of Civil Engineering, Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong, PRC*

Received 1 January 2005; accepted 23 June 2005

Abstract

Existing transportation network design studies focus on optimizing the network for a certain future time but without explicitly defining the time dimension within the formulation. This study extends the consideration by formulating the time-dependent network design problem. With this extension, one can plan for the optimal infrastructure improvement timetable, the associated financial arrangement, and tolling scheme over the planning horizon. In addition, this extension enables the pursuit of important considerations that are otherwise difficult, if at all possible, with the traditional timeless approach. Through the time-dependent framework, this study examines the issue of intergeneration equity according to the user and social perspectives. Basically, should the present generation build the full-blown network, or should users at the time pay for future incremental upgrades? Using a gap function to measure the degree of intergeneration equity achieved, this study illustrates that there are tradeoffs between societal and individual perspectives. Nevertheless, this study suggests ways whereby the planner can trade the level of equity to be attained with the overall network performance. In this way, some gradual measures can be introduced to the network design to compromise between these two perspectives.

Keyword: Time-dependent transportation network design

* Corresponding author. Tel.: +353 1 6083646; fax: +353 1 6773072.
E-mail addresses: ceszeto@yahoo.com.hk (W.Y. Szeto), cehklo@ust.hk (H.K. Lo).

0965-8564/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.tra.2005.06.004

1. Introduction

Transportation infrastructure is in an active phase of planning and development in many parts of the world, especially Asia. Transportation projects are expensive. In times of constrained government expenditures, they must be carefully scrutinized for cost-effectiveness. Traditionally, this analysis belongs to the discipline of transportation network design (e.g. LeBlanc, 1975; Boyce and Janson, 1980; Marcotte, 1986; Chen and Alfa, 1991; Friesz et al., 1993; Davis, 1994; Meng et al., 2001; Chiou, 2004). Transportation network design is usually formulated as a special class of bi-level programming problems, in which the upper level problem is the network planner's problem and the lower level problem is either the traffic assignment problem or the trip distribution/assignment problem.

The existing network design studies focus on optimizing the network for a certain future time but without explicitly defining the time dimension within the formulation. To broaden the consideration, Lo and Szeto (2003) introduced the time dimension to the network design problem and analyzed the continuous network design problem (CNDP) by focusing on network link capacity expansions. For simplicity, we refer to this extension as the CNDP-T. The time scale considered in the CNDP-T is typically in years, as compared with the second-to-second scale of traffic dynamics, or the day-to-day scale of route choice dynamics. With this extension, one can plan for the optimal infrastructure improvement timetable, associated financial arrangements, and tolling schemes over the planning horizon, while allowing for the consideration of time-variant elastic travel demands (Szeto and Lo, 2003, 2004; Lo and Szeto, 2004, 2005). In this extension, the basic question is to determine the optimal capacity improvements (and tolls, if any) over time.

With the time dimension explicitly defined, the CNDP-T formulation enables the pursuit of important considerations that are otherwise difficult, if at all possible, with the traditional timeless approach. One interesting question is on the issue of intergeneration equity. Within the context of network design, the time scale of each generation is taken to be five to ten years, and intergeneration considerations to be twenty to thirty years involving a few generations, as consistent with the time scale of transportation infrastructure planning. In each generation, we consider the impacts of network improvement and tolling schemes on the transportation facility users and non-users, developers, as well as the overall society. It can happen that a scheme that achieves optimal combined benefits across generations may not benefit each generation equally. With the current trend of financing transportation infrastructure through user-pay principles, cost recovery (or even profit making) through toll revenue over the planning horizon becomes an important question. The basic question is who should pay for the transportation infrastructure project? Should the present generation build the full-blown network, or should users at the time pay for future incremental upgrades? Which approach is more efficient, and more equitable between the present and future generations? Is it possible to plan for transportation improvement and tolling schemes over the planning horizon so as to maintain a similar level of social welfare for all generations? In short, introducing the time dimension opens up a new way to study this set of theoretically interesting, yet practically important questions.

In this paper, we develop a set of time-dependent network design formulations via the framework by Lo and Szeto (2005) to account for the present values of costs and benefits incurred on

the present as well as future generations, and analyze the issue of intergeneration equity. We define intergeneration equity with two aspects, namely social equity and user equity. The former is measured by the total social surplus per capita, discounted for the time effect, which represents the gross average net benefit of the transportation infrastructure projects on society from generation to generation. The latter measures the direct impact on users, as reflected by the toll and generalized travel cost, both discounted for the time effect. In a network context, these two measures do not necessarily go hand in hand. Toll facility users may benefit other users of the network by spreading traffic and lowering congestion, possibly leading to the situation that the overall society gains, at the expense of individual toll facility users. The extent and occurrence of this situation depends on the network structure and the tolls to be collected. It suffices to say that there may be a tradeoff between these two types of equity considerations. This study then defines a gap function to measure the degree of intergeneration equity achieved and employs the generalized reduced gradient method to solve the formulations. A scenario is set up to illustrate the problem and the tradeoff among these equity measures.

This study includes a discussion on two possible extensions of the models developed herein. Specifically, multi-objective network design formulations are provided for addressing the tradeoff between overall social benefit and intergeneration equity. Moreover, a discussion on dynamic demand evolution is provided to illustrate how these formulations can be further extended to capture dynamic demand changes in response to travel cost changes in previous generations. These are important extensions to make the formulations more reflective of reality and are left for future studies.

The outline of this paper is as follows. Section 2 develops the network design formulations with intergeneration equity considerations. Section 3 provides the numerical studies. Section 4 depicts possible extensions to the models. Finally, Section 5 provides some concluding remarks.

2. Formulations

We consider a general transportation network with multiple Origin–Destination (OD) flows over the planning horizon $[0, T]$. The horizon is divided into N equal design periods. The following intergeneration equity measures are focused in this paper:

2.1. Intergeneration equity measures

Two aspects of equity are considered in this study, namely user equity and social equity.

2.1.1. Intergeneration user equity

Intergeneration user equity is measured by the link tolls and discounted OD travel costs. The tolls are the most direct measure felt by users in each planning period. However, this measure ignores the service quality, namely travel time or congestion, which travelers are also concerned with. Moreover, for the same toll charge on a particular link, due to demand increases, the resultant travel times usually increase with time. To account for this, the more appropriate measure of discounted OD travel cost is used, which includes both travel times and tolls, discounted for their future values.

2.1.2. Intergeneration social equity

Intergeneration social equity is measured by the discounted social surplus (DSS) per capita, denoted as SE_τ, which captures the effect of the value of time, societal benefit, and the population size in each planning period. Mathematically, it can be formulated as:

$$SE_\tau = \frac{DSS_\tau}{\sum_{rs} \tilde{q}_\tau^{rs}}, \tag{1}$$

where DSS_τ is the DSS in period τ, and $\tilde{q}_\tau^{rs}$ is the potential travel demand between OD pair rs in period τ.

The denominator in (1) sums over the potential demands for all OD pairs, representing the total capita for the region under planning. The numerator, DSS in period τ, is defined as the sum of discounted consumer surplus DCS_τ and discounted profit DP_τ in that period, written as:

$$DSS_\tau = DCS_\tau + DP_\tau. \tag{2}$$

DCS_τ and DP_τ are to be precisely defined in the following subsections. In this formulation, in addition to measuring the societal measure of consumer surplus, we also include profits of tolled facility operations as part of social surplus.

2.1.2.1. Discounted consumer surplus (DCS). DCS in period τ, DCS_τ, is the sum of consumer surpluses CS_τ^{rs} for all OD pairs in that period, adjusted to their present values:

$$DCS_\tau = \sum_{rs} \frac{CS_\tau^{rs}}{(1+i)^{\tau-1}}, \tag{3}$$

where $\frac{1}{(1+i)^{\tau-1}}$ is the discount factor for period τ and i is the interest rate. The consumer surplus CS_τ^{rs} in (3) is defined as:

$$CS_\tau^{rs} = n\left[\int_0^{q_\tau^{rs}} D_\tau^{rs-1}(v)\,dv - \pi_\tau^{rs} q_\tau^{rs}\right], \tag{4}$$

where n is a factor converting hourly consumer surplus to the consumer surplus for the whole period, and $D_\tau^{rs-1}(\cdot)$, q_τ^{rs}, and π_τ^{rs} are, respectively, the inverse demand function, travel demand, and travel cost for OD pair rs in period τ. The first term in the square bracket in (4) is the hourly total travel cost the demand q_τ^{rs} would be willing to pay, whereas the second term is the hourly total travel cost they actually pay. Consumer surplus internalizes the effect of network congestion and the public's propensity to travel. For the same network and demand characteristics, a higher consumer surplus implies a better performing system.

2.1.2.2. Discounted profit (DP). The discounted profit DP_τ is the sum of the discounted profit $P_{b,\tau}$ of each link, in which $P_{b,\tau}$ is the difference between the discounted revenue $R_{b,\tau}$ of toll link b in period τ and its discounted cost $C_{b,\tau}$. Mathematically, DP is expressed as:

$$DP_\tau = \sum_b P_{b,\tau} = \sum_b [R_{b,\tau} - C_{b,\tau}]. \tag{5}$$

The discounted toll revenue $R_{b,\tau}$ on link b in period τ in (5) is the product of the toll $\rho_{b,\tau}$ and the volume $nv_{b,\tau}$ on that link in that period multiplied by the discount factor $\frac{1}{(1+i)^{\tau-1}}$, and can be written as:

$$R_{b,\tau} = \frac{\rho_{b,\tau}(nv_{b,\tau})}{(1+i)^{\tau-1}}, \tag{6}$$

where n is a factor converting the revenue from an hourly basis to a period basis.

The discounted cost $C_{b,\tau}$ in (5) is related to the source of capital for the construction or improvement works and the actual improvement cost of that link. Here, we assume the funding for the construction works is raised through loans and is returned through fixed-payment installments. Hence, there are two time indexes associated with this financial arrangement: loan time and payment time. In this subsection, we denote t as the former and τ as the latter. For the sake of simplicity, the loan time is taken to be the time of enhancements of the link. It is no more difficult to relax this assumption to some other time within this framework.

The loan $\tilde{C}_{b,t}$ for the improvement on link b in loan or enhancement period t is based on the following construction cost function:

$$\tilde{C}_{b,t} = \mu t_b^0 y_{b,t}(1+r)^{t-1}, \tag{7}$$

where μ is a construction cost parameter; t_b^0 denotes the free-flow travel time of link b; $y_{b,t}$ is the capacity gain on link b in enhancement period t; The term $(1+r)^{t-1}$ represents the inflation factor: for the same capacity enhancement, the improvement cost increases by r% each period. Eq. (7) states that the discounted improvement cost of link b is directly proportional to the extent of the widening (and hence capacity gain $y_{b,t}$), its length (as represented by its free-flow travel time t_b^0), the construction cost parameter μ, and the inflation factor $(1+r)^{t-1}$.

Let the number of installments be M; the loan $\tilde{C}_{b,t}$ is related to the *constant* payment $P_{b,t}$ of each installment for the improvement on link b in enhancement period $t \leqslant \tau$ by:

$$\tilde{C}_{b,t} = \sum_{\tau=t}^{t+M-1} \frac{P_{b,t}}{(1+i)^{\tau-1}}. \tag{8}$$

According to (8), the loan is equal to the total discounted payments over M installment periods, starting from t.

Let $\tilde{P}_{b,t,\tau}$ be the indicator representing the payment returned to the bank in payment period τ due to the improvement of link b in enhancement period t with $\tilde{P}_{b,t,\tau} = \begin{cases} P_{b,t} & t \leqslant \tau \leqslant t+M-1 \\ 0 & \text{otherwise} \end{cases}$. The indicator $\tilde{P}_{b,t,\tau}$ is equal to the constant payment $P_{b,t}$ within the installment period and zero otherwise. The discounted cost $C_{b,\tau}$ can thus be obtained by summing over all the payments to the bank in period τ due to the improvement of link b from $t=1$ to $t=\tau$, discounted to present value terms:

$$C_{b,\tau} = \sum_{t=1}^{\tau} \frac{\tilde{P}_{b,t,\tau}}{(1+i)^{\tau-1}}. \tag{9}$$

2.2. Time-dependent network design formulation

For the sake of completeness, this section reviews the time-dependent network design formulation proposed in Lo and Szeto (2005). The time-dependent network design problem includes N time-dependent traffic assignment sub-problems and a set of design constraints.

2.2.1. Traffic assignment constraints

For each traffic assignment sub-problem, there are Wardropian and flow conservation conditions, elastic demand functions, and non-negativity and definitional constraints. Each traffic assignment sub-problem is linked to each other by potential demand constraints.

2.2.1.1. Wardropian conditions. At each time period τ, we assume each traveler follows Wardrop's principle (1952), which states that the travel costs of all used routes between the same OD pair are equal and minimal. This principle can be expressed as the following nonlinear complementarity conditions:

$$f_{p,\tau}^{rs}[\eta_{p,\tau}^{rs} - \pi_{\tau}^{rs}] = 0, \quad \forall rs, p, \tau, \tag{10}$$

$$\eta_{p,\tau}^{rs} - \pi_{\tau}^{rs} \geqslant 0, \quad \forall rs, p, \tau, \tag{11}$$

where $f_{p,\tau}^{rs}$ and $\eta_{p,\tau}^{rs}$ are, respectively, the representative hourly flow and the path travel cost on path p between OD pair rs in period τ. From (10), if some flows are on path p in period τ, (i.e, $f_{p,\tau}^{rs} > 0$), the condition $[\eta_{p,\tau}^{rs} - \pi_{\tau}^{rs}] = 0$ is satisfied, implying that the path cost $\eta_{p,\tau}^{rs}$ must equal the minimum travel cost π_{τ}^{rs}. If path p carries no flow in period τ, the term $[\eta_{p,\tau}^{rs} - \pi_{\tau}^{rs}]$ in (10) is unrestricted and the travel cost $\eta_{p,\tau}^{rs}$ can be greater than or equal to π_{τ}^{rs} according to (11).

2.2.1.2. Definitional constraints. Definitional constraints define the relationships among the hourly path flow $f_{p,\tau}^{rs}$, toll $\rho_{b,\tau}$, capacity enhancements $y_{a,\tau}$, path travel costs $\eta_{p,\tau}^{rs}$, hourly link flow $v_{a,\tau}$, and link travel time $t_{a,\tau}$. The relationship between path flows and link flows is:

$$v_{a,\tau} = \sum_{rs} \sum_{p} f_{p,\tau}^{rs} \delta_a^p, \quad \forall a, \tau, \tag{12}$$

where δ_a^p is a link-path incidence indicator, which equals one if link a is on path p, and zero otherwise. Eq. (12) states that the flow on link a is obtained by summing the corresponding path flows on that link.

The link travel time is given by the link performance function:

$$t_{a,\tau} = t_a^0 \left[1 + \bar{\alpha} \left(\frac{v_{a,\tau}}{c_a + \sum_{i=1}^{\tau} y_{a,i}} \right)^{\bar{\beta}} \right], \quad \forall a, \tau, \tag{13}$$

where t_a^0 and c_a denote the free-flow travel time and initial capacity of link a, respectively; $\bar{\alpha}, \bar{\beta}$ are parameters of the link performance function. The summation term in (13) represents the total capacity enhancements of link a up to period τ. Therefore, the denominator inside the bracket denotes the link capacity in period τ after implementing the enhancements before and inclusive of period τ. When $\bar{\alpha} = 0.15$ and $\bar{\beta} = 4$, Eq. (13) is reduced to the typical Bureau of Public Roads (BPR) function.

Path travel cost is the sum of the link costs on that path, expressed as:

$$\eta_{p,\tau}^{rs} = \sum_{a \in A/B} \psi t_{a,\tau} \delta_a^p + \sum_{b \in B} (\psi t_{b,\tau} + \rho_{b,\tau}) \cdot \delta_b^p, \quad \forall rs, p, \tau, \tag{14}$$

where ψ is the cost of unit travel time; $\rho_{b,\tau}$ is the toll on toll link b in period τ; A and B denote, respectively, the set of links and toll links, with $B \subset A$. The first term in (14) is the sum of travel-time costs $\psi t_{a,\tau}$ on toll-free links along path p, whereas the second term is the sum of travel costs on toll links along the same path p which includes the link tolls $\rho_{b,\tau}$.

2.2.1.3. Flow conservation and non-negativity conditions. Like other traffic assignment problems, flow conservation must be satisfied in each traffic assignment sub-problem, which requires the sum of flows between the same OD pair in a particular period equal to the corresponding demand. Mathematically, this can be expressed as:

$$\sum_p f_{p,\tau}^{rs} = q_\tau^{rs}, \quad \forall rs, \tau, \tag{15}$$

where q_τ^{rs} is the travel demand of OD pair rs in period τ. Path flows, by definition, must be non-negative:

$$f_{p,\tau}^{rs} \geqslant 0, \quad \forall rs, p, \tau. \tag{16}$$

2.2.1.4. Elastic demand functions. The travel demand q_τ^{rs} of each OD pair is a decreasing function of its minimum travel cost π_τ^{rs}. For simplicity, we adopt a linear demand function:

$$q_\tau^{rs} = \tilde{q}_\tau^{rs} - \gamma^{rs}\pi_\tau^{rs}, \tag{17}$$

where γ^{rs} is the parameter of the travel demand function of OD pair rs, and $\tilde{q}_\tau^{rs}$ is the potential demand between OD pair rs in period τ.

2.2.1.5. Potential demand constraints. Potential demand constraints describe the relationship between the potential demands $\tilde{q}_\tau^{rs}$ in the two successive periods. The potential demand difference in the two successive periods represents the potential travel growth due to population growth. For simplicity, this paper employs the linear potential demand function, defined as:

$$\tilde{q}_\tau^{rs} = \tilde{q}_{\tau-1}^{rs}[1 + h^{rs}], \tag{18}$$

where h^{rs} is the growth rate of potential demand between OD pair rs.

2.2.2. Design constraints

2.2.2.1. Link improvement constraints. For practical and physical reasons, roads or highways rarely have more then a few lanes. Therefore, the formulation includes link improvement constraints, expressed as:

$$c_a + \sum_{i=1}^{\tau} y_{a,i} \leqslant u_a, \quad \forall a, \tau, \tag{19}$$

$$y_{a,\tau} \geqslant 0, \quad \forall a, \tau. \tag{20}$$

The maximum allowable capacity constraint (19) limits the total capacity $c_a + \sum_{i=1}^{\tau} y_{a,i}$ of link a in year τ to be less than its maximum allowable capacity u_a. Eq. (20) is the non-negativity condition of capacity improvements.

2.2.2.2. Cost recovery constraint. The cost recovery condition ensures the total discounted cost not to be greater than the total discounted toll revenue. This requires that the sum of the discounted profit in each period, the total discounted profit, is non-negative:

$$\sum_{\tau} \mathrm{DP}_{\tau} \geqslant 0, \tag{21}$$

where DP_{τ} follows the definitions in (5)–(9).

2.2.3. Objective function

The objective is defined from the perspective of society, measured by the total discounted social surplus (TDSS) for the entire planning horizon (by summing the DSS in each period):

$$\mathrm{TDSS} = \sum_{\tau} \mathrm{DSS}_{\tau}. \tag{22}$$

2.2.4. Basic network design formulation

The time-dependent network design formulation can be expressed as:

Basic model: $z^{*}_{\mathrm{basic}} =$ Maximize TDSS
Subject to Time-dependent traffic assignment constraints (10)–(18), and
Design constraints (19)–(21),

where TDSS follows the definitions in (2)–(9) and (22). This formulation basically designs the network from the perspective of society subject to time-dependent traffic assignment constraints, cost recovery constraints, and link improvement constraints. In this basic model, the issue of intergeneration equity is not taken into account. In the following subsection, a general framework is proposed to capture the intergeneration equity considerations as discussed in Section 2.1.

2.3. Intergeneration equity-based formulation

Generally, the equity consideration can be incorporated in the formulation via two approaches. One is to embed a mathematical expression for equity in the objective function, forming a bi-objective function with the original objective. The bi-objective formulation is subsequently dealt with via the method of preemptive priority (lexicographic ordering), the method of weighted objectives and metrics, and hybrid ones (Miettinen, 1999). The second approach is to capture the equity consideration as constraints in the formulation, forming the minimum requirement to be fulfilled. This is referred to as the ε-constraint method and is adopted in this study. Before describing this approach in some details, we define a gap function to capture the equity consideration:

$$G = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \|\mathbf{w}_i - \mathbf{w}_j\|^2,$$

where $\mathbf{w}_{\tau}$ denotes the vector of the equity measures at period τ such as link tolls, discounted OD travel costs, TDSS per capita, and others. The gap function is obtained by summing the squares of

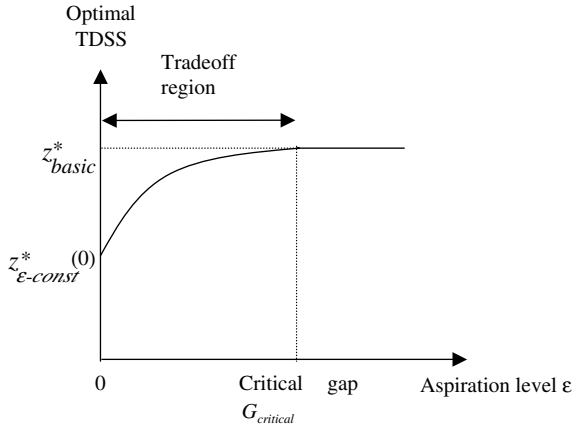

Fig. 1. Maximizing TDSS versus maintaining intergeneration equity.

the differences among all equity performance measures. This gap function is always non-negative and attains zero when all $\mathbf{w}_\tau$'s are equal, implying that the measure of interest is maintained the same over time. Moreover, the value of this function measures the degree of equity achieved over generations. The smaller is the gap; the more equitable is the solution.

The equity-based time-dependent network design model is written as:

$$\begin{aligned} &\varepsilon\text{-constraint model:} && z_{\varepsilon-\text{const}}(\varepsilon) = \text{Maximize TDSS} \\ &\text{Subject to} && \text{the same constraint as in the basic model, and} \\ & && G \leqslant \varepsilon, \end{aligned} \tag{23}$$

where the aspiration level ε is a non-negative number controlling the size of the gap and hence represents the minimum degree of equity to be achieved. When ε is zero, G equals zero, leading to $\mathbf{w}_\tau = \mathbf{w}_{\tau+1}$, $\forall \tau = 1, N-1$ or that absolute intergeneration equity is guaranteed. In expressing equity as a constraint, as one may expect, the tighter is the aspiration level, the objective TDSS to be achieved may be lowered, depending on whether the equity constraint is active or not, as illustrated in Fig. 1. The x-axis in Fig. 1. is the aspiration level whereas the y-axis is the corresponding optimal TDSS. z^*_{basic} is the optimal objective value under no equity consideration. $z^*_{\varepsilon\text{-const}}(0)$ is the optimal objective value when $\varepsilon = 0$. The critical gap G_{critical} is the value of the aspiration level that renders the equity constraint binding. If ε is large, the equity constraint is not binding; TDSS remains the same as the optimal objective value z^*_{basic} of the basic model. However, if ε is gradually reduced, the constraint becomes tighter, which in turn reduces the optimal TDSS to be achieved.

3. Numerical example

A scenario is set to illustrate three aspects of the equity-based network design formulation: (i) Intergeneration equity issues and the existence of multiple solutions, (ii) the tradeoff among different equity measures, and (iii) the tradeoff between maximizing TDSS and maintaining intergeneration equity. For ease of results exposition, a simple network is selected, although the

formulations are applicable for general networks. The network comprises 4 nodes and 6 links, as shown in Fig. 2. Nodes 1 and 4 are origins and node 2 is the destination. Existing and proposed new links are represented by solid and dashed lines respectively. The planning horizon is 30 years. The parameters in this scenario are as follows:

(a) Supply side parameters
- Initial link capacities: $c_2 = c_4 = 3600$ vph, $c_5 = c_6 = 1800$ vph.
- Maximum allowable capacities: $u_1 = u_3 = 4000$ vph.
- Free-flow travel times: $t_1^0 = 12$ min, $t_2^0 = t_3^0 = t_4^0 = 5$ min, $t_5^0 = 25$ min, $t_6^0 = 15$ min.
- Link performance function parameters: $\bar{\alpha} = 0.15$, $\bar{\beta} = 4$.

(b) Demand side parameters
- Potential demands at period 1: $\tilde{q}_1^{12} = 7000$ vph, $\tilde{q}_1^{42} = 5000$ vph.
- Growth rates: $h^{12} = 0.04$, $h^{42} = 0.02$.
- Parameters of the travel demand functions: $\gamma^{12} = \gamma^{42} = 100$ veh/h^2.

(c) Design parameters
- Value of time: $\psi =$ HK$ 30/h.
- Interest and inflation rates: $i = 0.03$, $r = 0.01$.
- Converting factor: $n = 87{,}600$.
- Construction cost parameter: $\mu =$ HK$ 6,000,000 /veh.
- Number of design intervals and installments: 3.

The results were obtained by solving the basic and ε-constraint models with zero aspiration level through the Generalized Reduced Gradient (GRG) algorithm (Abadie and Carpentier, 1969).

Multiple solutions can be found for the basic model without equity consideration. Fig. 3 gives the construction costs and toll revenues over the planning period for two possible solutions with the same optimal TDSS. According to this figure, both designs have similar improvement costs but the toll revenue of design 1 is higher than that of design 2. In other words, design 1 achieves the optimal TDSS by increasing the profits of the tolled facilities whereas design 2 achieves the same optimal TDSS by increasing the consumer surplus. As social surplus includes both profit and consumer surplus, even for the same TDSS, the beneficiaries are likely to be different. Methodologically, this finding highlights the non-uniqueness of solutions for this problem and that many network improvement and tolling plans are possible to achieve the same TDSS. One needs to sort out the most appropriate strategy among the multiple solutions for the local situation and context. This brings up the need of defining secondary objectives to facilitate the selection process.

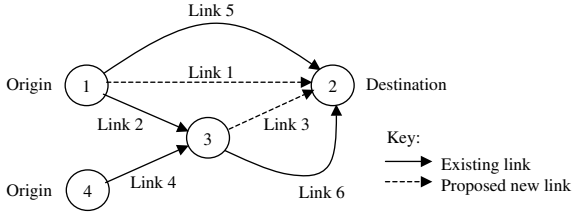

Fig. 2. The example network.

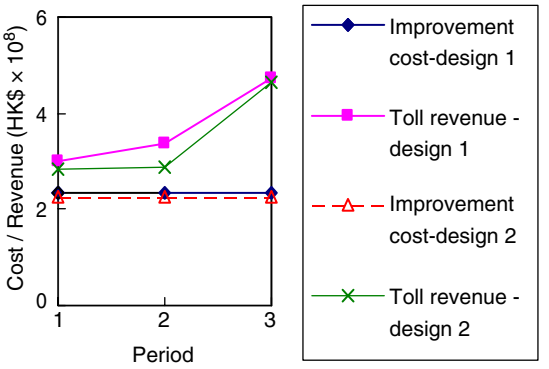

Fig. 3. Improvement costs and toll revenues over time.

Fig. 4 shows the equity measures for the two corresponding designs (as shown in Fig. 3) obtained from the basic model. As can be seen, inequity over generations does exist. The future generation will obtain much better benefits as measured by its DSS per capita. The DSS per capita for the third planning period is about 13% higher than that of the first in both cases. This increase is mostly due to the relatively higher profits earned by the tolled facilities in the third period. As far as the users (or travelers) are concerned, the ones in the third period would perceive that their benefits are diminishing as both the tolls and generalized travel costs are increased. As compared with the first period, the tolls and discounted travel costs in the third period increase, respectively, by around 34% and 15%.

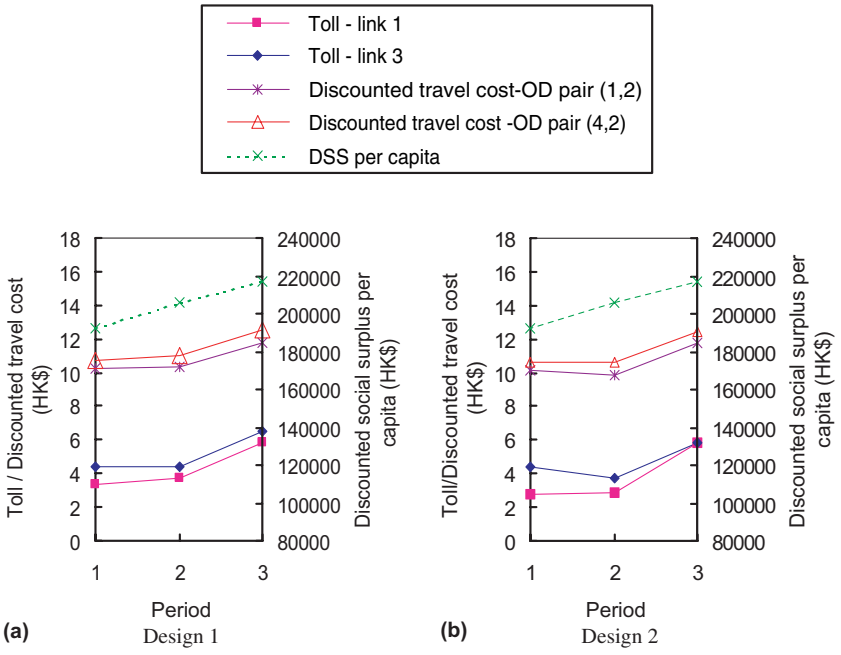

Fig. 4. Equity measures for designs 1 and 2.

One may introduce three types of constraints for intergeneration equity: constant tolls, constant discounted OD travel costs, and constant DSS per capita. The first two focus on the users' perspectives, whereas the last one on the overall societal impact. The results in Fig. 5 show that controlling for one measure of intergeneration equity would lead to inequity in other measures. Or, there are tradeoffs among these measures for intergeneration equity. Maintaining intergeneration equity for all three measures is unlikely to succeed. For the case of constant tolls (Fig. 5a), some travelers in the third planning period get a slightly lower discounted travel cost whereas the DSS per capita is much higher whose value increases steadily with time by about 7%. Moreover, as compared with the case of constant cost, the system charges higher tolls all through the three planning periods, resulting in generally higher generalized travel costs. For the case of maintaining constant travel costs over time (Fig. 5b), both the tolls and DSS per capita swing over time. In particular, the system does not achieve the same level of DSS per capita. Finally, for the case of constant DSS per capita (Fig. 5c), both the tolls and travel costs vary substantially over time.

The capacity improvement strategies for the three equity-based designs together with design 1 (without equity consideration) are shown in Fig. 6. Whereas design 1 frontloads all the improve-

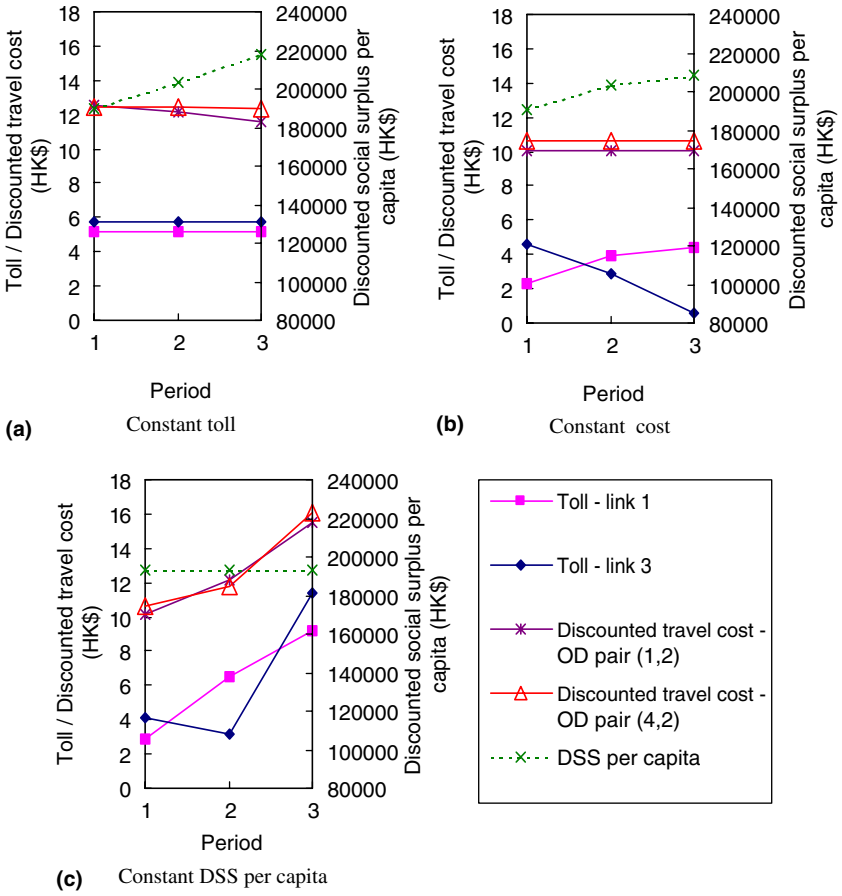

Fig. 5. Performance measures under three equity-based designs.

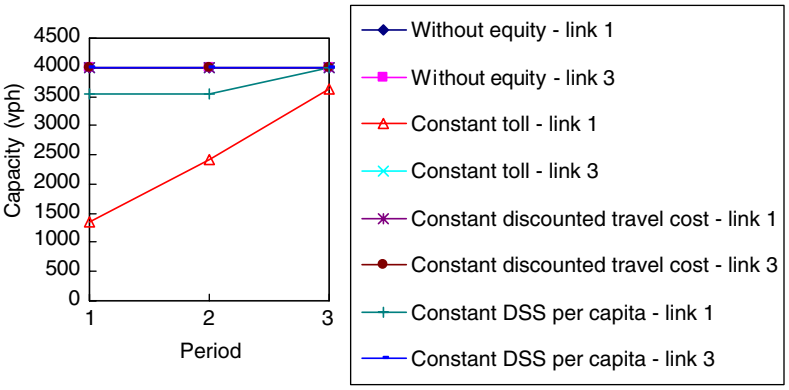

Fig. 6. Capacities over time comparison.

ments up to the maximum allowable capacity of 4000 vph in the first planning period, the equity considerations (constant toll and constant DSS per capita) do spread out the capacity improvements over time, implying that frontloading capacity improvements may not be good from the equity viewpoint. Nevertheless, to accomplish these various types of equity measures, it involves the tradeoff in achieving the overall TDSS for the whole network for the entire planning horizon, as shown in Fig. 7.

As shown in Fig. 7, the equity considerations do reduce the optimal TDSS achieved. In particular, the case of constant toll slightly lowers TDSS; the case of constant cost decreases TDSS by about 2%. The case of constant TDSS per capita reduces TDSS the most, about a 6% decrease. According to this result, it appears that the case of constant TDSS per capita, apparently the most equitable definition (as it considers both producer and consumer surpluses), exerts the tightest constraint. Whether this is always the case or specific to this example is something to revisit in the future.

In fact, in this formulation, one does not need to consider equity as an absolute requirement. One can relax the equity definitions through suitably defining the aspiration level ε to a positive number. For example, if we set $\varepsilon = 10^{10}$, then in effect the equity constraint becomes non-binding. Thus the optimal TDSS achieved in all three cases will be equal to TDSS achieved in the case without equity consideration, at HK\$ 8.02×10^9.

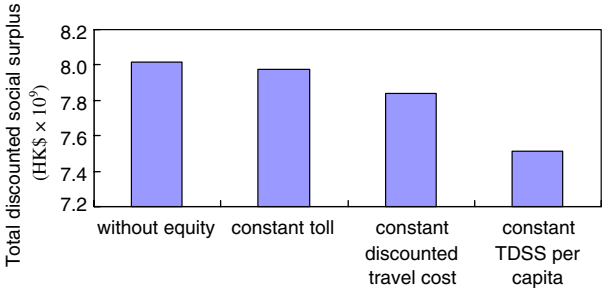

Fig. 7. Total discounted social surplus comparison.

4. Model extensions

4.1. Multi-objective formulations

As shown in the results, multiple solutions do exist in the network design problem, both with and without equity considerations. One can take advantage of this property of multiple solutions to address the equity concern. Basically, one can select from among the multiple solutions that best achieves the equity concern. For this purpose, three approaches can be considered: preemptive priority, method of weighted objectives, and hybrid.

4.1.1. Preemptive priority approach

The preemptive priority method prioritizes each of the objectives and optimizes them in order of priority. For this problem, we can first maximize TDSS to obtain z^*_{basic} from the basic model, and minimize the equity gap while fixing the optimal TDSS determined in the basic model. The preemptive priority formulation can be expressed as:

$$\begin{aligned}&\text{Preemptive priority model:} && G_{\text{critical}} = \text{minimize } G = \textstyle\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\|\mathbf{w}_i - \mathbf{w}_j\|^2\\ &\text{Subject to} && \text{the constraints as in the basic model, and}\\ & && \text{TDSS} = z^*_{\text{basic}}.\end{aligned}$$

This model selects from among the multiple solutions of the basic model to one that best accommodates the equity measure of interest while not sacrificing the optimal TDSS achieved.

4.1.2. The method of weighted objectives approach

In this approach, we combine the equity gap and the overall TDSS objective into one function with weights that sum to one. Let M be the weight between zero and one inclusively. Then the formulation can be expressed as:

$$\begin{aligned}&\text{Weighted objective model:} && \text{minimize } (1-M)\tfrac{z^*_{\text{basic}}-\text{TDSS}}{z^*_{\text{basic}}-z^*_{\varepsilon\text{-const}}(0)} + M\cdot G\\ &\text{Subject to} && \text{the constraints as in the basic model, and}\\ & && \text{TDSS} \geqslant z^*_{\varepsilon\text{-const}}(0). \qquad (24)\end{aligned}$$

This model minimizes the two objectives, $\frac{z^*_{\text{basic}}-\text{TDSS}}{z^*_{\text{basic}}-z^*_{\varepsilon\text{-const}}(0)}$ and the gap, simultaneously. As $z^*_{\varepsilon\text{-const}}(0)$ and z^*_{basic} are constants, minimizing the first objective is equivalent to maximizing TDSS. Constraint (24) ensures the solution to be within the feasible region, with TDSS between $z^*_{\varepsilon\text{-const}}(0)$ and z^*_{basic} inclusively. The weight M determines the tradeoff between maximizing TDSS and minimizing the gap. Basically, if M is zero, the weighted objective model reduces to the basic model. On the other hand, if M equals one, only G is minimized, which attains the value of zero at optimality, leading to $\mathbf{w}_\tau = \mathbf{w}_{\tau+1}$, $\forall\tau = 1,\ N-1$, implying absolute equity is the first priority. This weighted objective model then becomes the ε-constraint model with zero aspiration.

4.1.3. Hybrid approach

In general, one may also devise methods that combine the ε-constraint, preemptive priority, and method of weighted objectives approaches to develop models and select solutions that satisfy prescriptive conditions on the acceptable level of TDSS and temporal swings in the desirable equity measures.

4.2. Dynamic demand evolution

The "*equilibrium*" demand function (formed by both the elastic demand and potential demand functions (17), (18)) in Section 2.2 is actually the special case of the following:

$$q_\tau^{rs} = F^{rs}(\tilde{q}_\tau^{rs}, \tilde{q}_{\tau-1}^{rs}, \ldots, \tilde{q}_1^{rs}, \pi_\tau^{rs}). \tag{25}$$

According to (25), the demand in period τ, q_τ^{rs}, depends on the set of potential demands $\tilde{q}_\tau^{rs}$ from the first period up to period τ (i.e. $\tilde{q}_\omega^{rs}$, $\forall \omega = 1, \ldots, \tau$), but only on the lowest travel cost π_τ^{rs} within period τ. This demand function has two properties: First, the demand in period τ is assumed to be in equilibrium with the lowest travel cost within that period, i.e., demand responds immediately for any change in lowest travel cost or that complete demand adjustment is finished within one period. Second, since the demand in period τ does not depend on the lowest travel cost before period τ, changes in the lowest travel cost before τ have no effect on the demand in period τ. In other words, the demand function has no memory on the travel cost prior to period τ.

In reality, people do remember their previous travel costs and have delays in their responses to the changes in travel cost (Dargay and Goodwin, 1995). To capture this, we can apply the approach adopted in Dargay and Goodwin (1995) and extend the equilibrium demand curve in each period by a set of demand functions representing the *time-dependent adjustment process*:

$$q_\tau^{rs} = F^{rs}(\tilde{q}_\tau^{rs}, \tilde{q}_{\tau-1}^{rs}, \ldots, \tilde{q}_1^{rs}, \pi_\tau^{rs}, \pi_{\tau-1}^{rs}, \ldots, \pi_1^{rs}). \tag{26}$$

In this demand function, the demand in the current period depends on the entire set of potential demands and lowest travel costs from the first period up to period τ. In (26), the influences of all the previous lowest travel costs are included in the consideration, which allows for the capturing of delay responses to changes in the lowest travel costs. By adjusting the parameter on delay response, we can control the pace of the response.

Alternatively, demand function (26) can be expressed as:

$$q_\tau^{rs} = F^{rs}(\tilde{q}_\tau^{rs}, \pi_\tau^{rs}, q_{\tau-1}^{rs}). \tag{27}$$

In this function, the demand in the current period depends on the potential demands and the lowest travel cost in the same period, as well as the demand in the previous period. Recursively, the demand in the previous period depends on the potential demands and the lowest travel cost in the period before that, and so on. One can interpret (27) as a shorthand formulation of (26). One example is the double-logarithmic demand function:

$$\ln q_\tau^{rs} = \beta_c + \beta_{\tilde{q}} \ln \tilde{q}_\tau^{rs} + \beta_\pi \ln \pi_\tau^{rs} + \beta_q \ln q_{\tau-1}^{rs}, \tag{28}$$

where β_c, β_q, β_π, and β_q are parameters depending on the speed of the adjustment process or the response pace, with β_π to be the short run elasticity with respect to the lowest travel cost $\left(\frac{\mathrm{d}\ln q_\tau^{rs}}{\mathrm{d}\ln \pi_\tau^{rs}}\right)$

and $\frac{\beta_\pi}{1-\beta_q}$ the long run elasticity (equal to the sum of the short run elasticity in each period). The method of estimating these parameters can be found in Dargay and Goodwin (1994).

With this extension, the dynamic evolution of demand from one period to the next, including response delays, can be captured in our time-dependent network improvement framework. We believe this extension is an important future research direction, as more realistic modeling of dynamic demand evolution is considered simultaneously with time-dependent network improvement planning.

5. Concluding remarks

This study introduces the consideration of intergeneration equity into the network design problem according to the user and societal perspectives. A gap function to measure the degree of intergeneration equity is proposed. Based on this gap function, general time-dependent network design models are developed to determine the equity-based optimal tolls and improvement strategies over time. Some examples of equity-based strategies are illustrated. The results show that there are tradeoffs among the equity-based strategies, and that if the planner considers intergeneration equity as an overriding issue to be resolved, then the overall societal benefit for the planning horizon is likely to be compromised. This study also suggests ways whereby one can trade the level of equity to be attained with the overall societal benefit. In this way, some gradual measures can be introduced to the network design to strike a balance between these two objectives. Finally, this study discusses some possible extensions to the models developed within this paper. One is to acknowledge the existence of multiple solutions for the network design problem and introduce multi-objective programs to incorporate the equity consideration. The other is to capture dynamic demand responses to changes in travel costs. In this way, one can develop more accurate models to combine the consideration of dynamic demand evolution within the time-dependent framework of network improvement planning. Both are interesting extensions to this study.

Acknowledgement

This research is sponsored by the Competitive Earmarked Research Grant HKUST6154/03E from the Hong Kong Research Grant Council. We are grateful to Professor Phil Goodwin and anonymous referees for constructive comments.

References

Abadie, J., Carpentier, J., 1969. Generalization of the Wolfe reduced gradient method to the case of nonlinear constraints. In: Fletcher, R. (Ed.), Optimization. Academic Press, New York, pp. 37–47.

Boyce, D.E., Janson, B.N., 1980. A discrete transportation network design problem with combined trip distribution and assignment. Transportation Research 14B, 147–154.

Chen, M.Y., Alfa, A.S., 1991. A network design algorithm using a stochastic incremental traffic assignment approach. Transportation Science 25, 215–224.

Chiou, S.W., 2004. Bilevel programming for the continuous transport network design problem. Transportation Research 39B, 361–383.

Dargay, J.M., Goodwin, P.B., 1994. Transport Evaluation in a Disequilibrium World: Some Problems in Dynamics. In: Proceedings of the 11th Annual conference on Transport Research, Linkoping Sweden.

Dargay, J.M., Goodwin, P.B., 1995. Evaluation of consumer surplus with dynamic demand. Journal of Transport Economics and Policy XXIX (2), 179–193.

Davis, G.A., 1994. Exact local solution of the continuous network design problem via stochastic user equilibrium assignment. Transportation Research 28B, 61–75.

Friesz, T.L., Anandalingam, G., Mehta, N.J., Nam, K., Shah, S.J., Tobin, R.L., 1993. The multiobjective equilibrium network design problem revisited: a simulated annealing approach. European Journal of Operational Research 65, 44–57.

LeBlanc, L.J., 1975. An algorithm for discrete network design problem. Transportation Science 9, 183–199.

Lo, H., Szeto, W.Y., 2003. Time-dependent transport network design: a study of budget sensitivity. Journal of the Eastern Asia Society for Transportation Studies 5, 1124–1139.

Lo, H., Szeto, W.Y., 2004. Planning transport network improvement over time. In: Lee, D.H. (Ed.), Urban and Regional Transportation Modeling: Essays in Honor of David Boyce. Edward Elgar, pp. 157–176 (Chapter 9).

Lo, H., Szeto, W.Y., 2005. Time-dependent transport network design under cost-Recovery. Transportation Research B, accepted.

Marcotte, P., 1986. Network design problem with congestion effects: a case of bilevel programming. Mathematical Programming 34, 142–162.

Meng, Q., Yang, H., Bell, M.G.H., 2001. An equivalent continuously differentiable model and a locally convergent algorithm for the continuous network design problem. Transportation Research 35B, 83–105.

Miettinen, K.M., 1999. Nonlinear Multiobjective Optimization. Kluwer Academic Publishers, Massachusetts, US.

Szeto, W.Y., Lo, H., 2003. Network improvements strategies across time. In: Proceedings of the 8th meeting of Hong Kong Society for Transportation Studies, pp. 161–170.

Szeto, W.Y., Lo, H., 2004. Strategies for road network design over time: Robustness under uncertainty. Transportmetrica 1, 47–63.

Wardrop, J., 1952. Some theoretical aspects of road traffic research. In: Proceedings of the Institute of Civil Engineers, Part II, pp. 325–378.