Fairness, efficiency and the simultaneity of pricing and infrastructure capacity choice

Andreas Kopp 1*

1 OECD/ECMT Transport Research Centre
2, rue André Pascal
75016 Paris
FRANCE

Abstract

The primary objective of infrastructure pricing in normative economics and policy discussions is economic efficiency. This focus has led to the proposal that charges for infrastructure use should be based on all internal and external marginal costs associated with the use of infrastructure services. Distributional considerations, of the “fairness” of infrastructure pricing often played a supplementary role to help the acceptance of infrastructure charging.

This paper sets out a simple framework for a quasi-market for infrastructure services with the perspective of simultaneously determining efficient prices and levels of infrastructure investment. It is shown that, depending on the intensity of infrastructure use, revenues generated by efficient prices do not in all cases cover the full costs of the services. Efficient cost recovery requires an additional fixed charge. Such a combination of a fixed charge and an efficient price per unit of service implies a distributional conflict if users differ substantially in their demand for infrastructure services. It is shown that methods to allocate fixed costs resolve this conflict applying standard norms of distributional justice and being compatible with a bargaining equilibrium among heterogeneous infrastructure users.

Keywords: Infrastructure pricing; Infrastructure investment; Cost recovery; Fairness.

1. Introduction

This paper discusses the conflict between allocative efficiency and “fairness” that arises from an optimal decentralized provision of infrastructure services. Pricing of infrastructure services and the notion of “fairness” is narrower than in current policy discussions (Commission of the European Communities, 1998). The paper starts out by focussing on the problem of efficiently providing infrastructure services for high levels of congestion and a homogeneous population of prospective users (Starrett, 1988, for

* Corresponding author: Andreas Kopp (Andreas.Kopp@oecd.org.)
example). It will be shown that with high levels of congestion, optimal prices for infrastructure services cover full costs. Congestion costs are represented as a disutility due to crowding. This contrasts with the standard literature on pricing and infrastructure investment, where congestion costs are included in a generalised cost function, making strong assumptions about its functional form (Mohring and Harwitz, 1962; Small, 1992; see the review in Verhoef, forthcoming). In contrast, with no or relatively low levels of congestion, optimal prices imply deficits in the provision of these services.

To pin down ideas on fairness, prominent principles of fairness will be discussed (for a review, see Moulin, 2003). One of these principles, the reward principle, will be argued to be the most important fairness principle in the discussion of the distributional effects of pricing rules. Applying the fairness notion of the reward principle, no conflict of compatibility between allocative efficiency and distributive justice arises in the benchmark case with strong congestion and optimal marginal cost prices. A genuine distributional conflict results in the cases of relatively low levels of congestion, with the implication of the non-coverage of fixed costs by the revenues from efficient prices, and heterogeneous users with respect to their demands for infrastructure services.

The paper concludes with a characterization of the allocation of fixed costs that satisfies widely supported axioms of fairness (Mirman and Taubman, 1982), and which can be interpreted as the outcome of an n-person cooperative bargaining game (Harsanyi, 1979) as well as an application of the Rawlsian theory of justice (Rawls, 1988).

2. The basic framework

In accordance with the public finance characterization of transport infrastructure goods, we focus on the fact that transport infrastructure in general and highways in particular have high fixed costs and relatively low costs that vary with the level of usage. Moreover, due to the indivisibility of infrastructure goods there is an under-utilization of the good and crowding at high levels of usage. Considering construction, maintenance and congestion costs due to crowding, the collection of users is confronted with decreasing average costs for low levels of usage due to the dominance of the fixed costs and with increasing costs per user for high levels of usage due to the dominance of the congestion costs. Due to the invisibility there is also a very low elasticity of substitution with other infrastructure objects, if this exists at all. These characteristics imply that private markets do not in general lead to optimal allocations, or that government interventions have the potential to lead to a higher degree of welfare for the collective of users. In implementing reforms of the provision of transport infrastructure services, income distribution effects seem to have had at least as strong a resonance politically as arguments concerning the efficiency of the transport sector. One of the reasons why infrastructure pricing is perceived to be unjust lies in the disconnection between the provision of services and payment brought about by tax financing. Some of the users seem to have interpreted this as receiving a free good.

To set out the analytical framework for the analysis of the distribution effects of highway pricing, we start with a very simple framework where income distribution is not an issue. A population of n users all have the same preferences for the consumption of infrastructure services and a private consumption good.
We start by defining the following variables:

c ≡ consumption good (actual consumption minus initial endowment), the consumption good is perfectly divisible

η ≡ individual use of the transport infrastructure good

n ≡ number of individuals, treated as a continuous variable

U ≡ utility of individual (all equal)

Γ ≡ costs of infrastructure provision

n*η ≡ G, total use of the public consumption good

The identity of the preferences is expressed by assuming that all users have the same utility function.

We disregard the production sector: All individuals are equipped with a certain amount of consumption goods and decide on how much of private consumption they would like to give up for using the infrastructure. This disregard of the production sector for the private good means that we implicitly assume that the private sector is perfectly competitive.\(^1\) Focussing on the allocation aspects in this section, all individuals are equal in terms of the initial endowment with the private consumption good \(c\).

2.1 Utility

Given that individuals are assumed to be homogeneous, all have the same utility function:

\[ U = U(c, \eta, n) = U(c, \eta, G) \]

\[ U_c \geq 0, U_\eta \geq 0, U_G \leq 0 \]  \hspace{1cm} (1)

All consumers, or the representative consumer, have binary preference orderings which are complete, transitive and continuous. The utility function is then a continuous, real-value function. It increases digressively with the private consumption good, and with the individual use of infrastructure, which might be the number of trips or the number of kilometres travelled. \(G\) denotes the total use of the infrastructure. The higher the total use the more individuals suffer from congestion costs. That is, an increase of \(G\) reduces utility. The second row of 1 denotes first derivatives of the utility function.

---

\(^1\) Otherwise the pricing of infrastructure services might have to consider second best pricing, taking account of the degree of monopoly in the private sector.
2.2 Costs

Costs of the facility increase with its total use, the “capacity”. We assume that costs increase with the capacity. At this stage it is not necessary to restrict admissible forms of the function, i.e. decreasing, constant or increasing average costs.

\[ \Gamma = \Gamma(n\eta) = \Gamma(G) \]
\[ \Gamma_G \geq 0 \quad (2) \]

3. The planner's problem with homogeneous users

The planner seeks to maximize the utility of the individuals. As all individuals are identical, this amounts to maximizing the utility of a “representative agent”. With perfect knowledge of the preferences of the infrastructure users, he will simultaneously optimise the supply of services and cost recovery through the revenues generated by pricing. The constraint he faces is the total availability of resources. The planner cannot spend more on infrastructure than the total amount of consumption goods the individuals do not want to use for private consumption. He or she then faces the following budget constraint:

\[ n(\bar{c} - c) - \Gamma(n\eta) = 0 \quad (3) \]

To find out how much of the endowments should go into private consumption and how much should be used for infrastructure, the planner solves a constrained optimization problem. She or he maximizes individual utility under the budget constraint:

\[ \max_{c,\eta,n} L = U(c,\eta,n\eta) + \lambda[n(\bar{c} - c) - \Gamma(n\eta)] \quad (4) \]

That is, the planner determines the optimal level of consumption, the optimal number of users of the facility and the optimal number of trips.

4. Optimal solutions for pricing and capacity

First-order conditions for the optimal solution:

\[ \frac{\partial L}{\partial c} = \frac{\partial U}{\partial c} - \lambda n = 0 \quad , \]

which implies

\[ \lambda = \frac{1}{n} \frac{\partial U}{\partial c} \quad (6) \]
$\lambda$ indicates what the social availability of one more unit of the consumption good means to the welfare of all users.

$$\frac{\partial L}{\partial \eta} = \frac{\partial U}{\partial \eta} + n \frac{\partial U}{\partial G} - \lambda \frac{\partial \Gamma}{\partial G} n = 0 \tag{7}$$

The first term on the right-hand side shows the increase in utility of having one more use of the infrastructure. The second term indicates the disutility of all the other members of society doing the same, with the consequence of increasing congestion. The sum of these effects should be equal to the marginal costs of all of the (equal) individuals taking the same decision, multiplied by the factor that transforms costs in terms of the consumption good into terms of utility.

Dividing by the expression for $\lambda$ and $n$ we obtain from (7):

$$\frac{\partial U / \partial \eta}{\partial U / \partial c} + n \frac{\partial U / \partial G}{\partial U / \partial c} - \frac{\partial \Gamma}{\partial G} = 0, \text{ or} \tag{8}$$

$$\frac{\partial U / \partial \eta}{\partial U / \partial c} = \frac{\partial \Gamma}{\partial G} - n \frac{\partial U / \partial G}{\partial U / \partial c} \tag{9}$$

The absolute value of (a) is identical to the Marginal Rate of Substitution between private consumption and use of the infrastructure facility. It indicates how much of private consumption individuals are prepared to give up to have one more unit of infrastructure use in equilibrium. It is equivalent to the willingness to pay for a unit of infrastructure use and is in a well-defined sense the "efficiency price". (b) is negative and indicates an individual's utility loss due to congestion if another individual increases the use of the facility by one unit: (-n) times this expression is then what the latter should pay to compensate all the other users for the increase in congestion. The first term on the right-hand side of (9) is the marginal cost of operating the facility due to the increase in infrastructure use by one unit.

Marginal operation costs plus the compensations for the disutility of increased congestion add up to the efficiency price.
That this just covers the total costs of the facility can be seen from the planner's answer to the question of how many users should be admitted to the facility. Given that he knows already the optimal quantity of demand per user, this amounts to determining the capacity.

Differentiating (4) with respect to \( n \), we have

\[
\frac{\partial L}{\partial n} = \eta \frac{\partial U}{\partial G} - \lambda (\bar{c} - c) - \lambda \frac{\partial \Gamma}{\partial G} \eta = 0. 
\]  

(10)

Dividing by the expression for \( \lambda \) according to equation (6) we obtain

\[
n \frac{\partial U}{\partial U/c} \eta - (\bar{c} - c) - \frac{\partial \Gamma}{\partial G} \eta = 0
\]

(11)

Multiplying both sides of the equation by \( n \) leads to

\[
n \frac{\partial U}{\partial U/c} G - (\bar{c} - c)n - \frac{\partial \Gamma}{\partial G} G = 0
\]

(12)

Minus \( n(\bar{c} - c) \) is equal to \( \Gamma \). That is

\[
\Gamma = - n \frac{\partial U}{\partial U/c} (c) + \frac{\partial \Gamma}{\partial G} (d) G = 0
\]

(13)
The right-hand side of (13) shows the total payments by the users of the infrastructure. (c) is the total payment by all users for causing congestion, (d) is the total payment for the marginal operation costs of all users. The right-hand side of (13) is exactly equal to the efficiency price of using one more unit (trip, hour of driving), which is equal to the right-hand side of (9), times the total use of the infrastructure G. It also shows that if congestion is sufficiently strong, transport infrastructure can be offered like a private good. Dividing (13) by G, we have the optimality condition that the marginal congestion costs plus the marginal infrastructure costs add up to the average infrastructure costs, i.e. the costs per unit of service. This is illustrated in Figure 1, with the slope of the ray from the origin representing the average costs. As this ray is tangent to the social cost curve, it represents the social marginal costs as well. Without congestion (c) in (13) is negligible, and the optimality conditions will always be violated with the usual assumptions about the cost function of infrastructure. With constant marginal costs and fixed costs, average costs will be decreasing throughout the relevant levels of usage.

The optimality conditions can be restored by fixed transfers to the infrastructure sector. These transfers are either financed by taxes or by fixed charges. In any case, they have to be unrelated to the use of the infrastructure as well as to the characteristics of the users so as not to violate the optimality conditions. It is this required independence that leads to distributional problems in the case of the heterogeneity of agents.

5. User heterogeneity, optimality and perceptions of distributional justice

To cast the problem of distribution effects and pricing in the simplest form, we assume that two groups exist that are still identical but for their endowment with initial income. We assume that the first group has \( n_1 \) members, all equipped with an initial income of \( c_1 \), the second group has \( n_2 \) members with an income of \( c_2 \). To avoid any discussion of the comparability of utilities, we assume that all individuals have the same utility function.

The total use of the infrastructure has then to be redefined to

\[
G = n_1 \eta_1 + n_2 \eta_2, \tag{14}
\]

and the cost function to

\[
\Gamma = \Gamma(n_1 \eta_1 + n_2 \eta_2) \tag{15}
\]

The planner's problem is then changed to

\[
\max L = U(c_1, \eta_1, G) + U(c_2, \eta_2, G) + \lambda \left[n_1 (c_1 - c_1) + n_1 (c_1 - c_1) - \Gamma(n_1 \eta_1 + n_2 \eta_2)\right] \tag{16}
\]

By the same analytical steps as in the case of homogeneous users, we arrive at two optimality conditions, one for each group.
The first term on the left-hand side of both equations indicates the willingness to pay of the user group members for an additional unit of infrastructure services. Despite the differences in income, these are equal because

\[
\frac{\partial U}{\partial \eta_i} + n_i \frac{\partial U}{\partial c_1} - \frac{\partial \Gamma}{\partial G} = 0 \tag{17}
\]

and

\[
\frac{\partial U}{\partial \eta_2} + n_2 \frac{\partial U}{\partial c_2} - \frac{\partial \Gamma}{\partial G} = 0 \tag{18}
\]

The first term on the left-hand side of both equations indicates the willingness to pay of the user group members for an additional unit of infrastructure services. Despite the differences in income, these are equal because

\[
\frac{n_i}{\partial U/\partial c_1} = \frac{n_2}{\partial U/\partial c_2} = \frac{1}{\lambda} \tag{19}
\]

as follows from the optimal values for the consumption of the private good.

That is, the optimal social organization of infrastructure provision implies the pricing of individual units of infrastructure use according to the social marginal costs of the services. If congestion is strong enough that the social marginal cost of infrastructure increases with an increasing number of users and/or an increasing number of kilometres travelled, the infrastructure is self-financing.

### 6. Principles of distributive fairness

Notions of “fairness” are not, of course, universal. They refer to underlying principles which are more or less able to claim universal support. Fairness naturally, and following Aristotelian philosophy, entails the equal treatment of equals. If two persons have identical characteristics in all dimensions relevant to an allocation problem at hand, they should receive the same treatment, i.e. the same share of income, voting power or costs of a service which is commonly enjoyed.

The unequal treatment of unequals is, in contrast, a vague concept, which is open to interpretation. That is, the difficulties with the notion of “unequal treatment of unequals” result from the heterogeneity of the population that the fair distribution of surplus or costs is designed for. Four different principles are central to the discussion:

1. exogenous rights,
2. fitness,
3. compensation,
4. reward.

These principles will be briefly discussed to argue that the potential conflict with the Pareto principle -- that allocations should be such that no reallocation of resources could improve the position of one party without worsening the position of another -- implies
that only the reward principle is important for the discussion of infrastructure pricing and distributive fairness.

The notion of fairness concerning exogenous rights is independent of the consumption of the relevant resources or the responsibility of the consumers in their production. A paramount instance of exogenous rights is the fairness principle of equality in the allocation of certain basic rights such as political rights, the freedom of speech, etc. The right to vote, for example, is equal for all voters regardless of their desire to vote or the rationality of the voters. Equal exogenous rights postulate equality ex ante in the sense of an equal claim to resources, regardless of the way they affect our welfare and that of others. For the provision of infrastructure services, this would imply an equal (and free) access to infrastructure that is financed by a lump sum tax, regardless of the endowments of the user or differences in individual demand.

The fitness principle postulates that resources must go to whoever potentially makes the best use of them. The fitness principle justifies an unequal allocation of resources, independently of needs, merit or rights.

Both the exogenous rights and the fitness principle are in sharp contrast to the compensation principle. The compensation principle is based on the idea that certain differences in individual characteristics are involuntary, morally unjustified and affect the distribution of a higher order characteristic that is to be equalized. This justifies unequal shares of resources in order to compensate for the involuntary difference in the primary characteristics. The compensation principle aims at an equal degree of satisfaction of consumers' needs ex post. For the consumption of infrastructure services, equality according to the compensation principle would require an equal sacrifice in utility terms for all users of the transport system. The compensation principle is relevant to the discussion of the fairness of pricing rules only to the extent that fiscal redistribution mechanisms are unable to correct a socially undesirable distribution of incomes or abilities. The unequal distribution of characteristics which induce undesired inequalities of higher order characteristics should focus on the correction of the unequal distribution of primary characteristics. More specifically, if the income distribution of a society is perceived to be too unequal, the unequal income distribution should be corrected by fiscal measures and not the consequent unequal distribution of opportunities to travel.

The most important principle of fairness for the discussion of infrastructure pricing roles and distributional fairness is the reward principle. According to the reward principle, individual characteristics are morally relevant when they are viewed as voluntary and consumers are held responsible for them. They justify unequal treatment. Due to the fact that individual demands do not lead to variations in total costs of infrastructure but might reduce the per capita costs for all consumers, the application of the reward principle is not straightforward.

7. Distributional conflict of fixed fee and optimal pricing

If infrastructure users are unequal, a distributional conflict might be introduced as a result of different demands by the users of the infrastructure. More specifically, the different interests of unequal users may manifest themselves in differences in the preferred pricing rule. Assume that the users have a choice between different pricing
rules to cover the full costs of infrastructure. A first option could be to postulate that the per-km user charges should cover the full costs of the infrastructure. The optimisation of the social planner, set out in section (2), would then have to be extended by a restriction that prices have to cover the costs of the infrastructure. Such an optimisation exercise would lead to a Ramsey price of \( p_0 \). The consumers might be offered the choice of paying a price \((p_0 - t)\) which is lower by the amount \( t \). The alternative expenditure functions would then be (cf. Willig 1978)

\[
E(p_0, n^*) = p_0 \eta, \text{ and}
\]

\[
E(p, n^*) = \gamma + (p_0 - t) \eta
\]

The user will prefer the pricing rule that will imply the higher indirect utility, denoted by \( V \). She or he would prefer a two-part tariff to a Ramsey price if

\[
V(p_0 - t, \bar{c} - \gamma) \geq V(p_0, \bar{c})
\]

For small \( t \), starting from full cost prices, the consumer prefers a two-part tariff if

\[
\left. \frac{dV}{dt} \right|_{t=0} = -\frac{\partial V}{\partial p} \frac{\partial V}{\partial \bar{c}} \gamma = \frac{\partial V}{\partial \bar{c}} (\eta - \gamma) > 0
\]

As the marginal indirect utility with respect to real income is always positive, the preference for a two-part tariff follows from \((\eta - \gamma)\) being positive. The higher the equilibrium demand for infrastructure services, the more the user will prefer a two-part tariff. The smaller the equilibrium demand, the more they will prefer Ramsey pricing. To implement full cost pricing to satisfy the political demands of the low demand group would lead to an aggregate welfare loss. If marginal costs were zero, as assumed in Figure 2, the triangle BCD would represent the loss of consumer rent which would result from full cost pricing.
8. Solution of the distributional conflict

The solution of the conflict between equity and efficiency proposed here relies on the well-known model of allocating the costs of a jointly used resource according to the cooperative game theory concept of the Shapley value (Shapley, 1953, Shubik, 1962). In terms of the general principles stated in section (6), the solution concept rests almost entirely on the reward principle, posing the question of a fair level of contribution to the joint costs in order to deserve the service enjoyed in equilibrium.

With marginal cost pricing, the cost allocation game is about access charges for different users whose demands add in different ways to the total costs of the infrastructure. This is a standard mechanism for joint fixed cost allocation mechanisms (Young, 1994; Sharkey, 1995). They have also been applied in transport economics, as a solution to charging in cases where there are only fixed costs and price inelastic demand (Littlechild and Thomson, 1977). The cost allocation method has even been proposed as a way to calculate full cost prices for infrastructure use (Doll, 2005). However, it has not been discussed as a solution to the distributional problem of two-part tariffs. To introduce the idea of the solution concept, consider the following example. There are three classes of vehicles A, B and D requiring different infrastructure designs, leading to different fixed, annual stand-alone costs:

\[ C(i) = 60, \text{ for } i = A, B, D \]  

(24)
The capital C indicates fixed annual costs for the different coalitions, assuming technical efficiency. To compute a fair allocation, a generalised principle of marginalism is applied. Adding, for example, vehicle class B to A, or D to A, or both B and D to A leads to additional demands for the infrastructure, implying additional costs of 60.

\[ C(A) = C(AB) - C(A) = C(AD) - C(A) = C(ABD) - C(A) = 60 \]  

(28)

The solution mechanism now orders the vehicle classes to randomly identify the expected additional fixed costs, for which the individual vehicle classes are responsible. For the coalition formation process B, A, D we obtain, for example, the following values \( x_i \), with \( i = A, B, D \):

\[ x_B = C(B) = 60, \quad x_A = C(AB) - C(B) = 60, \quad x_D = C(ABD) - C(A) = 60 \]  

(29)

This process is repeated for all possible sequences to form the “coalition” of all vehicle classes. This leads to the following orderings and additional fixed costs:

<table>
<thead>
<tr>
<th>Ordering</th>
<th>Class A</th>
<th>Class B</th>
<th>Class D</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABD</td>
<td>60</td>
<td>60</td>
<td>0</td>
</tr>
<tr>
<td>ADB</td>
<td>60</td>
<td>60</td>
<td>0</td>
</tr>
<tr>
<td>BAD</td>
<td>60</td>
<td>60</td>
<td>0</td>
</tr>
<tr>
<td>BDA</td>
<td>60</td>
<td>60</td>
<td>0</td>
</tr>
<tr>
<td>DAB</td>
<td>0</td>
<td>60</td>
<td>60</td>
</tr>
<tr>
<td>DBA</td>
<td>60</td>
<td>0</td>
<td>60</td>
</tr>
<tr>
<td><strong>Sharpley value</strong></td>
<td>50</td>
<td>50</td>
<td>20</td>
</tr>
</tbody>
</table>

That is, the fair allocation as defined by the Shapley value assigns the average of the marginal contributions to total costs in the process of the formation of the all-player coalition to the individual vehicle types. This model of a random formation of the all-player coalition mimics, in a sense, the notion of the Rawlsian theory of justice that fairness considerations are based on the expectation that the individual might end up in a socially disadvantaged position. Furthermore, it has been shown by Harsanyi that the cost allocation solution presented above generalizes the two persons bargaining game of Nash to an arbitrary number of players. That is, the notion of fairness presented here does not depend on a “public interest” view of politics, where a benevolent dictator decides on the assignment of costs following a universally accepted principle of fairness. Rather, the solution concept can be interpreted as the anticipated outcome of a bargaining process between all parties involved.

An often-raised counterargument against the Shapley value is the high level of information requirements, either for planners or bargaining partners. In the specific

\[ C(AB) = C(AD) = 120 \]  

\[ C(BD) = 60 \]  

\[ C(ABD) = 120 \]  

(25) (26) (27)
In the context of infrastructure provision there is, however, a way of identifying types of consumers by vehicle types. In many countries, an approximate solution could be implemented by designing or re-designing a vehicle tax according to the presented cost allocation mechanism.

9. Conclusion

The paper has discussed the conflict between efficient pricing and fairness. It has been shown that the conflict arises in cases where marginal congestion costs are too low to allow for coverage of full costs by marginal cost pricing. Conventional solutions to solve the distributional problem would entail efficiency losses. The paper proposes an allocation mechanism for the fixed costs that would resolve the conflict between efficiency and distributional equity.

References


