Institute of Transport Economics
Norwegian Centre for Transport Research

Transport cost benefit analysis

Basic assumptions and accounting rules

Harald Minken

1934/2023

Title:	Transport cost benefit analysis. Basic assumptions and accounting rules
Tittel:	Nyttekostnadsanalyser i transport. Grunnleggende forutsetninger og regneregler
Author:	Harald Minken
Date:	02.2023
TØI Report:	1934/2023
Pages:	189
ISSN Electronic:	2535-5104
ISBN Electronic:	978-82-480-1991-6
Project Number:	4330
Funded by:	TØI
Project:	4330 – Samfunnsøkonomiske metoder
Project Manager:	Askill Harkjerr Halse
Quality Manager:	Askill Harkjerr Halse
Research Area:	Economic Analysis
Keywords:	Economic efficiency, Sustainability, Conceptual models, Discount rate

Summary

This report is a collection of papers on the methods of transport project appraisal and transport cost benefit analysis. The first paper defends the normative use of transport economics provided equity issues are addressed. What follows are derivations of transport cost benefit rules and proposals concerning the discount rate, the treatment of sustainability, and project selection. The next papers treat specific issues such as industrial reorganisation benefits, freight values of time, scheduled transport costs, how congestion pricing affects the cost of capital, and how to find the benefit cost ratio of a marginal increase in the maintenance budget. Finally, environmental input output analysis is proposed as a general approach to take account of impacts on sectors or areas outside the transport model.

Items 1, 7 and 8 are published papers. The others are conference papers, excerpts from project deliverables or texts never published at all. Most of them require some knowledge of mathematics.

Kort sammendrag

Denne rapporten er en samling artikler om metodespørsmål i samfunnsøkonomisk analyse av transporttiltak. Den første artikkelen er et forsvar for en normativ bruk av transportøkonomien, dvs. å bruke den til å si hva som bør gjøres eller bygges. Denne bruken forutsetter imidlertid at fordelingsspørsmål og rettferdighetsspørsmål blir behandlet og tatt hensyn til separat. Deretter følger artikler som utleder regler for nyttekostnads-analyser i transportsektoren og om kalkulasjonsrenta, hvordan vi kan ta hensyn til bærekraftighet og valg av prosjekter. Så tar vi opp noen mer spesielle emner, slik som virkninger på omorganiseringer i industrien, tidsverdier for godstransport, kostnader i rutegående transport, hvordan køprising påvirker skyggeprisen på investeringene, og hvordan man kan beregne nyttekostnadsbrøken av en liten økning av vedlikeholdsbudsjettet. Til slutt følger en skisse til hvordan en kan beregne virkninger på sektorer eller områder som faller utafor området som transportmodellen dekker. De fleste artiklene krever matematikkunnskaper.

Preface

This report is a collection of papers by Harald Minken on the methods of transport project appraisal and transport cost benefit analysis. Selected published papers are supplemented by conference papers, excerpts from TOI reports and project task reports, and even old manuscripts never published before in any form.

The language is English throughout. However, a TOI report in Norwegian with papers and manuscripts on the same subject, TØI-rapport 1936/2022, is published simultaneously. None of them are translations of the papers in English published here, and none of the papers and reports in Norwegian have been translated into English.

The selection of content for both reports was made by the author. At his retirement the author wants to thank everybody at TOI for their friendship, support and help through nearly 30 years. In particular, he thanks Farideh Ramjerdi, Lasse Fridstrøm and the former TOI library for teaching him transport economics, and Olav Eidhammer and Kjell Werner Johansen, who followed each other as leaders of the institute's Department of Transport Economics, for providing room for him to practice what he had learnt, each in their own way.

Other works by Minken or other TOI researchers can be found at https://www.toi.no/publikasjoner/. (Toggle between Norwegian and English in the top right corner of the window.)

Oslo, February 2023
Institute of Transport Economics

Bjørne Grimsrud
Managing Director

Kjell W. Johansen
Director of Research

Contents

Institute of Transport Economics
Norwegian Centre for Transport Research

ENGLISH

Summary

Transport cost benefit analysis

Basic assumptions and accounting rules

TØI Report 1934/2023 • Author: Harald Minken • Oslo 2023 • 189 pages

This report is a collection of papers on the methods of transport project appraisal and transport cost benefit analysis. The first paper defends the normative use of transport economics provided equity issues are addressed. What follows are derivations of transport cost benefit rules and proposals concerning the discount rate, the treatment of sustainability, and project selection. The next papers treat specific issues such as industrial reorganisation benefits, freight values of time, scheduled transport costs, how congestion pricing affects the cost of capital, and how to find the benefit cost ratio of a marginal increase in the maintenance budget. Finally, environmental input output analysis is proposed as a general approach to take account of impacts on sectors or areas outside the transport model.

Items 1, 7 and 8 are published papers. The others are conference papers, excerpts from project deliverables or texts never published at all. Most of them require some knowledge of mathematics.

Transportøkonomisk institutt
Stiftelsen Norsk senter for samferdselsforskning

NORSK

Sammendrag

Nyttekostnadsanalyser i transport

Grunnleggende forutsetninger og regneregler

TØI rapport 1934/2023 • Forfatter: Harald Minken • Oslo, 2023 • 189 sider

Denne rapporten er en samling artikler om metodespørsmål i samfunnsøkonomisk analyse av transporttiltak. Den første artikkelen er et forsvar for en normativ bruk av transportøkonomien, dvs. å bruke den til å si hva som bør gjøres eller bygges. Denne bruken forutsetter imidlertid at fordelingsspørsmål og rettferdighetsspørsmål blir behandlet og tatt hensyn til separat.

Deretter følger først en artikkel som utleder regler for nyttekostnadsanalyser i transportsektoren, og deretter tre artikler om kalkulasjonsrenta, hvordan vi kan ta hensyn til bærekraftighet, og hvordan vi skal velge ut prosjekter. Så tar vi opp noen mer spesielle emner, slik som virkninger på omorganiseringer i industrien, tidsverdier for godstransport, kostnaden ved å drive transport i rute, hvordan køprising påvirker skyggeprisen på prosjektene som finansieres med inntektene av køprisingen, og hvordan man kan beregne nyttekostnadsbrøken av en liten økning av vedlikeholds-budsjettet. Til slutt følger en skisse til hvordan en kan beregne virkninger på sektorer eller områder som faller utafor området som transportmodellen dekker. De fleste artiklene krever matematikkunnskaper.

1 Introduction

A common feature of all the texts in this report (and in an accompanying report in Norwegian) is that they derive rules of some sort from simple mathematical models. They may be rules that describe the consequences of government actions and changes in external factors. But clearly, the intention behind these models is for the most part prescriptive: They should be taken as rules that tell government what to do in the situations described. Explicitly or implicitly, the objective is to maximise welfare in the short run, or possibly in the very long run (sustainability).

The basic element of the mathematical models used is a demand function, showing aggregate demand to be a function of generalised cost. Many questions can legitimately can raised about this. Is it generally possible to add time use and costs to a generalised cost? Is it generally possible to form an aggregate demand function and assume that it behaves in the same way as the individual consumer does? And if it does, may we legitimately draw conclusions from observed regularities to what we as regulators should do?

Generally, many economists answer these questions in the negative. At the same time, most economists embrace the axiomatic foundations of consumer theory and continues to use it not only as a theory of the behaviour of the single consumer, but also as a guide to the behavior of aggregate demand. And most transport economists make some use of the concept of generalised cost. This author agrees with them. It seems that aggregate demand in transport markets generally may be modelled fairly accurately using such simplifications. And if so, it must make perfectly good sense to use welfare functions and concepts derived from consumer theory, like the consumer surplus, to guide transport policy. We may indeed make the step from describing transport markets to making transport policy advises.

We are not the first to derive policy guidelines by mathematical reasoning. “Ethica ordine geometrico demonstrata”, or ethics derived by mathematical reasoning, was the title of the magnum opus of the philosopher Spinoza. Who says it cannot be done?

There are however theoretical restrictions. In the first chapter, *“The Pareto criterion and the Kaldor Hicks criterion”*, we show that for the transport users to be added together to a representative consumer, their utility functions must have a certain structure. When this is not the case, we can usually make do with the average of all consumers, the average consumer. Since she does not behave like the representative consumer, equity issues will be a concern. We may therefore not always follow the advice of the average consumer, at least not unless equity issues are addressed at the same time. Explicitly or implicitly, then, this is supposed in the remaining chapters.

With this qualification, the policy recommendations of the remaining chapters are thought to be valid. The advice of the average consumer is thought to be worth following. Not always, but as a rule.

The report consists of 13 papers

“Transport cost benefit rules” is a revised version of a paper originally presented to the NTF Workshop on the marginal cost of public funds, held in Copenhagen on October 25, 2005.

The main purpose of the paper is to derive rules for entering investment costs in transport cost benefit analyses.

"A note on the effects of the marginal cost of funds and the indirect tax correction factor" studies the wider economy impacts of using non-transport taxes in transport, and the impacts of transport policy on non-transport tax revenue.

The paper "Systematic risk, and how it is taken account of in the discount rate of Norwegian transport cost benefit analyses" derives an explicit formula for the risk premium of the discount rate used to compute the present value of future costs and benefits and applies it to projects in the transport sector. The paper was presented to an ECTRI working group meeting in Madrid in 2008. It must be said that the approach was abandoned four years later, when the Ministry of Finance introduced a simplified approach applicable to all sectors. Nevertheless, the approach of the paper met some interest internationally.

"Appraising the sustainability of urban land use and transport strategies" outlines the approach to urban land use and transport planning that was used in a string of European research projects led by ITS Leeds and with TOI as a partner. One of these was the Prospects project, where TOI was in charge of one of the main deliverables, the methodological guidebook.

The next paper, *"A finite horizon model of an exhaustible resource"*, treats the approach taken to assess long term sustainability from transport model runs of the situation in the not too distant future.

"Project selection with sets of mutually exclusive alternatives" was published in Economics of Transportation in 2016. It treats the actually very common situation where projects exist in alternative variants, possibly with very different costs, and a national or local transport plan is to be formed by choosing under a budget constraint – choosing, that is, not only projects but also their exact variant. The simple iterative procedure for doing so should really become standard for urban and national transport plans.

"Industrial reorganisation benefits revisited" was published in Journal of Transport Economics and Policy in 2014. It shows that, given the free market conditions that is assumed in most discussions about the potential industrial reorganisation benefits of transport improvements, there are too many firms and too little transport. Preferably, we should eliminate the inefficiency rather than counteract it with transport infrastructure building. A uniform price of the commodity at all locations might do the trick.

The unpublished papers *"A theory of freight values of time and reliability"* and *"Scheduled transport cost functions"* are examples of the axiomatic approach to transport policy mentioned above.

The short note *"Congestion pricing should affect the cost of capital"* was written in 2008 for the EU-funded project ENACT. It shows that financing infrastructure building with the revenue from congestion pricing might cause problems in the long run, as adjusting the congestion price to the reduced congestion makes it more difficult to repay loans.

"Assessing the benefit cost ratio of a marginal increase in the maintenance budget", provides a tool to increase the economic efficiency of maintenance plans. It may also make it possible to introduce maintenance plans as alternatives to infrastructure projects in urban transport plans.

The final paper, *"Environmental input output analysis and ecological footprints"* was written for the Prospects project as a way of following up on the hitherto mostly unquantified environmental consequences of transport plans in a systematic way. It was not followed up in Prospects, but similar approaches are probably in use elsewhere.

2 The papers

2.1 The Pareto Criterion and the Kaldor Hicks Criterion

The Pareto Criterion and the Kaldor Hicks Criterion[1]

Harald Minken

Institute of Transport Economics

Oslo, Norway

ABSTRACT

We define and discuss the Pareto and the Kaldor Hicks criteria, and show how, if we care about distributional issues or the feasibility of compensating the losers, they both have a role to play in transport cost benefit analysis.

To the extent that travelers in our models are too diverse to be represented by a representative consumer, their consumer surpluses cannot be added without strong ethical assumptions, and so the Kaldor Hicks criterion cannot legitimately be applied. We mention some ways to evade this problem, including the widely used "average consumer" approach or reverting to voting theory.

Keywords:

Cardinal/ordinal, Compensating variation, Consumer surplus, Congestion charging, Cost benefit analysis, Equivalent variation, Pareto improvement, Potential Pareto improvement, Representative consumer, Value of time, Voting theory, Welfare function, Willingness to pay, Utilitarian

---

[1] This article was published in: Vickerman, Roger (eds.) International Encyclopedia of Transportation, vol. 1, pp. 190-194.

Contents

The Pareto criterion and the Kaldor Hicks criterion

Definitions

These criteria will apply to a society (usually taken to be a country) consisting of a given number of individuals. Different policies or options are open to the society. Given the policy chosen, a given amount of goods and services will be available at the aggregate level. Either by the action of the individuals themselves or government, or by the interaction of both, each individual will get a certain share of it. An *allocation* is a distribution of the total amount of goods and services to each of the individual members of society. Each policy results in an allocation.

A *reallocation* is a measure taken to change the allocation after a policy has been implemented.

This highly abstract view of society abstracts from time (time only exists in the form of before and after a policy is implemented), and it says nothing about the mechanisms that bring about a certain allocation, be it production, trade, or the individuals' own choices of how much to work, what to consume and how to spend their time in general. In particular, utility maximization and profit maximization have not been assumed.

The individuals are however thought to be able in all circumstances to tell if a bundle of goods and services is better, worse or just as good as another. That is, they have a *complete and transitive preference ordering* over such bundles.

Definitions concerning the Pareto criterion

The Pareto criterion provides the following sufficient condition for an allocation to be better than another:

- An allocation is better than another (constitutes an improvement from the base case allocation) if at least one member of society is better off and nobody is worse off.

If this is the case, the allocation is called a *Pareto improvement* and is deemed worth carrying out. Note that if only one single person is worse off, the allocation is not a Pareto improvement, even if all others are overjoyed by it. In such situations – certainly the majority of cases – the Pareto criterion cannot decide whether or not a policy should be implemented.

When no Pareto improvements are possible, i.e., nobody can be made better off without making someone else worse off, we have a *Pareto maximum*.

Definitions concerning the Kaldor-Hicks criterion

A *potential Pareto improvement* is a situation where those that are made better off by the policy measure, can in principle compensate those who are made worse off (so that they are at least as well off as before), and still be better off themselves. Put otherwise: In the situation after the policy has been implemented, there is at least one reallocation that can make at least one individual better off, while making no one worse off than before the policy was implemented.

The Kaldor-Hicks criterion simply says

- A policy which brings about a potential Pareto improvement is worth carrying out.

We note that the concept of economic efficiency as defined and computed in transport cost benefit analysis, is essentially the same as to satisfy the Kaldor-Hicks criterion.

The Pareto criterion

History

To the early utilitarianists, like Jeremy Bentham and John Stuart Mills, it seemed obvious that the best action was the action that produces the greatest well-being to the greatest number of people, and that somehow, the level of attainment of this goal was measurable.

The Italian economist and sociologist Vilfredo Pareto (1848-1923) thought otherwise. To him, the utility of one individual and that of another could never legitimately be added or compared. Every person is indeed able to compare, rank and choose among different outcomes for herself, but interpersonal comparisons make no sense.

So how can we pass judgment on policies that affect more than one person? Pareto came up with a criterion that does not require more than that each affected individual is able to rank any set of policies presented to her (her preferences induces a complete and transitive ordering on the set of policy options), and that she prefers more rather than less of any good. This is the Pareto criterion, and it may be shown that it is indeed the strongest possible criterion that does not involve interpersonal comparisons.

Utility may be measured on an ordinal or cardinal scale. Measurements are on an ordinal scale if they are used for ranking purposes only. Cardinal measures on the other hand are based on an underlying scale which makes it meaningful to do computations. Your utility is twice the size of mine, the difference between yours and mine is 23, etc.

Pareto's concept of utility was ordinal – utility differences exists, but the level of utility, the size of utility differences and the concept of marginal utility have no intrinsic meaning. Consequently, interpersonal comparisons are also meaningless. Other important economists, such as Hicks (1939), Samuelson (1956) and Sen (2018), came to exactly the opposite conclusion: Even if a cardinal and measurable concept of utility is not needed to derive the demand functions of individuals, interpersonal comparisons not only make sense, they are also necessary for policy purposes. If a cardinal utility concept is needed for such purposes, so be it.

Both approaches have lived on, but it is fair to say that the cardinal approach has lost much ground in academic circles, while the ordinal approach has become more and more formalized. Perhaps to the point where, as K.J. Lancaster once remarked, it now stands as an example of how to extract a minimum of results from a minimum of assumptions. In the applied fields of economics, however, and perhaps more than anywhere else in transport economics, the cardinal approach has spread and developed over the last half century in the form of cost benefit analysis.

Transport applications of the Pareto criterion

The Pareto principle is seldom applied in transport economics. The main reason is that major transport infrastructure projects are nearly always financed by taxes. Among the taxpayers, there will usually be many that will make no use of the new infrastructure, and so there is a large group of losers that cannot possibly be compensated without undermining the financial basis for the whole project. The possibility of applying the Pareto principle might be a little better if we consider a more comprehensive urban or national transport plan, consisting not only of road projects, but also projects concerning public transport, walking and cycling. But even then, there will be many people who stand to lose.

Research on the political processes that produces such transport plans have shown that there is a tendency to distribute projects among districts in such a way that those left out of the current plan will be first in line for projects in the next plan. In the end, every taxpayer will get something back, it is

thought (and sometimes even openly said or formally agreed). Such processes are often accused of wasting taxpayers' money by including a lot of economically inefficient projects. Could it possibly be, however, that in the best of cases, they do at least produce a Pareto improvement? If the best plan according to a cost-benefit analysis (the Kaldor-Hicks criterion) is infeasible because the losers understand that compensation could not possibly be paid (or at least will not actually be paid) there might still be some room for rational bargaining for a Pareto improvement.

If we turn from infrastructure plans to pricing and regulation, the potential for actual Pareto improvements increases considerably. More efficient ways of regulating traffic, providing information to users, and utilizing existing infrastructure and rolling stock may cost little and (in the best of cases) impose no extra inconveniences on anybody. Manuals with ideas for such "small projects" exist.

Congestion charging and other forms of marginal cost pricing split the users into three groups: The first group opts out, and will incur a loss the size of which depends on the best remaining alternative. The second group also stands to lose but finds that paying the price and staying in is a lesser evil than the other remaining alternatives. Members of the third group are the only ones who find themselves better off in the new situation, for instance because they have a very high willingness to pay for time savings on the road, for getting a seat on the bus, etc.

A fourth group – the second winning group – is the public sector, who gets the revenue from the charges. It is almost a matter of definition of marginal cost pricing that this last group could reimburse the three others and still have something left. (It will probably not be able to save some money by identifying the members of the third group.) But to do so would be meaningless: If all travelers understand that they will get their money back regardless, they will not change their travel habits. The potential Pareto improvement in this case actually could not be made a reality, at least not in the straightforward way by reimbursing the tolls. The compensation must either be made in kind, for instance by improving public transport, or paid out with an equal amount to all inhabitants in the district. Either way, there will still be losers, but not so many, and not by so much.

We see that the Pareto principle has a role to play in transport as a complement to the Kaldor-Hicks criterion. There is a moral reason for this: We feel obliged to reduce the number of losers from a policy and the severity of their loss. Potentiality becomes a hollow phrase if it cannot ever be transformed to reality. And there is a practical side: All our congestion charging plans tend to be turned down if we do not pay attention to compensating the losers.

The Kaldor-Hicks criterion

Theoretical problems

To apply the Kaldor Hicks criterion, we must be able to express the losses of some and the gains of others on a common scale, preferably money. According to microeconomic consumer theory dating back to John Hicks (1939), this can be done in two different ways: the equivalent variation (EV), and the compensating variation (CV). The first is the answer to the question of how much money one would be willing to pay to get the benefits of a proposed policy, and the second answers the question of how much one must have to forego these benefits.

There is also a third way, commonly associated with Alfred Marshall, but actually dating back to the French engineer Jacques Dupuit (1844). This is the consumer surplus (CS), consisting of the area under the ordinary demand curve for the good that the policy provides. This measure lies in between the two Hicksian measures, both of which coincide with the CS under special circumstances.

It turns out that there are theoretical problems with each of these. For instance, it could be shown that using EV or CV, it might sometimes happen that starting from the "do nothing" situation, doing "something" is judged to be an improvement, while starting from the situation where this "something"

has already been implemented, going back to the original situation is also judged to be an improvement (Scitovsky reversals). The CS, on the other hand, lacks the straightforward interpretation and general applicability that EV and CV have.

In the fifties, it was shown by W.F. Gorman (1953) that for consistent interpersonal comparisons to be made, and consequently for computing the aggregate losses of the losers and the aggregate gains of the winners and compare them with one another, the utility functions of all individuals must be of a particular mathematical form, the so-called Gorman Polar Form.

Let **p** be the vector of prices of the goods and services that concern us here, and let R_h be the income of individual *h*. The indirect utility function $V_h(\mathbf{p}, R_h)$ of individual *h* is of the Gorman Polar Form if and only if it takes the form

$$V_h(\mathbf{p}, R_h) = a_h(\mathbf{p}) + b(\mathbf{p})R_h,$$

The function a_h (**p**) is specific for individual *h*, while the function b(**p**) is the same for all individuals. Summing over individuals, we get the (utilitarian) welfare function:

$$V(\mathbf{p}, R) = \sum_h a_h(\mathbf{p}) + b(\mathbf{p})R$$

where *R* is the sum of all individual incomes. It is seen that this welfare function has the same form as the individual indirect utilities. That the utility functions of all are of the Gorman Polar Form is not only the condition for consistent interpersonal comparisons to be made, it is also the condition for there to be a representative consumer, i.e., for the aggregate decisions of all consumers to correspond to consumer theory applied at the aggregate level.

It is evident from the formula that welfare, as measured in this way, stays the same regardless of changes in the income distribution. Another property of this function is that even if the individuals differ in their income and tastes, they are all equal in the sense that if they get a marginal increase in income, they will all spend it in the same way. (The derivative of the demand functions with respect to income are the same for all individuals.)

It has been shown by Chipman and Moore (1979, 1994) that to apply the Kaldor-Hicks criterion in a situation where reversals cannot exist and to apply the consumer surplus (CS) in the situation where a representative consumer exists, is basically one and the same thing. Chipman and Moore called the Kaldor-Hicks criterion restricted to situations without reversals of any kind for the Kaldor-Hicks-Samuelson criterion (KHS-criterion for short).

Where does that leave us?

It might seem that the conditions for applying the Kaldor-Hicks criterion in a consistent way are so special that they almost never will be fulfilled. But given the models that we have, this is not necessarily true. Take general equilibrium models as an example – they almost always apply representative consumers, and even if there are more than one of them, they may be sufficiently similar to be compared.

Partial equilibrium analyses are often even more close to the conditions for the Kaldor Hicks criterion to apply. Because we only model a small part of total consumption, we abstract from the prices of most goods and services, and often even from the income effect in the markets of interest. That is, demand in these markets are not supposed to be a function of income. This is really to apply the Gorman Polar Form.

It may be objected that even if the necessary conditions apply to the model, they certainly do not apply to reality. But if we find the models useful and enlightening, there is no intrinsic reason why their

consumer surpluses cannot be useful and enlightening too. Entirely accurate they will never be, but they may at least be good indications of the true consumer preferences.

Quasilinear utility: Problem solved, but reappearing

Setting $b(\mathbf{p}) = 1$ in the equations above, we get the quasilinear indirect utility function. The direct utility function in this case is $U = z + u(\mathbf{x})$, where *z* is some composite good and **x** is a vector of the goods and services we want to model in more detail. This model is of the Gorman Polar Form, a representative consumer exists, the indirect utilities of all individuals can be added, and the Kaldor Hicks criterion can legitimately be applied. Furthermore, the model divides the economy nicely in a sector of interest where a diverse set of goods **x** belong (the transport sector, for instance), and a single composite good *z* representing everything else.

But this is only as long as we maximize *U* subject to one single constraint – the budget constraint. But often, we want to apply constraints on total available time as well (the time budget), and maybe on the various forms of time use. The minimum number on working hours, for instance. The de Serpa model of the allocation and valuation of time (de Serpa 1971), that has formed the basis for most studies on the valuation of travel time savings in transport, is a case in point. The two constraints are reduced to one if we define total income not as *R*, as usual, but as $R + wt_w$, where *w* is the hourly wage and t_w is the number of working hours. Since t_w also figures in the time constraint, we may solve that second constraint for t_w and insert the result in the budget constraint. Maximum potential income is then $R + wT$, where *T* is total available time, while the trips all have generalized costs, i.e. costs composed of a monetary cost and a time cost (trip time times the wage rate, possibly modified by the shadow prices of additional constraints on time uses).

The existence of the additional form of income wt_w disrupts the separation of the indirect utility function in one part that is a function of income and another part that is not. So the Gorman Polar Form no longer applies here, and conditions for The Kaldor Hicks criterion to be used are not fulfilled.

The Kaldor Hicks criterion and standard transport appraisal

The key issue in the appraisal of transport plans and projects is often if the time savings cover the monetary costs. To judge about that, we need to know the value of an hour of travel time savings in different modes and under different circumstances. National and international guidance on this issue is based on average values from stated preference value-of-time studies. Already here it becomes apparent that there can be no question of identifying winners and losers among the travelers based on their individual values of time. They will all be versions of the average traveler in the different settings that he experiences – short and long trips, work and leisure trips, car and public transport trips.

It might of course be quite clear that a given policy benefits individuals using public transport, for instance, while car users stand to lose. But this is on the average and might be different in individual cases. And indeed, the same person might win as a public transport user and lose as a car user. Apart from that, the Kaldor-Hicks criterion boils down here to the question if the “multiplied average traveler” can compensate the public sector, the transport operators and the environment for the costs they incur.

It seems that to substitute the average traveler for the representative traveler does require some sort of justification. We are certainly not compelled for theoretic reasons to follow the average traveler everywhere he goes, but his views might nevertheless be of considerable interest. And in fact, as the distribution of individual values of time in stated preference studies is always skewed to the right

(most people have lower values than the average), assuming that those who benefit from a project are also those that have to pay, more projects get selected using the average value of time than by voting (the mean is to the right of the median).

But by the very nature of things, a cost benefit analysis using values of time from national guidebooks (that is, average values), should always be accompanied by at least some thoughts about distributional consequences – and possibly by some forms of compensation to the losers.

The Kaldor-Hicks criterion and random utility models

If the model is of the GEV family and the marginal utility of money is constant, the logsum formula is a valid welfare measure. It is basically identical to the consumer surplus (CS) welfare measure in the deterministic case. The conditions for it to apply are the same, and when the conditions are met, the EV and CV coincides with the CS, just as in the deterministic case.

The problem is however if the marginal utility of money is indeed constant in the case at hand, and what to do if this is not the case.

It may be shown that when the marginal utility of money is not constant, no representative individual exists, and the Kaldor Hicks criterion does not apply. However, Anders Karlström (1999) has derived an exact and computable measure of expected user benefits (expected equivalent and conditional variations) that may be used for cost benefit analysis with GEV models in which marginal utility of money is not assumed to be constant. So if instead of deterministic values we are willing to make do with expectations, and if the concept of the average (mean) individual is accepted, a meaningful welfare measure exists even in such models.

Alternative approaches

Aggregate willingness-to-pay

Seeing that individual consumer surpluses, equivalent variations or compensating variations cannot legitimately be added without explicit or implicit assumptions on the weight of each individual in the welfare function, some economists have pointed to the following solution: If the aggregate willingness-to-pay in the concrete situation at hand is higher than the costs, the project or the plan should be carried out.

This criterion is impractical, because it requires a local valuation study to be carried out for each alternative in each project proposal. That would be expensive. And unless the stated amounts of money could actually be collected afterwards, the respondents would have every incentive to exaggerate their willingness to pay. A compromise solution could be to regularly carry out valuation studies of a more general type at the local level. But even that would be very expensive, and the chances would be high that the incentive problems would persist.

Voting theory

Transport project appraisal consists of a cost benefit analysis summing up the monetarized impacts plus an analysis of non-monetary impacts. If we let every impact (or suitable aggregate of impacts) be a “voter”, as it were, and let each of these voters rank the alternatives, there will be many different ways of using this information to pick a winning alternative or form an aggregate ranking. Voting theory is both a very old field of research and a very alive one. (See https://en.wikipedia.org/wiki/Category:Voting_theory.) If one does not want to apply the Kaldor Hicks criterion and do a cost benefit analysis, some form of voting theory might be considered, or at least it

can be applied to the non-monetary impacts. We cannot get around Arrow's impossibility theorem[2], but we may for instance be able to define sets from which the winner must be chosen.

Another possible approach is to find a ranking of the alternatives that minimizes the distance to all the individual rankings, as measured for instance by the minimum number of permutations needed to transform all individual rankings to an aggregate one.

Both voting theory and rank aggregation by permutations can be seen as forms of multicriteria analysis, but with set rules instead of subjective expert opinion.

Conclusion

The strict conditions for the aggregation over individual consumer surpluses to form a consistent measure of how much a policy is worth to society are not often met in real life. This means that the Kaldor Hicks criterion cannot be applied without simplifying assumptions, either on the form of the demand functions or on the legitimacy of interpersonal comparisons. But even if it can, compensations to loser are often not feasible, not even in theory. This implies that there is a role for the Pareto criterion to be applied in conjunction with the Kaldor Hicks criterion, to avoid policies with adverse distributional impacts or design compensating schemes to accompany them. There might also be a role for methods based on theories with less strict assumptions than the consumer theory of economics.

References

Chipman, J.S and Moore, J.C. (1979). On social welfare functions and the aggregation of preferences. *Journal of Economic Theory* 21, 111-139.

Chipman, J.S and Moore, J.C. (1994). The measurement of aggregate welfare. In Eichhorn, W. (ed.). *Models and measurement of welfare and inequality*, pp. 552-592. Berlin, Springer Verlag.

de Serpa, A (1971). A theory of the economics of time. *The Economic Journal* **81**, 828-845.

Dupuit A J E J (1844). De la mesure de l'utilité des travaux publics. *Annales des ponts et chaussées, Série II*, 8. Reprinted 1995 in *Revue française d'économie* **10**(2): 55-94.

Gorman, W.M. (1953). Community preference fields. *Econometrica* 21, 63-80.

Hicks, J.R. (1939). The foundations of welfare economics. *Economic Journal* 49, 696-712.

Karlström, A. (1999). *Four essays on spatial modelling and welfare analysis. Royal Institute of Technology.* Department of Infrastructure and Planning. Stockholm.

Samuelson, P.A. (1956). Social indifference curves. *The Quarterly Journal of Economics* **70**(1), 1-22.

Sen, A. (2018). Social choice. In: *The New Palgrave Dictionary of Economics*. London: Macmillan.

[2] Arrow postulated four very reasonable properties that a social welfare function should possess, and showed that no welfare function could possess them all.

Further reading

Anderson, S.P., de Palma, A. and Thisse, J.F. (1992). *Discrete Choice Theory of Product Differentiation*. Cambridge: MIT Press.

Blackorby, C. and Donaldson, D. (1990). A review article: The case against the use of the sum of compensating variations in cost-benefit analysis. *The Canadian Journal of Economics* **23**(3), 471-494.

Chipman, J.S. (2018) Compensation principle. In: *The New Palgrave Dictionary of Economics*. London: Palgrave Macmillan.

Chipman, J.S and Moore, J.C. (1979). On social welfare functions and the aggregation of preferences. *Journal of Economic Theory* 21, 111-139.

Chipman, J.S and Moore, J.C. (1994). The measurement of aggregate welfare. In Eichhorn, W. (ed.). *Models and measurement of welfare and inequality*, pp. 552-592. Berlin, Springer Verlag.

Dahl, G. and Minken, H. (2010). A note on permutations and rank aggregation. *Mathematical and Computer Modelling* **52** (1-2), 380-385.

de Serpa, A (1971). A theory of the economics of time. *The Economic Journal* **81**, 828-845.

Gorman, W.M. (1953). Community preference fields. *Econometrica* 21, 63-80.

Hicks, J.R. (1939). The foundations of welfare economics. *Economic Journal* 49, 696-712.

Karlström, A. (1999). *Four essays on spatial modelling and welfare analysis. Royal Institute of Technology.* Department of Infrastructure and Planning. Stockholm.

Kreps, D.M. (1990). *A course in microeconomic theory*. New York: Harvester Wheatsheaf.

Samuelson, P.A. (1956). Social indifference curves. *The Quarterly Journal of Economics* **70**(1), 1-22.

Sen, A. (2018). Social choice. In: *The New Palgrave Dictionary of Economics*. London: Macmillan.

Varian, H.R. (1978). *Microeconomic Analysis* (3rd edn.). New York: W.W. Norton & Company.

2.2 Transport cost benefit rules

Transport cost benefit rules[3]

Harald Minken

Institute of Transport Economics

Oslo

ABSTRACT

A simple general equilibrium model with transport demands similar to standard transport models is used to derive rules concerning the use of an indirect tax correction factor (ITCF), marginal cost of funds (MCF) and the treatment of taxes in transport cost benefit analysis based on standard transport models. Although different in some respects, the derived rules broadly agree with Swedish and Danish national guidelines, but to a lesser degree with British and Norwegian.

Key words: Transport, cost benefit analysis, marginal cost of funds

[3] This is a revised version of a paper originally presented to the NTF Workshop on the marginal cost of public funds, held in Copenhagen on October 25, 2005. It might have been submitted to the Journal of Transport Economics and Policy, but if so, it obviously must have been rejected.

Contents

1 Introduction

By transport cost benefit rules we understand rules for the formation of the welfare function used to appraise transport projects and transport policies with respect to economic efficiency. This includes rules for the valuation of resources consumed and goods produced in the project, and rules for entering costs and benefits for each class of agents and for treating transfers between them, including the treatment of taxes. Examples are the use of a marginal cost of funds (MCF) or an indirect tax correction factor (ITCF). The choice of a discount rate is another example, but will not concern us here.

Practice and official guidance with respect such rules vary considerably across countries. With varying justifications, the UK and Sweden and a few other countries use the ICTF. The UK justification stems from Professor Sugden (DfT 2002 a,b). According to him, an ITCF should be used to account for the fact that whereas consumers express their valuations in terms of prices that include indirect taxes, firms, being able to reclaim the value added tax and possibly other indirect taxes, do not. Likewise, government, the receiver of indirect tax revenue, should be able to take these revenues into account when assessing the cost of the resources they themselves expend. Thus different agents use different units of account (price indices with and without indirect taxes) when assessing their costs and benefits, whereas a cost benefit analysis, it is argued, should use the same unit of account throughout.

The Swedish justification (Bruzelius 1980 and Halloff and Bruzelius 1986) conforms to former UK thinking on the ITCF, dating back at least to the early seventies.[4]

The use of a marginal cost of public funds is rather more widespread. The concept was introduced in Atkinson and Stern (1974) and seems to have been given its present name by Browning (1976). It is part of standard theory in public economics. It is used to account for the fact that taxes causes private agents to equate their marginal utility to prices that exceed marginal rates of transformation in production, thus creating inefficiencies in the economy. Since public expenditure will have to be financed by taxes, such effects should be taken into account even in cost-benefit analyses that otherwise confine themselves to impacts in the transport sector. Whereas the new UK ITCF makes for different treatment of consumers on the one hand and firms and government on the other, the MCF makes for different treatment of private agents and government agencies.

Swedish and Danish practice is to use both the ITCF and the MCF, while the UK practice is to use the ITCF only, and Norwegian practice is to use the MCF only.[5] Not everybody can be right. The aim of the present paper is to derive a set of rules that can be used to assess and compare the varying practices. Obviously, the first step will be to construct a more or less drastically simplified general equilibrium model of the whole economy, then introduce a transport project that requires funding from outside the transport sector. By studying the

---

4 Harrison (1974, p 153) says:

> … it has become common practice in UK transport appraisals to make an allowance for so-called tax benefits and to adjust the cost of resources entering into the appraisal to allow for the taxation paid on them in other uses. (…) The general procedure (…) is to eliminate significant transport-specific taxes from the estimates of resource cost use within the transport sector (…) and then attach a shadow price to these resources equivalent to the average rate of taxation in other sectors.

5 Current Swedish practice is described in SIKA (1999, 2002). Current Danish practice is described in Finansministeriet (1999), UK practice in DfT (2004) and Norwegian practice in Finansdepartementet (2000) and the cost benefit manuals of the transport sector adminstrations.

economywide impacts of the project it might be possible to express them in terms of data that can be obtained from a partial equilibrium transport analysis and thus to enter these impacts in the form of modifications to the partial equilibrium welfare function. The modifications – be it shadow prices for the resources consumed in the project, a MCF or a ITCF, or rules for entering taxes and other transfers – are our transport cost benefit rules.

We draw on a recent public economics literature (van Dender 2003, Parry and Bento 2001, see also Mayeres and Proost 2001, 1997, Mayeres 2001, 1999) that shows the need to take labour market impacts into account when setting transport taxes and deciding on the use of the revenue. Out aim is however not to measure such impacts, but to find out how to go about an ordinary transport cost benefit analysis without ignoring completely that transport is part of a wider economy. Thus, many of the finer points of this literature are dropped here. In particular, issues of congestion and environmental externalities are not tackled through the MCF but are assumed to be appropriately handled by the transport model system and possibly by environmental post-models. Issues of equity are ignored.

Some general principles will guide us in the formulation of the general equilibrium model and the appraisal framework. First, we do not want to eliminate transfers (from consumers to firms, from consumers and firms to the government and vice versa) at a too early stage, but to enter them explicitly. The reason is that consumer surplus must necessarily be expressed in terms of the market prices that consumers actually face. For instance, it will not do to enter variable public transport operating cost instead of the ticket price in the consumer surplus expression. The fare is a transfer to the public transport operator, but it is not a "mere" transfer. It will have to be entered as a cost in the consumer surplus expression and as a benefit to the operator. Furthermore, if indeed there is a MCF, taxes will also have to be entered explicitly as transfers. This requires us to keep separate accounts for these three sectors of agents and to start from the market prices that faces each of them. Total welfare will be the (possibly weighted) sum of the net benefits of all three sectors.[6]

Second, we recognise that firms are owned by consumers or the government, and their surplus will eventually show up as non-labour income. Third, as far as possible, appraisal should be consistent with the model used to compute the impacts. This is the key to the approach taken here. Models used in partial equilibrium transport analysis are almost universally based on utility functions with strong separability between transport services and other goods. Thus income is not an argument in transport demand functions. We will not be able to use the exactly same form of utility function in the general equilibrium model, since more goods and services (at least, leisure, labour and a general consumption good) will have to be included. But for consistency, transport services should enter the utility function in as much the same way as possible in the two utility functions, including the separability. If in a wider context, separability is seen as unrealistic, it is the transport models that should change. Appraisal should not change until the models do. It is at this point that the present paper differs from most national guideline derivations of cost benefit rules, which often employ a framework that is rather remote from the standard transport modelling framework and contains little of the specific features of transport demand and transport project impacts.

Section 2 describes the simple economy, and section 3 sets out and solves the utility maximisation problem of a representative consumer who derives utility from a general comsumption good, leisure and commuting. Section 4 introduces a marginal reform consisting of a transport infrastructure improvement, partly financed by increasing the labour

[6] For simplicity, we ignore accidents and environmental external effects here. They might affect a fourth party.

tax. Section 5 and 6 specify the government budget and derives the MCF, section 7 derives tax rules, while section 8 summarises the transport cost benefit rules and compares results to current practice. Section 9 concludes.

2 The economy

Consider an economy with consumption, production, export and import, and with a private and public sector. The consumption goods are a composite consumption good x_0, leisure z and trips $\mathbf{x}$. The composite good is produced by private firms with labour as the sole input and sold to households or exported at a fixed export price. Transport service providers may be privately or publicly owned. Transport services are produced using labour, vehicles, and kilometre dependent inputs in fixed proportions. Vehicles and the kilometre dependent inputs are imported at fixed world market prices. The service is sold to households, who use this purchased good plus own time to produce public transport trips. Car trips are also produced in the household, using own time and a third imported good.

The public sector provides non-transport public services in a fixed amount at a fixed cost B. It also provides transport infrastructure g, using labour as the only input. Transport infrastructure is provided for free, and its effect is to reduce transport time. Finally, government may subsidise the private transport providers to the amount S. To pay for these three types of expenses, government levies taxes of three kinds: a value added tax on the composite good, a labour tax, and transport taxes. For some reason (i.e., the analyst's convenience) the value added tax cannot be changed, while the other taxes can.

From the above, we see that the labour supplied by households have three uses: in the composite good sector, the public transport sector and in infrastructure provision. Equilibrium in the labour market and the other markets is assumed. Likewise, we assume that the government maintains a balanced budget and that foreign trade is balanced.

To make the required contact between transport modelling and public economics, we adopt one main simplifying assumption from each field and add one more. In line with standard transport modelling, we assume that a representative consumer with a quasi-linear utility function exists. The effect of this is that no aggregate demand functions except demand for the composite good will be functions of non-labour income. Note in particular that labour supply will not be a function of non-labour income. The labour supply function will be exceedingly simple in other respects as well, due to the similarity between people in an economy where a representative consumer exists and the lack of constraints in the labour market.

In line with much of the public economics literature on the MCF, we assume constant returns to scale and perfect competition in production. Thus, producer prices are unaffected by policy measures. (The public transport sector, while exhibiting constant returns to scale, is however an exception with respect to perfect competition, since it may earn profits or incur losses.) The implication of this is that in the absence of "feedback externalities" like congestion, the total welfare effect of a policy intervention can be analysed using only two equations: the government budget balance and the welfare function.[7] With congestion, a set of relations (for instance, variational inequalities) securing user equilibrium transport times and volumes must

[7] A proof of this for the particular economy at hand can be had from the author on request.

be added, or user equilibrium must be assumed to apply or be built into the welfare function as in combined models.

The additional simplifying assumption is that the value of time in commuting will equal the net wage. This assumption is built into the representative consumer's utility maximisation problem. If both the number of working days and the working hours per day can be freely chosen, as in our model, it might not be that unrealistic.

3 The representative consumer

All demand and supply is measured per week. As already indicated, x_0 is the composite good and z is leisure. There are four types of trips: Commuting by car x_1, commuting by public transport x_2, leisure trips by car x_3 and leisure trips by public transport x_4. Thus $\mathbf{x} = (x_1, x_2, x_3, x_4)$. This is not restrictive: The model can be extended to include more modes and destination choice without altering our results. We also introduce the number ℓ of commuting trips (equals the number of working days) and the number n of leisure trips as choice variables. Labour supply per week is L.

Transport times are $\mathbf{t} = (t_1, t_2, t_3, t_4)$. If there is congestion, **t** and **x** are determined simultaneously, as for instance by **t** = **t**(**x**(**t**)). We do not model this explicitly, but simply *assume* that user equilibrium always applies. But to uphold separability between travel purposes in the face of congestion, we also assume that commuting and leisure trips take place at different times.

Transport prices (the price of the purchased good or service) are $\mathbf{p} = (p_1, p_2, p_3, p_4)$. The price p_0 of the composite goods is set to 1. The net wage is w and non-labour income is R.

The representative consumer's direct utility function is

$$U = x_0 - \frac{1}{b_0} z\left(\ln\frac{z}{z_0} - 1\right) - \left(\frac{1}{b_2} - \frac{1}{b_1}\right)\ell\left(\ln\frac{\ell}{\lambda} - 1\right) - \left(\frac{1}{c_2} - \frac{1}{c_1}\right) n\left(\ln\frac{n}{\eta} - 1\right) + \sum_{i=1}^{4} h_i x_i - \frac{1}{b_1}\sum_{i=1}^{2} x_i(\ln x_i - 1) - \frac{1}{c_1}\sum_{i=3}^{4} x_i(\ln x_i - 1) \tag{1}$$

We require the parameter b_0 to be non-zero, the parameters b_1, b_2, c_1, c_2 to be strictly positive and $b_1 > b_2$, $c_1 > c_2$. The parameters h_i represent the attractiveness of each mode. They may take on any value and one of them for each mode may be set to 0. The three last parameters, z_0, λ and η, are strictly positive and may serve calibration purposes.

The function U with its entropy terms resembles the utility function of a representative traveller in Oppenheim (1995). Apart from the inclusion of leisure and two travel purposes (taking place at different times of day), other differences will appear in the constraints, which are:

$$x_0 + \sum_{i=1}^{4} p_i x_i - wL = R \quad (\mu) \tag{2}$$

$$\sum_{i=1}^{2} t_i x_i + z + L = T_1 \quad (\theta_1) \tag{3}$$

(4) $$\sum_{i=3}^{4} t_i x_i = T_2 \quad (\theta_2)$$

(5) $$x_1 + x_2 = \ell \quad (\varepsilon_1)$$

(6) $$x_3 + x_4 = n \quad (\varepsilon_2)$$

Equation (2) is the money budget. Equations (3) and (4) shows that the consumer splits his time budget in two separate parts, implicitly assumed to take place at different times of day. This is done to achieve indirect strong separability between the travel purposes (the values of time will be different for the two purposes, and demand functions will only be functions of own mode variables). Such separability is implicitly assumed in transport model estimation, as well as in cost benefit analysis when user benefits from different travel purposes are added.

What makes the most difference to the mathematical form of the solution is however that the high level variables *z*, ℓ and *n* are not tied together or constrained separately, not even by inequalities. It is also this feature that makes the model suitable as a model that incorporates free labour supply (both with respect to working days and hours per day) within what is essentially a transport model. For the derivations to follow, the constraints are the most important, while any quasi-linear utility function exhibiting direct strong separability, i.e. of the form $U = x_0 + u^0(z) + u^1(x_1, x_2, \ell) + u^2(x_3, x_4, n)$, could do.

The solution to the problem to maximise *U* subject to (2)-(6) is given in the appendix. As is seen there, the transport part of it consists of hierarchical trip frequency/mode choice models for the two purposes, with binomial logit mode choice. What we need for derivation of cost benefit rules is however only the knowledge of which prices and other variables that enter the different demand functions, and the optimal values of the three first Lagrange multipliers (given in brackets in (2)-(4)). These multipliers are: $\mu^* = 1$, $\theta_1^* = w$ and θ_2^*. No explicit solution can be given for θ_2^* (see appendix equations (A1)-(A3)). With these values, the envelope theorem gives us the following partial derivatives of the indirect utility function *V*:

(7) $$\frac{\partial V}{\partial R} = 1, \ \frac{\partial V}{\partial p_i} = -x_i \ (i = 1,2,3,4), \ \ \frac{\partial V}{\partial t_i} = -w x_i \ (i = 1,2),$$
$$\frac{\partial V}{\partial t_i} = -\theta_2^* x_i \ (i = 3,4), \ \ \frac{\partial V}{\partial w} = L$$

4 A marginal policy reform

A marginal policy reform is a small change in infrastructure provision, taxation, and subsidies to public transport, causing small changes in trip times. It is to be implemented under a balanced government budget constraint. To assess the impact on indirect utility, we will have to specify the tax structure. Let s_0 be the value added tax rate, *s* the (marginal and average) labour tax and τ_i the transport taxes (*i* = 1,2,3,4). We denote the gross wage by ω and the net of taxes/production prices of the consumption goods by π_i (i = 0,1,2,3,4). Assume:

(8) $$p_0 = 1 = (1 + s_0)\pi_0$$

(9) $$w = (1 - s)\omega$$

(10) $$p_i = \tau_i + \pi_i, \quad i = 1,2,3,4$$

The meaning of (8) is that we use the market price of the composite good as our unit of account and that units of this good is measured such that one unit costs 1. Equation (9) implies that the marginal tax rate equals the average rate, which is also a gross simplification from the standpoint of advanced MCF literature such as Snow and Warren (1996). The transport taxes of (10) may or may not include a value added tax.

For convenience, we assume that $\tau_1 = \tau_3, \pi_1 = \pi_3, \tau_2 = \tau_4, \pi_2 = \pi_4$ and consequently $p_1 = p_3, p_2 = p_4$. This rules out congestion pricing unless the mix of travel purposes are the same in congested and uncongested periods, but makes the derivations a little less tedious without affecting the main structure of the cost benefit rules.

The subsidy *S* and infrastructure provision *g* was introduced in chapter 2. We may now specify a marginal policy reform as a change $(ds, d\tau_1, d\tau_2, dS, dg)$ with impacts (dt_1, dt_2, dt_3, dt_4) on trip times. Using (7)-(10) and seeing that *S* and *g* are not variables in the indirect utility function *V*, the impact on *V* is found to be

$$dV = -\omega L ds - (x_1 + x_3)\, d\tau_1 - (x_2 + x_4)\, d\tau_2 + \sum_{i=1}^{4} \frac{\partial V}{\partial t_i} dt_i$$

The last three terms is the marginal user benefit *dUB* as it would have been assessed in a transport analysis. Thus

(11) $$dV = dUB - \omega L ds$$

As we have said, public transport may be privately or publicly owned, exhibits constant returns to scale and uses inputs in fixed proportions. On the average, or sooner or later, the households who own shares in public transport companies will be able to pocket the surplus (or, perhaps more realistically, will have to bear the deficit). This income will be part of their non-labour income *R*. It is rational for the individual households to ignore the effect of their choices on the profits of the public transport companies that they own, since their own public transport trips are only a very tiny part of all public transport trips. Consequently, the representative consumer also ignores the possibility to influence these profits. But the analyst, taking the point of view of society, will have to take them into account. If the share of public ownership is *a* and the constant marginal cost of a public transport trip is *c*, the analyst will recognise that *R* can be written

(12) $$R = R_0 + (1-a)\left[(\pi_2 - c)(x_2 + x_4) + S\right]$$

where *S* is the subsidy given to public transport companies by the government. The social welfare function *W* that we will use to appraise reforms is the indirect utility function of the representative consumer with *R* written in the form of (12). Now by (7), (11) and (12):

$$\begin{aligned} dW &= dUB + (1-a)\left[(\pi_2 - c)(dx_2 + dx_4) + dS\right] - \omega L ds \\ &= dUB + (1-a)\left[(\pi_2 - c)(Dx_2 + Dx_4) + dS\right] \\ &\quad - \left[\omega L - (1-a)(\pi_2 - c)\frac{\partial x_2}{\partial s}\right] ds \end{aligned} \tag{13}$$

where we have introduced the operator D defined by

(14) $$Dy = \sum_{i=1}^{2} \frac{\partial y}{\partial \tau_i} d\tau_i + \sum_{i=1}^{4} \frac{\partial y}{\partial t_i} dt_i$$

The middle term of the last two lines of (13) is the marginal benefit to private public transport operators as it would have been entered in a transport analysis. We have however assumed that ultimately it will accrue to households and rationalised its inclusion in the cost benefit calculation on that basis. Either way you see it – before or after dividends have been paid or share values have gone up – the value in the cost benefit calculation is the same. Thus, there is no basis for an indirect tax correction factor that treats benefits to firms and to households differently. Instead of recognising that firms are owned by households, this type of indirect tax correction factor treats firms as a special class of consumers with tax exemptions.

5 Benefits to government

From (8) it is seen that the tax revenue from the value added tax is $s_0\pi_0 = s_0\left(1+s_0\right)^{-1}$. The government budget, consisting of fixed outlays *B* and transfers and infrastructure costs, and financed by taxation and profits from publicly owned firms, is then:

(15) $$B = \frac{s_0}{1+s_0} x_0 + s\omega L + \tau_1\left(x_1+x_3\right) + \tau_2\left(x_2+x_4\right) + a\left[\left(\pi_2 - c\right)\left(x_2+x_4\right) + S\right] - S - \omega g$$

If we move the last two terms on the right over to the left, (15) says that expenses equal revenues, or the budget is balanced. The term ωg is the infrastructure cost. The assumptions are constant returns to scale in infrastructure building and upkeep, that the only input is labour, and that output is measured in the same units as input. The costs are measured by the gross wage as it would in private firms, while the labour tax revenue of infrastructure workers is a part of $s\omega L$.

Since the budget should always be balanced, we may interpret *B* to be a function of the right hand variables. Bearing in mind, however, that we do not want *B* to change, a marginal policy reform must fulfil the requirement

$$0 = \frac{\partial B}{\partial s} ds + \sum_{i=1}^{2} \frac{\partial B}{\partial \tau_i} d\tau_i + \sum_{i=1}^{4} \frac{\partial B}{\partial t_i} dt_i + \frac{\partial B}{\partial S} dS + \frac{\partial B}{\partial g} dg$$

Solving for ds and using the operator D:

(16) $$ds = -\frac{1}{\frac{\partial B}{\partial s}}\left[DB + \frac{\partial B}{\partial S} dS + \frac{\partial B}{\partial g} dg \right]$$

While (13) is a perfectly correct formula to appraise the economic efficiency of a marginal policy reform in the simplified general equilibrium framework we have developed, it is of little use if *ds* cannot be assessed. This is the case if (13) is applied to a partial (transport only) analysis. Using (16) to substitute for *ds* in (13) might help us out of the predicament, provided (a) that the right hand side of (16) can be assessed with transport model data, and (b) that the transport part of the general equilibrium framework is sufficiently similar to the partial transport model we are using, so that one the one hand, the partial model would have been capable of being extended to a general equilibrium model, and on the other hand, the general equilibrium framework is capable of producing transport demands of a similar type to the demand functions we have. These two provisions are the reasons why we have made the

assumptions about consumption and the economy that we have. If they are found to be utterly unrealistic, we had perhaps better not use (13) and (16) for appraisal.

It is standard in public economics to define the marginal cost of funds associated with using the tax *s* to finance public spending, MCF_s, by

$$MCF_s = -\frac{\frac{\partial W}{\partial s}}{\frac{\partial B}{\partial s}} \tag{17}$$

In our case, *s* is the labour tax rate. We may also use transport taxes, but provided (16) can be assessed from transport model data, their impact in our framework is fully accounted for, and they do not give rise to unknown wider impacts. Dropping the subscript on MCF and recognising by (13) that the partial derivative of *W* with respect to *s* is $-\omega L + (1-a)(\pi_2 - c)\frac{\partial x_2}{\partial s}$, we get

$$\begin{aligned} dW &= dUB + (1-a)\left[(\pi_2 - c)(Dx_2 + Dx_4) + dS\right] \\ &+ MCF\left[DB + \frac{\partial B}{\partial S}dS + \frac{\partial B}{\partial g}dg\right] \end{aligned} \tag{18}$$

According to (18), marginal welfare consists of marginal user benefits, marginal public transport surplus after subsidies and marginal government surplus times MCF. A fourth class, non-feedback externalities, has been ignored here. Our two remaining tasks are to evaluate MCF and the bracketed terms of the last line.

6 Evaluating MCF

Considering the grossly simplified representation of labour supply in our framework, there is perhaps little to be said for using it to actually estimate MCF. Nevertheless, the structure of it might be of use to us.

Applying our model to (17), $-\partial W/\partial s$ is the terms in the bracket of the last line of (13). To evaluate $\partial B/\partial s$, we start by eliminating x_0 from (15). Using (2), (8)-(10) and (12), we have

$$x_0 = R_0 + (1-a)S + (1-a)(\pi_2 - c)(x_2 + x_4) + (1-s)\omega L - \sum_{i=1}^{4}(\tau_i + \pi_i)x_i \tag{19}$$

Differentiating *B* with respect to *s* using the above expression for x_0 gives

$$\frac{\partial B}{\partial s} = \frac{1}{1+s_0}\left[M_1\frac{\partial x_1}{\partial s} + M_2\frac{\partial x_2}{\partial s} + \omega L + (s+s_0)\omega\frac{\partial L}{\partial s}\right] \tag{20}$$

where

$$M_1 = \tau_1 - s_0\pi_1, \quad M_2 = \tau_2 - s_0\pi_2 + (s_0 + a)(\pi_2 - c) \tag{21}$$

Inserting in (17) and rearranging:

(22) $$MCF = \frac{\omega L - (1-a)(\pi_2 - c)\frac{\partial x_2}{\partial s}}{\omega L} \cdot \frac{1+s_0}{1 - \frac{s+s_0}{1-s} El_w L + \frac{M_1 \frac{\partial x_1}{\partial s} + M_2 \frac{\partial x_2}{\partial s}}{\omega L}}$$

In (22), we have replaced the elasticity of labour supply with respect to the labour tax, which occurs naturally in the derivation, by the elasticity of labour with respect to the net wage. ($El_s L = -s(1-s)^{-1} El_w L$.)

The first factor of MCF is likely to be very close to 1 (since expenditure on public transport tickets is a small fraction of gross wage income ωL and the *change* in public transport surplus is probably a small fraction of expenditure on tickets). We propose to ignore it.

By the same reasoning, the third term in the denominator of the second factor is probably small, although possibly not ignorable. It is the marginal change in government revenue from commodity taxation induced by the work trip consequences of a marginal change in labour taxation, measured as a proportion of all gross wages. If the first factor is set to 1, it is by this term that mode split, transport taxes, transport input production prices and in fact transport ownership influences the MCF and might call for a specific MCF for each study area or model. Note that it is only the commuting trips that may influence MCF – leisure trips are not a function of *s* in our model. Even if our MCF is essentially specific for a study area, it is not dependent on the particular policy reform to be appraised.

So far results could be derived with the more general quasi-linear and strongly separable utility function, and we have not made use of the particular mathematical form of (1). It might however be used to assess the transport-dependent term of MCF. Taking the derivatives of x_1 and x_2 as given in the appendix with respect to *s*, the third term of the denominator of the second factor becomes

(23) $$\frac{M_1 \frac{\partial x_1}{\partial s} + M_2 \frac{\partial x_2}{\partial s}}{\omega L} = \frac{b_1 \frac{x_1 x_2}{x_1 + x_2}(M_1 - M_2)(t_1 - t_2) + b_2 \frac{M_1 x_1 + M_2 x_2}{x_1 + x_2} \sum_{i=1}^{2} t_i x_i}{L}$$

Average travel statistics and average working hours for a study area, tax rates and some knowledge about elasticities to determine b_1 and b_2 may be used to assess (23). The same data and parameter values may also be used to evaluate the second term of the denominator, the net wage elasticity of labour supply. The formula is given in the appendix.

The main determinants of our MCF will be the value added tax rate and the net wage elasticity of labour supply. If labour supply is inelastic and trips are not functions of *s*, MCF = $1 + s_0$, which is around 1.2 in most countries. If $s_0 = 0.2$, $s = 0.3$, the labour supply elasticity is 0.1 and the third term of the denominator is negligible, MCF = 1.29.

7 Rules for entering taxes

We turn now to the evaluation of the bracketed terms of the last line of (18). We start by rewriting (15) with the new expression for x_0 and the definitions given above. Collecting terms we get

(24) $$B = \frac{1}{1+s_0}\left[s_0 R_0 + (1-a)S + (s+s_0)\omega L + M_1(x_1 + x_3) + M_2(x_2 + x_4)\right] - \omega g$$

Totally differentiating:

$$(25)\qquad \begin{aligned} -\frac{\partial B}{\partial s}ds &= \sum_{i=1}^{2}\frac{\partial B}{\partial \tau_i}d\tau_i + \sum_{i=1}^{4}\frac{\partial B}{\partial t_i}dt_i + \frac{\partial B}{\partial S}dS + \frac{\partial B}{\partial g}dg = DB + \frac{\partial B}{\partial S}dS + \frac{\partial B}{\partial g}dg \\ &= \frac{1}{1+s_0}A - \frac{1-a}{1+s_0}dS - \omega dg \end{aligned}$$

where A is defined by

$$(26)\qquad A = D\left[M_1\left(x_1+x_3\right)+M_2\left(x_2+x_4\right)\right]-\left(s+s_0\right)\omega\left(t_1Dx_1+t_2Dx_2\right)$$

With this A, and using P as shorthand for the marginal producer surplus, (18) becomes

$$(27)\qquad dW = dUB + P + MCF\left[\frac{1}{1+s_0}A - \frac{1-a}{1+s_0}dS - \omega dg\right]$$

The first term of A is the marginal change in government revenue from transport taxes and publicly owned transport firms, minus the lost revenue from value added taxes on other commodities and services because of the shift in consumption from general consumption to purchased goods from transport or vice versa. This term accomplishes the "indirect tax correction" as far as consumption goods are concerned. Namely, it means to enter transport taxes in the government account of the cost benefit calculation not at face value, but at the shadow prices given by (21). Prices including taxes have been used (and must be used if there is a non-marginal change) in the calculation of user benefits. By entering the full transport tax revenue as a benefit in the calculation of benefits to government, we recognise that the inclusive price as perceived by consumers is too high from society's point of view. But this "correction" is too drastic, because consumption on transport crowds out other consumption. The correct correction is achieved if we enter transport taxes as a benefit to government at the modified rates of (21), divide by $1 + s_0$ and multiply by the MCF as seen by (27).

The second term of A introduces a "correction" to the value of the resource of time consumed in transport. Labour tax revenue is lost if commuting is taking more time, and there is also a loss of value added tax caused by the consumers' labour income loss when transport takes more time. Actually, if the cost benefit calculation uses the actual market prices that each of the two private sectors (households and firms) face, uses the same prices for government as for firms, and enters transfers between the three sectors explicitly, there is no need to talk about corrections at all. What is seen as corrections is better seen as keeping track of the tax revenue implications for government. All of the tax implications can be computed from a transport model since the operator D recognises only that trips are functions of trip prices and trip times.

It might be more transparent to split the shadow prices M_i and perform the calculations of the terms separately. Thus

$$(28)\qquad \begin{aligned} &A = D\left[\tau_1\left(x_1+x_3\right)+\tau_2\left(x_2+x_4\right)\right]-\left(s+s_0\right)\omega\left(t_1Dx_1+t_2Dx_2\right) \\ &-s_0\left[\pi_1\left(Dx_1+Dx_3\right)+\pi_2\left(Dx_2+Dx_4\right)\right]+\left(s_0+a\right)\left(\pi_2-c\right)\left(Dx_2+Dx_4\right) \end{aligned}$$

The fourth term here is the increase in government income from public transport firms – the ones it owns itself and the taxes collected from the privately owned. Equation (27) shows that all of the tax income and profits, as well as lump-sum transfers to the private sector, should be divided by one plus the value added tax rate, while the cost of public production of public goods should not. It is interesting to note that if government owns public transport ($a = 1$) its

costs and revenue from this should be treated on a par with infrastructure production, according to the fourth term of *A*. Thus, all public production, whether of public goods or private goods, should be treated the same, and differently from taxes and transfers. Perhaps the most convenient way of summarising this is to say that a different MCF should be applied to taxes and transfers on the one hand and public production on the other.

This finishes the derivation of results for this model. What are its weakest or most debatable points, and how do they influence the results? There are at least two such points. First, the second term of *A* in (26), the deduction of taxes lost due to transport taking more time, appears because since optimal leisure is not a function of transport taxes, the *only* alternative use of transport time when transport taxes change is labour. Although some such effect is to be expected, the second term of *A* takes it too far. Empirical values of time in commuting are lower than *w*, indicating that leisure is the most probable alternative to spending time commuting – perhaps due to constraints on work hours in the short run. Second, the two different MCF's must be caused by public production and taxes and transfers eliciting different behavioural responses in the private sector, in particular with respect to labour and general consumption. Consider a one euro increase in public production versus a one euro increase in transfers, both financed by an increase in *s*. In both cases, the increase in *s* has the impact (-1)*MCF on the private sector. But while the subsidy is wholly transformed into an increase in x_0 and does not affect *L* according to our model, infrastructure production affects both labour supply and general consumption through its impact on transport times. While little is known about the elasticity of labour supply with respect to commuting time, it would be a wonder if it conformed to formulas (A25) and (A26) of the appendix. Thus, there are reasons to be cautious with respect to the factor $(1+s_0)^{-1}$ applied to modify MCF in the case of subsidies and taxes.

8 Cost benefit rules

Subject to the above qualifications, the following transport cost benefit rules can be extracted:

1. There is no need for an indirect tax correction factor, and to the extent that it adjusts the market prices confronting firms and households differently, it is wrong.

2. A marginal cost of funds should be applied to all government revenue and cost. However, the MCF for taxes and transfers is less than the MCF for public production by a factor of $(1+s_0)^{-1}$.

3. Transport taxes should be entered as government revenue, and deductions should be made for the loss of income tax and indirect tax due to more time being used in transport instead of work, and for the loss of indirect tax due to more produced resources being consumed in transport instead of in the production of non-transport commodities and services.

4. Economic efficiency may be calculated as the sum of three sectors (four if non-feedback externalities are included). They are: households, private firms and government. Prices to use for the two first sectors are the market prices confronting each sector, while transfers between all sectors (tickets, taxes, subsidy) are entered explicitly. For government production, use the prices confronting private firms. The structure of the welfare function may be written

$$W = UB + P + MCF\left[\frac{1}{1+s_0}A - \frac{1-a}{1+s_0}S - \omega g\right]$$

where P is the private operator surplus including transfers and A is transport tax revenue minus lost non-transport tax revenue and plus surplus of publicly owned firms.

5. These rules apply for cost benefit analysis based on transport models that can be derived from the utility maximisation problem of a representative consumer with a quasi-linear utility function. If models change, so might appraisal.

We now briefly compare these rules to the official guidance and actual practice in the UK, Sweden, Denmark and Norway. A distinction should be made between inputs used in public production and resources used by private agents when they make use of the public good or otherwise adapt to the project. It turns out that much guidance issued by ministries of finance implicitly think of the project as the provision of a pure public good, the consumption of which requires no additional private resources. With respect to inputs used in public production, Danish and Swedish guidance is clear: Strip them of taxes and adjust the result upward by an ITCF of 1.2, say.[8] This is really the same as our rule to apply a different MCF to public production and other public income and expenditure. However, an inconsistency results when a MCF conforming to (17) is applied with the "multiply input production prices by 1.2" rule, because the 1.2 is really taken from the MCF, as shown by (27) together with (22).

In Denmark and Sweden and most other countries, consumer goods are valued at market prices. Turning to the tax revenue consequences of private use of resources, guidance from ministries of finance is much more unclear. Danish guidance on tax revenue is to ignore it. Happily, Swedish guidance is made by transport agencies. It is to use the M_i basically as we do. That is, the ITCF is extended to privately consumed resources. They also use an MCF, and so the only difference concerns the second factor of A. However, current practice lags somewhat behind guidance in so far as project costs, but not tax revenue, is subjected to MCF. Danish guidance on MCF seems to be similar to Swedish practice.

Norwegian guidance is issued by the Ministry of Finance and therefore somewhat detached from transport cost benefit analysis problems. An MCF is applied to project costs, and it is stressed that to achieve economic efficiency in production, project inputs should be valued at the same prices as used in competing private uses. ITCF is not used and current practice in transport analysis is to enter the full tax revenue as government benefit. Clearly, this should be reconsidered.

UK guidance does not use the MCF but the ITCF as currently understood does perform a similar role with respect to the private sector as a whole versus the public sector as a whole. This is seen if we realise that if labour supply and commuting trips are inelastic with respect to s, MCF in (22) becomes $1 + s_0$. In fact, the basic model used to justify the ITCF in Sugden (2002a) assumes inelastic labour supply and contains no trips. An appeal is made to the same reasoning from the UK seventies that is behind the Swedish ITCF. Tax revenue is entered in the public accounts with a correction for the crowding out of general consumption (DfT 2004). However, UK guidance also makes benefits to firms more valuable relative to benefits

[8] The Danish ITCF is below 1.2 and the Swedish and UK ITCF are above. Labour wages are not stripped of taxes but valued at the price including taxes and social costs, then adjusted upward by the ITCF.

to households. The reasoning leading to this result involves an interpretation of the ITCF in terms of equalisation of units of account across sectors. We claim to have shown that in a more realistic model with many goods, this equalisation (if it is called for at all) is accomplished by entering tax revenue to government in the right way and by fully taking in that the profit of firms accrue to households.

9 Conclusion

A simple general equilibrium model with transport demands similar to standard transport models has been used to derive rules concerning the use of an indirect tax correction factor (ITCF), marginal cost of funds (MCF) and the treatment of taxes in transport cost benefit analysis based on standard transport models. It is found that it is wrong to apply an ITCF that adjusts prices facing households and firms differently. The MCF should be used, but with a different value for taxes and transfers on the one hand and public production on the other. From government transport tax revenue should be deducted labour and value added taxes lost due to time and purchased goods being redirected to transport instead of other production. Although different in some respects, these rules broadly agree with Swedish and Danish national guidelines, but to a lesser degree with British and Norwegian

10 References

Atkinson, A.B and Stern, N.H. (1974) Pigou, taxation and public goods. *Review of Economic Studies* **41**(1), 119-128.

Browning, E.K. (1976) The marginal cost of public funds. *Journal of Political Economy* **84**, 283-298.

Bruzelius, N. (1980) *Samhällsekonomiska kostnads-inntäktskalkyler: Teori och tillämpning på investeringar i transportsektorn*. BFR rapport R 97:1980.

Bruzelius, N. and U. Halloff (1986) *Lönsamhetsbedömning av investeringar i transportsektorn*. Rapport 1986:21, Transportforskningsberedningen, Stockholm.

Department for transport (DfT) (2002a) *The Treatment of Taxation in the Cost-Benefit Appraisal of Transport Investment.* Published 28 March 2002.. http://www.dft.gov.uk/stellent/groups/dft_econappr/documents/page/dft_econappr_504910.hcsp

Department for transport (DfT) (2002b) Developing a Consistent Cost-Benefit Framework for Multi-Modal Transport Appraisal. Published 28 March 2002. http://www.dft.gov.uk/stellent/groups/dft_econappr/documents/page/dft_econappr_504897.hcsp

Department for Transport (DfT) (2004) Transport Analysis Guidance. www.webtag.org.uk/index.htm.

Finansdepartementet (2000) Veileder i samfunnsøkonomiske analyser.

Finansministeriet (1999) *Vejledning i udarbeidelse af samfundsøkonomiske konsekvensvurderinger.* http://www.fm.dk/db/filarkiv/2950/hele.pdf.

Harrison, A.J. (1974) *The Economics of Transport Appraisal*. Croom Helm, London.

Mayeres, I. (2000) The efficiency effects of transport policies in the presence of externalities and distortionary taxes. *Journal of Transport Economics and Policy* **34**, 233-259.

Mayeres, I. (1999) *The Control of Transport Externalities: A General Equilibrium Analysis.* Ph.D thesis no. 126 1999, Katholieke Universiteit Leuven.

Mayeres, I. and S. Proost (2001) Marginal tax reform, externalities and income distribution. *Journal of Public Economics* **79**(2), 343-363.

Mayeres, I. and S. Proost (1997) Optimal Tax and Public Investment Rules for Congestion Type of Externalities. *Scandinavian Journal of Economics* **99**, 261-279.

Oppenheim, N. (1995) *Urban travel demand modeling*. John Wiley & Sons, New York.

Parry, I.W.H. and A.M.R Bento (2001) Revenue Recycling and the Welfare Effects of Road Pricing. *Scandinavian Journal of Economics* **103**(4), 645-671. Also: Policy Research Working Paper 2253, The World Bank, Washington DC.

SIKA (2002) *Övergripande kalkylparametrar*. SIKA Rapport 2002:7. http://www.sika-institute.se.

Snow, A. and R.S. Warren (1996) The marginal welfare cost of public funds: Theory and estimates. *Journal of Public Economics* **61**, 289-305.

Van Dender, K. (2003) Transport Taxes with Multiple Trip Purposes. *The Scandinavian Journal of Economics* **105**(2), 295-310.

APPENDIX

First order conditions for the problem to maximise $U = x_0 + u^0(z) + u^1(x_1, x_2, \ell) + u^2(x_3, x_4, n)$ *subject to (2)-(6):*

(A1) $$\frac{\partial \mathsf{L}}{\partial x_0} = 1 - \mu = 0$$

(A2) $$\frac{\partial \mathsf{L}}{\partial z} = \frac{\partial u^0}{\partial z} - \theta_1 = 0$$

(A3) $$\frac{\partial \mathsf{L}}{\partial L} = \mu w - \theta_1 = 0$$

(A4) $$\frac{\partial \mathsf{L}}{\partial \ell} = \frac{\partial u^1}{\partial \ell} + \varepsilon_1 = 0$$

(A5) $$\frac{\partial \mathsf{L}}{\partial x_i} = \frac{\partial u^1}{\partial x_i} - \mu p_i - \theta_1 t_i - \varepsilon_1 = 0 \quad i = 1,2$$

(A6) $$\frac{\partial \mathsf{L}}{\partial n} = \frac{\partial u^2}{\partial n} + \varepsilon_2 = 0$$

(A7) $$\frac{\partial \mathsf{L}}{\partial x_i} = \frac{\partial u^2}{\partial x_i} - \mu p_i - \theta_2 t_i - \varepsilon_2 = 0 \quad i = 3,4$$

(A8) (A1) and (A3) applied to (A2) shows that z* is a function of w alone. Likewise, if solutions from (A1) and (A3) are used in (A4) - (A5), it is seen that (x_1, x_2, ℓ) depend only on (p_1, p_2, t_1, t_2, w). Similarly, (x_3, x_4, n) depend only on (p_3, p_4, t_3, t_4, w). (Congestion may however create a link between x_1, x_2 and x_3, x_4, which is why we assume commuting and leisure to take place at different times.)

First order conditions for the problem to maximise U of (1) subject to (2)-(6):

(A9) $$\frac{\partial \mathsf{L}}{\partial x_0} = 1 - \mu = 0$$

(A10) $$\frac{\partial \mathsf{L}}{\partial z} = -b_0 \ln \frac{z}{z_0} - \theta_1 = 0$$

(A11) $$\frac{\partial \mathsf{L}}{\partial L} = \mu w - \theta_1 = 0$$

(A12) $$\frac{\partial \mathsf{L}}{\partial \ell} = \left(\frac{1}{b_1} - \frac{1}{b_2}\right) \ln \frac{\ell}{\lambda} + \varepsilon_1 = 0$$

(A13) $$\frac{\partial \mathsf{L}}{\partial x_i} = h_i - \frac{1}{b_1} \ln x_i - \mu p_i - \theta_1 t_i - \varepsilon_1 = 0 \quad i = 1,2$$

(A14) $$\frac{\partial \mathsf{L}}{\partial n} = \left(\frac{1}{c_1} - \frac{1}{c_2}\right) \ln \frac{n}{\eta} + \varepsilon_2 = 0$$

(A15) $$\frac{\partial \mathsf{L}}{\partial x_i} = h_i - \frac{1}{b_1} \ln x_i - \mu p_i - \theta_2 t_i - \varepsilon_2 = 0 \quad i = 3,4$$

The solution to the problem to maximise U subject to (2)-(6) is:

(A16) $$z^* = z_0 \exp\left(-b_0 w\right)$$

(A17) $$\ell^* = \lambda\left[\lambda^{-1}\left(\sum_{j=1}^{2}\exp\left(b_1\left(h_j-\left(p_j+wt_j\right)\right)\right)\right)\right]^{\frac{b_2}{b_1}}$$

(A18) $$x_i^* = \ell^* \cdot \frac{\exp\left(b_1\left(h_i-\left(p_i+wt_i\right)\right)\right)}{\sum_{j=1}^{2}\exp\left(b_1\left(h_j-\left(p_j+wt_j\right)\right)\right)} \quad i=1,2$$

(A19) $$L^* = T_1 - z^* - \sum_{i=1}^{2} t_i x_i^*$$

(A20) $$n^* = \eta\left[\eta^{-1}\left(\sum_{j=3}^{4}\exp\left(c_1\left(h_j-\left(p_j+\theta_2^* t_j\right)\right)\right)\right)\right]^{\frac{c_2}{c_1}}$$

(A21) $$x_i^* = n^* \cdot \frac{\exp\left(c_1\left(h_i-\left(p_i+\theta_2^* t_i\right)\right)\right)}{\sum_{j=3}^{4}\exp\left(c_1\left(h_j-\left(p_j+\theta_2^* t_j\right)\right)\right)} \quad i=3,4$$

(A22) $$\theta_2^* \text{ solves } \sum_{j=3}^{4} t_j x_j^* = T_2$$

(A23) $$x_0^* = R + wL^* - \sum_{i=1}^{4} p_i x_1^*$$

(A24) $$V = R + wT_1 + \theta_2^* T_2 + \frac{1}{b_0} z^* + \frac{1}{b_2}\ell^* + \frac{1}{c_2} n^*$$

Some elasticities of L:

(A25) $$El_w L = \frac{w}{L}\left[b_0 z + \frac{1}{x_1+x_2}\left(b_1 x_1 x_2\left(t_1-t_2\right)^2 + b_2\left(t_1 x_1 + t_2 x_2\right)^2\right)\right]$$

(A26) $$El_{t_1} L = \frac{t_1 x_1}{L}\left\{-1+\left(1-s\right)\omega\left[b_1\frac{x_2\left(t_1-t_2\right)}{x_1+x_2} + b_2\frac{t_1 x_1 + t_2 x_2}{x_1+x_2}\right]\right\}$$

(A27) $$El_{t_2} L = \frac{t_2 x_2}{L}\left\{-1+\left(1-s\right)\omega\left[b_1\frac{x_1\left(t_2-t_1\right)}{x_1+x_2} + b_2\frac{t_1 x_1 + t_2 x_2}{x_1+x_2}\right]\right\}$$

Full expansion of A:

$$
\begin{aligned}
A = {} & d\tau_1\left[-\left(s+s_0\right)\omega\left(t_1\frac{\partial x_1}{\partial \tau_1}+t_2\frac{\partial x_2}{\partial \tau_1}\right)+M_1\left(\frac{\partial x_1}{\partial \tau_1}+\frac{\partial x_3}{\partial \tau_1}\right)+\left(x_1+x_3\right)+M_2\left(\frac{\partial x_2}{\partial \tau_1}+\frac{\partial x_4}{\partial \tau_1}\right)\right] \\
& +d\tau_2\left[-\left(s+s_0\right)\omega\left(t_1\frac{\partial x_1}{\partial \tau_2}+t_2\frac{\partial x_2}{\partial \tau_2}\right)+M_1\left(\frac{\partial x_1}{\partial \tau_2}+\frac{\partial x_3}{\partial \tau_2}\right)+\left(x_2+x_4\right)+M_2\left(\frac{\partial x_2}{\partial \tau_2}+\frac{\partial x_4}{\partial \tau_2}\right)\right] \\
& +dt_1\left[-\left(s+s_0\right)\omega\left(t_1\frac{\partial x_1}{\partial t_1}+t_2\frac{\partial x_2}{\partial t_1}+x_1\right)+M_1\frac{\partial x_1}{\partial t_1}+M_2\frac{\partial x_2}{\partial t_1}\right] \\
& +dt_2\left[-\left(s+s_0\right)\omega\left(t_1\frac{\partial x_1}{\partial t_2}+t_2\frac{\partial x_2}{\partial t_2}+x_2\right)+M_1\frac{\partial x_1}{\partial t_2}+M_2\frac{\partial x_2}{\partial t_2}\right] \\
& +dt_3\left[M_1\frac{\partial x_3}{\partial t_3}+M_2\frac{\partial x_4}{\partial t_3}\right]+dt_4\left[M_1\frac{\partial x_3}{\partial t_4}+M_2\frac{\partial x_4}{\partial t_4}\right]
\end{aligned}
\tag{A28}
$$

2.3 A note on the effects of the marginal cost of funds and the indirect tax correction factor

A note on the effects of the marginal cost of funds and the indirect tax correction factor on optimal user charges in transport, with a caution[9]

Harald Minken

Institute of Transport Economics

Contents

[9] This is a TOI working paper produced for the PROSPECTS project. (Working paper TØ/1898/2006, dated August 31st 2006.)

1 Introduction

From an economic viewpoint, the two main types of objectives are economic efficiency and equity. In transport, there are usually also environmental and accident objectives that cannot be subsumed under economic efficiency or equity. Financing is an objective in the sense that all infrastructure investment and other costly measures must be financed. User charges and investment are policy instruments to achieve these objectives.

Even if financing is an objective, it is not an objective to achieve it by a particular instrument alone. Financing may be achieved by non-transport taxes, transport taxes, user charges or private contributions. The mix of revenue generating instruments that best fulfils the other objectives should be chosen. To impose restrictions on the use of funds from outside the transport sector in transport, or on the use of funds from transport taxes and charges outside the transport sector, can only reduce the level of fulfilment of the objectives. Thus the need to finance transport investments in an optimal way creates a link between transport and the wider economy and may induce changes in the whole tax system.

Taking such a broad view on financing, however, requires us to take account of the wider economy impacts of using non-transport taxes in transport, and of the impacts of transport policy on non-transport tax revenue. Ignoring industrial reorganisation benefits, land use changes and other long-term impacts of transport improvements, these two are the most important ways in which the economic efficiency of a transport policy is influenced by feedback to the wider economy.

In the Scandinavian countries, manuals and official guidance in transport cost benefit analysis take account of these two impacts in slightly different ways. In all three countries, the wider economy impact of using non-transport taxes in transport is summarised by multiplying the net result for the government sector by a Marginal Cost of Funds (MCF). This is because any negative net result will have to be financed by general taxation, which brings about inefficiency in the economy by forcing prices to exceed marginal costs. Conversely, any positive net result will improve efficiency by saving on general taxation. Following the theoretical public economics literature of the last 30 years, the use of MCF in cost benefit analysis is now quite common across the world.

The impacts of transport policy on non-transport tax revenue are the following: When consumers use more money in transport, they will have to cut back on other expenditure. This entails a loss of government revenue from the value added tax. Also, using more time in transport may mean using less time for work, with the effect that the government loses revenue from labour tax and from the value added tax on commodities that would have been bought if the wage income was higher. Following practice that originated in the UK around 1970 (now called the Indirect Tax Correction Factor, ITCF), the Scandinavian countries take these effects into account in varying degree and in different ways. The UK, New Zealand, Sweden and Denmark use the ITCF in about the same way, but elsewhere it has not been used until very recently. (It was adopted in a couple of important EU research projects, and has been introduced in another form in the new CBA manual of the Norwegian Road Administration.)

In this paper, I concentrate on road transport for simplicity. Environmental and accident objectives that cannot be subsumed under economic efficiency are ignored, and I will only touch on equity issues in passing. My first issue is how the optimal gasoline tax is affected by the inclusion of MCF and ITCF in the economic efficiency objective function. Next, I address the issue of optimal congestion charges and discuss once more an issue that was brought up a long time ago, namely: If optimal congestion charges are used and road capacity is expanded

in an optimal way, will the congestion charges cover the costs of expanding capacity? The answer provided by Mohring (1965) and Strotz (1965) was that, provided there are constant returns in capacity provision, it will. In fact, the revenue from user charges in this case will exactly finance the optimal road capacity expansion. The subsequent discussion (e.g., Keeler and Small (1977), Kraus 1981) concentrated on whether or not there were in fact constant returns to scale. None of this literature took account of the non-transport impacts captured by the MCF and the ITCF. Will their results still hold in this case? My third issue is toll financing on uncongested roads. Taking account of the MCF and ITCF, is there a role for user charges in financing investment in that case?

The paper is organised as follows: In section 2, I introduce the form of ITCF that was derived in Minken (2006) and is currently applied in road cost benefit analyses in Norway. The general form of the welfare function to be optimised is then introduced. In section 3, this welfare function is optimised by setting a gasoline tax (or more generally, a distance dependent tax) in the uncongested case. In section 4, congestion charges and capacity expansion are used to optimise the welfare function in the congested case. This model closely resembles the Mohring model. It turns out that the optimal charges may be divided in two parts. The first part is identical to the optimal gasoline tax. This part captures all the impacts of the MCF and ITCF. Thus the gasoline tax should be kept as it is in the first model. The second part, the congestion charge proper, consists of the marginal cost of congestion plus the marginal collection costs. Under constant returns, the revenue from this part will pay for collection plus about 1.2 times the optimal investment. The 20 per cent more will however not be available for additional projects; it covers losses on non-transport taxes. Thus the result is basically the same as in Mohring (1965). In section 5, we find that there might be a case for tolling on uncongested roads, but probably not if the gasoline tax is set optimally.

The optimal gasoline tax is disturbingly sensitive to parameters that we know too little about. We round off by discussing possible ways of overcoming this predicament.

2 The welfare function

Minken (2006) used a simple general equilibrium model to derive an expression for the MCF in transport cost benefit analyses as well as a rule for entering transport taxes in the "government account". The model assumes constant returns to scale in all production. The representative consumer derives utility from general consumption, leisure and four types of trip (trips to work and leisure trips, each by two modes). She maximises a quasi linear, strongly separable utility function subject to a constraint on expenditure and two time constraints, one involving leisure, work, and work trips and the other involving leisure trips.

Suppose π is the resource price (price net of all taxes) of some input to a trip, such a gasoline; the tax rate on this input is τ; and the average rate of the value added tax on non-transport consumption is s_0. According to this model, the revenue per trip from the tax τ should be entered in the government account as

$$R_1 = \frac{(\tau - s_0)\pi}{1 + s_0} \tag{1}$$

Likewise, suppose t is the time taken by the trip, ω is the cost of labour and s is the proportion of ω that consists of taxes. If on the margin, an increase in the time spent in transport leads to a corresponding decrease in labour time, the loss in tax revenue from an additional trip would be

(2) $$R_2 = -\frac{s + s_0}{1 + s_0}\omega t$$

Since next to nothing is known of the rate of substitution between labour time and trips as trips become more expensive or time consuming, R_2 is ignored in the following.

Like tax revenue, subsidies and grants between the public and the private sector should be divided by $1 + s_0$ before being entered in the government accounts. Contrary to this, the inputs of all public production should be entered with the same prices as is used in the private sector and no further adjustments. Finally, all public revenue and expenditure are to be multiplied by the MCF. These are the rules that follow from the model.

Assume for a start that the generalised cost c of a trip consists of operating cost $(1 + \tau)\pi$ plus time cost vt, where v is the value of time:

(3) $$c = (1 + \tau)\pi + vt$$

π and $1 + \tau$ might be vectors of (perceived) resource prices and tax mark-ups. Ignoring public transport and private toll collection, the welfare function will consist of benefits to users, plus the government budget surplus, minus external costs. Assuming no traffic growth and an infinite time horizon, the welfare function can now be written

(4) $$W = \frac{1}{r}\left\{\int_c^\infty D(y)\,dy + (1 + \lambda)\left(\frac{(\tau - s_0)\pi}{1 + s_0} - m\right)D(c) - kD(c)\right\}$$

where $D(c)$ is the demand function, $1 + \lambda$ is the MCF, m is the marginal cost of road wear and tear pr. trip and k is the environmental and accident costs of a trip. Later, we will expand W with investments, tolls and toll collection costs, but at the moment, this is not necessary.

3 Optimal gasoline tax

As a rough approximation, we may assume that all monetary inputs to a trip except gasoline carry the average value added tax. Thus these other inputs vanish from W, π becomes the net price of gasoline and only the gasoline tax is available as a policy instrument. The optimal gasoline tax is found by maximising W with respect to c and τ subject to (3). Eliminating the Lagrangian multiplier from the two first order conditions and rearranging, we arrive at the following expression for the optimal tax:

(5) $$\tau\pi = s_0\pi + (1 + s_0)\left(m + \frac{k}{1 + \lambda}\right) + \frac{s_0 - \lambda}{1 + \lambda}\frac{D(c)}{D'(c)}$$

Actually, this is not an explicit solution, because demand depends on τ. If we may assume the price elasticity of demand to be constant, the explicit solution is obtained by rewriting the last term and rearranging:

(6) $$\tau\pi = \frac{s_0\pi + (1 + s_0)\left(m + \frac{k}{1 + \lambda}\right) + \frac{s_0 - \lambda}{1 + \lambda}\cdot\frac{\pi}{El_{(1+\tau)\pi}D}}{1 - \frac{s_0 - \lambda}{1 + \lambda}\cdot\frac{\pi}{El_{(1+\tau)\pi}D}}$$

Obviously, the optimal tax does more than just cover the external costs *m* and *k*. As expected, an allowance must also be made for the revenue lost on general consumption. Furthermore, value added tax is applied to the external costs, while the external costs falling on others than the government is adjusted down by the factor $(1 + \lambda)^{-1}$.

The influence of the gasoline elasticity depends on the sign of the factor $(s_0 - \lambda)$. The more positive it is, the more the optimal gasoline tax is reduced. The reduction is less the more inelastic transport is. We might say that in this case, heavy taxation of car use is not a very good idea. Consequently, a good deal of the funds for transport infrastructure will have to come from outside transport. If, on the other hand, $(s_0 - \lambda)$ is negative, the optimal gasoline tax becomes higher, and higher the more inelastic transport is. In this case, the car is a good object of taxation, and a fair amount of funds for transport improvements may come from the gasoline tax.

It is not easy to say which of these cases apply. The reason is that at present, estimates of the MCF are very uncertain.

However, according to the Norwegian Ministry of Finance, MCF is 1.2 ($\lambda = 0.2$). The value added tax on most commodities is 0.25, while food has 0.13 and (importantly) public transport has 0.08. Books, newspapers and some cultural services are exempted from the value added tax. On average, s_0 may be sufficiently close to 0.2 for the whole elasticity term to be ignored. However, if public transport is available as a good alternative to car travel, the factor $(s_0 - \lambda)$ is definitely negative, calling for an upward adjustment of the optimal gasoline tax. This is the case in many urban areas.

The external cost of accidents, air pollution and noise for driving a light vehicle in sparsely populated areas (*k*) may be set to NOK 0.42 per kilometres (Norwegian Railway Authority 2005). These cost are very uncertain. Road wear and tear for light vehicles (*m*) may be set to 0. Fuel consumption per kilometre for light gasoline cars may be set to 0.28 litres, and the production price per litre (π) to NOK 3.89 (Samstad et al 2005). Assuming $s_0 = \lambda = 0.2$, the optimal tax per litre is calculated to be NOK 6.61, producing a total price per litre of NOK 10.50. The actual tax as of 2005 was NOK 8.24 and the full price NOK 12.13.

The long-term elasticity of car kilometres with respect to the fuel price is -0.26 for Norway as a whole (Fridstrøm 1999). In large urban areas this will be higher, so let us assume an elasticity of -0.4. We make a guess at s_0 in this case of 0.18. The external costs will also be higher in large urban areas. According to the Norwegian Rail Authority (2005), they are NOK 0.67 per kilometre, which means NOK 9.38 per litre. This produces an optimal gasoline tax for urban areas of NOK 12.03, nearly two times higher than the result for sparsely populated areas, yielding a full fuel price of NOK 15.92! A caution is in order here: This result is largely driven by two extremely uncertain factors, the factor $(s_0 - \lambda)$ and the external costs.

For technical and equity reasons, it is not recommendable to differentiate the gasoline tax according to the type of region. But the tentative result for the urban area might be kept in mind when, next, we turn to congestion charges.

4 Optimal congestion charges

Let travel time *t* be a monotonously increasing, convex function of the ratio between demand *D* and road capacity *K*, $t = t(D/K)$. There are two time periods, the high demand period and

the low demand period. High demand variables are indexed by *H* and low demand variables are indexed by *L*. Road capacity is provided by government according to the production function $K = f(N)$, where *N* is a basket of inputs whose price is *w*. The gasoline tax is assumed to be fixed. The policy variables are the level of investment in the production of capacity, *N*, and the time-differentiated congestion charges b_H and b_L. To collect these charges, government incurs a collection cost *h* per trip.

The welfare function becomes

$$(7)\quad W = \frac{1}{r}\sum_{j\in\{H,L\}}\left\{\int_{c_j}^{\infty} D_j(y)\,dy + (1+\lambda)\left(\frac{(\tau - s_0)\pi + b_j}{1+s_0} - (m+h)\right)D_j(c) - kD_j(c)\right\} - (1+\lambda)wN$$

The problem is to maximise *W* with respect to c_H, c_L, b_H, b_L and *N* subject to constraints

$$(8)\qquad (1+\tau)\pi + b_j + v_j\, t\left(D_j(c_j)/f(N)\right) = c_j, \quad j = H, L$$

Again, eliminating the Lagrangian multipliers from the first-order conditions and rearranging, we get for *j = H, L*

$$(9)\qquad \tau\pi + b_j = \left\{s_0\pi + (1+s_0)\left(m + \frac{k}{1+\lambda}\right) + \frac{s_0 - \lambda}{1+\lambda}\cdot\frac{D_j}{D_j'}\right\} + \left\{(1+s_0)h + v_j D_j\frac{\partial t}{\partial D_j}\right\}$$

By comparison with (5), we see that if the gasoline tax is set optimally (for the region) the congestion charges in the two periods should cover $(1+s_0)$ times the collection cost plus the marginal congestion costs imposed on others. If not, (6) shows us that there will be an addition Δ,

$$(10)\qquad \Delta = s_0\pi + (1+s_0)\left(m + \frac{k}{1+\lambda}\right) + \frac{s_0 - \lambda}{1+\lambda}\cdot\frac{\pi}{El_{(1+\tau)\pi}D} - \left(1 - \frac{s_0 - \lambda}{1+\lambda}\cdot\frac{\pi}{El_{(1+\tau)\pi}D}\right)\tau\pi$$

So far we have derived expressions for the optimal charges obtaining at the optimal level of road capacity. To be able to determine that optimal road capacity, we need also to utilise the first order condition derived from setting the partial derivative of the Lagrangian with respect to *N* to zero. Having eliminated the Lagrangian multipliers, this equation can be written

$$(11)\qquad (1+s_0)rw\frac{f(N)}{f'(N)} = \left(v_H D_H\frac{\partial t}{\partial D_H}\right)D_H + \left(v_L D_L\frac{\partial t}{\partial D_L}\right)D_L$$

The right hand side of (11) consists of part of the congestion charge revenue, namely the marginal congestion externality of each period (the terms in parentheses) times the demand in each period. (The other parts of revenue from the congestion charge is, as we saw, connected to the collection costs and Δ.) Now if there is constant returns to scale in the provision of capacity ($f(N) = aN$) the left hand side becomes $(1+s_0)rwN$. Except for the value added tax, this is the investment cost. Thus the congestion charges will cover the investment cost (as in Mohring (1965)), but with something to spare. To be more exact, it will cover the investment cost and the collection cost and the loss of tax revenue caused by drawing resources from the private to the public sector.

5 Toll financing

Reverting to the uncongested case, we ask if user charges have a role to play in the financing of new infrastructure in that case. The level of investment has already been decided upon and is not an issue, we assume. The impact of the investment is to reduce generalised cost from c_0 to c_1. The only remaining question is how best to finance it, by tolls or by taxes. In conformity with Norwegian regulation, we assume tolling is only allowed for a maximum of T_{max} years and can not go on after the infrastructure has been financed. For simplicity, we ignore the possibility of starting tolling before the infrastructure has been built. Toll collection costs consist of a fixed cost h_0 to set up the system plus a cost h per passage. Let a be the proportion of investment costs that is financed by tolls, and let δ be 1 if a toll system is established, 0 otherwise.

Contrary to the previous example, we assume that the tolls are collected by a non-profit private firm who incurs the investment cost and collection costs and gets partially refunded by government. The remainder of its income is the toll revenue. We want to use the policy instruments a, B and T (all positive) to maximise the function W:

$$W = W(a,B,T) =$$

$$(12)\qquad \left\{ \int_{c_1+B}^{c_0} D(y)\,dy + (1+\lambda)\left(\frac{(\tau - s_0)\pi}{1+s_0} - m \right) D(c_1+B) + (B - \delta h - k) D(c_1+B) \right\} \cdot \int_0^T e^{-rt} dt$$

$$-(1+\lambda)(1-a) I - (aI + \delta h_0)$$

subject to constraints

$$(13)\qquad (B - \delta h) D(c_1+B) \cdot \int_0^T e^{-rt} dt = aI + \delta h_0$$

$$(14)\qquad T \le T_{\max}$$

$$(15)\qquad a \le 1$$

Assuming for the moment that $a > 0$, the Kuhn-Tucker conditions for a solution can be written

$$(16)\qquad (\lambda + \eta_1) I = \eta_3$$

$$(17)\qquad \left[(R-k) + (1-\eta_1)(B-h) \right] D'(c_1+B) = \eta_1 D(c_1+B)$$

$$(18)\qquad \left\{ \int_{c_1+B}^{c_0} D(y) + \left[(R-k) + (1-\eta_1)(B-h) \right] D(c_1+B) \right\} e^{-rt} = \eta_2$$

where a slash denotes differentiation, η_1, η_2 and η_3 are the Lagrangian multipliers connected to constraints (13), (14) and (15) respectively, and R is defined by

$$(19)\qquad R = (1+\lambda)\left(\frac{(\tau - s_0)\pi}{1+s_0} \right) - m$$

It turns out that there are only three candidate solutions, since the case of both (14) and (15) non-binding produces a contradiction. Thus if $a^* < 1$, we must have $T^* = T_{\max}$, but if $a^* = 1$,

i.e. 100 per cent toll finance, T^* could be equal to T_{max} or less. The parameters of the problem decide which candidate is the best.

Suppose now that an optimal solution for the $a > 0$ case has been found. To test if toll financing should be used at all, we want to compare the solution with the $a = 0$ case. The case of $a = 0$ entails $B = 0$ by (13). Formally, W in this case is maximised by setting $T = T_{max}$. But since this is the case of 100 per cent tax finance, the annual benefits are unaffected by the length of the tolling period. To compare this candidate solution with other candidates, we may use the fact that the two candidates are equal after the end of the tolling period and look only at the difference $W\left(a^*, B^*, T^*\right) - W\left(0, 0, T^*\right)$, where the starred variables are the optimal values according to the other candidate.

We get

$$W\left(a^*, B^*, T^*\right) - W\left(0, 0, T^*\right) =$$

$$(20)\quad \left\{ -\int_{c_1}^{c_1+B^*} D(y)\,dy - (R-k)\left(D(c_1) - D\left(c_1+B^*\right)\right) + \left(B^* - h\right) D\left(c_1+B^*\right) \right\} \int_0^{T^*} e^{-rt} dt$$

$$+\lambda a^* I - h_0$$

Now by (13), the time integral may be eliminated. Rearranging, we get:

$$W\left(a^*, B^*, T^*\right) - W\left(0, 0, T^*\right) > 0$$

$$(21)\quad \Leftrightarrow$$

$$\frac{\int_{c_1}^{c_1+B^*} D(y)\,dy + (R-k)\left(D(c_1) - D\left(c_1+B^*\right)\right)}{\left(B^* - h\right) D\left(c_1+B^*\right)} < (1+\lambda)\frac{a^* I}{a^* I + h_0}$$

The nominator on the left hand side of (21) is the annual welfare loss of travellers and lost government tax revenue due to tolling, minus the external costs reduction caused by tolling. Put briefly, the sum of these is the disbenefit of tolling. The denominator is annual net toll revenue. Obviously, for tolling to be a good idea, the left hand side should be small. If (21) does *not* apply, toll financing should not be used. It appears that tolling might sometimes be a good solution even in uncongested cases, in particular if demand is rather inelastic, transport taxes are low compared to the external costs, the MCF is high or the fixed cost of toll collection is low. However, by (5) and (19), it might be argued that if the gasoline tax has been set optimally, the term involving $R - k$ in (21) is proportional to the notoriously uncertain and possibly very small term $\left(s_0 - \lambda\right) \cdot \left(1 + s_0\right)^{-1}$. Since the integral is at least as large as the denominator, it would be hard to argue for toll financing in that case.

In a more simplified framework ignoring externalities, fixed collection costs and the effects of transport policy on general government tax revenue, Larsen (1986) and Ramjerdi (1995) have derived similar results with respect to optimal tolls. Ramjerdi (1995) derives an expression for the optimal toll in terms of "the marginal cost of toll financing", which is basically the terms in curly brackets in (20) divided by net toll revenue.

Others have suggested that tolling should be chosen if and only if $h/B < \lambda$. This is generally not correct, as we can see if we rearrange the inequality of (21) further to get

$$\frac{h}{B} < 1 - \frac{a^* I + h_0}{(1+\lambda) a^* I} \cdot \frac{\int_{c_1}^{c_1+B^*} D(y)\,dy + (R-k)\left(D(c_1) - D(c_1+B^*)\right)}{B^* D(c_1+B^*)} \tag{22}$$

6 Discussion

The discomforting feature of the optimal gasoline tax is its high degree of sensitivity to factors about which we know too little. Actually, it seems almost immoral to ask millions of people to pay such large sums with so scant justification. The problem is not just that we know too little, it is also that the model may be too simple, and that a more elaborate model might be more robust with respect to the uncertain factors. Models like the one applied here have served economists well for decades in arguing for charges that internalise the external costs connected to accidents, air pollution, noise and congestion. Recently, their insistence has paid off well with the successful congestion charging schemes of London and Stockholm. But obviously, the inclusion of a marginal cost of funds and the effects of transport-induced shifts in the consumption pattern on government revenue is more than this simple conceptual model can bear.

A more elaborate model might include car ownership choices and provide a role for car taxes. The model would also benefit from the inclusion of mode choice. Thus as road transport becomes more costly, the rise in the own price elasticity would act as a brake. More fundamentally, we need to know more about the interaction between the labour market and transport markets. Most fundamentally, the uncertainty surrounding the MCF needs to be reduced.

7 Conclusion

The optimal gasoline tax and optimal congestion charges and tolls have been derived from a simple conceptual model where a marginal cost of funds and the effects of transport-induced shifts in the consumption pattern on government revenue have been included. The results appear to be simple modifications of old wisdom. However, the optimal gasoline tax is extremely sensitive to very uncertain parameters. A more elaborate model, giving the travellers more options, is called for if the MCF and ITCF are to be used for optimisation, and more needs to be known about the size of MCF and the labour market – transport market interaction.

References

Fridstrøm, L. (1999) *Econometric models of road use, accidents, and road investment decisions. Volume II.* TØI report 457/1999.

Keeler, T.E. and K.A. Small (1977) Optimal peak-load pricing, investment, and service levels on urban expressways. *Journal of Political Economy* **85**(1), 1-25.

Kraus, M. (1981) Scale economies analysis for urban highway networks. *Journal of Urban Economics* **9**, 1-22.

Larsen, O.I. (1986) Bompenger som finansieringsform. *Sosialøkonomen* **40**(4), 9-11.

Norwegian Rail Authority (2005) Samfunnsøkonomiske analyser for jernbanen. Metodehåndbok JD 205.

Minken, H. (2006) Transport cost benefit rules.

Mohring, H. (1965) Relation between optimum congestion tolls and present highway user charges. *Highway Research Record* 47, 1-14.

Ramjerdi, F. (1995) *Road Pricing and Toll Financing*. Essay no 3. Ph.D dissertation at the Royal Institute of Technology, Stockholm.

Samstad, H., M. Killi and R. Hagman (2005) *Nyttekostnadsanalyse i transportsektoren: Parametre, enhetskostnader og indekser*. TØI-rapport 797/2005.

Strotz, R.H. (1965) Urban transportation parables. In Margolis, J. (ed.) *The public economy of urban communities*, 127-169. Resources for the Future, Washington D.C.

2.4 Systematic risk, and how it is taken account of in the discount rate of Norwegian transport cost benefit analyses

Systematic risk, and how it is taken account of in the discount rate of Norwegian transport cost benefit analyses[10]

Harald Minken

Institute of Transport Economics, Oslo

Contents

ABSTRACT

According to a regulation from the Norwegian Ministry of Finance, the discount rate to use in cost benefit analyses should consist of a risk-free rate plus a risk premium. The covariance of project returns with the returns on all national assets should determine the size of the risk premium. Subject to a few general rules, the sector authorities are permitted to apply the regulation to the specific circumstances in their sector. The Ministry of Transport and Communications commissioned the Institute of Transport Economics to suggest rates to use in transport. Based on a general model of when to invest in non-tradable assets (Lund 1993), the present author developed a criterion for investments under uncertainty in the transport sector and used it to estimate and propose discount rates for road, rail, air and coastal services. With a minor exception, the proposal was accepted by the Ministry and implemented from 2006.

This paper treats the general principles, develops the investment criterion and briefly touches on estimation issues.

[10] Paper presented at the ECTRI TWG G on Transport Economics and Policy, Madrid September 2008.

1 Background

A transport cost benefit analysis (CBA) of infrastructure improvements consists of an initial investment cost and a number of future annual net benefits. Typically, the annual benefits are computed from some transport model. The transport model takes as input a set of exogenous variables, such as

- Demographic forecasts and assumptions on land use and location of residential areas, workplaces and other attraction factors,
- Income, which in turn influences car ownership rates and other factors such as values of time and other non-market costs,
- Vehicle fuel efficiency,
- National tax policy, the price of oil and other factors influencing market prices,
- The coded infrastructure measure itself and its consequences for volume-delay functions and other level-of-service characteristics, including public transport supply characteristics.

All of the exogenous factors are – in different ways and to a different degree – uncertain. Furthermore, there is uncertainty about the goodness of the transport model itself and the methodological principles of the CBA. Needless to say, investment costs are also uncertain. So how do we, as transport economists, handle uncertainty? I am afraid that basically, we limit ourselves to a few sensitivity tests and a final warning that all future predictions are uncertain. Could we do better?

For a start, a distinction should be made between risk and uncertainty. Risk is the probability that the realised outcome deviates from the expected value. Thus when we talk about risk, we are able to assess (probably subjectively) the probability of different outcomes and to compute the expected value, standard deviation etc. Uncertainty in a deeper sense is when even this is impossible. A first move would be to require, as far as possible, all of our exogenous variables, unit costs etc. to be expected values. For instance, there exists some knowledge on how values-of-time evolve with income, and even if official forecasts of income growth are not usually presented with probabilities attached to a worst case, a best case and a most probable case, it might be possible to infer subjective probabilities from comments in official documents and to compute an expected value of time for future years. That would be much better than to keep the value of time constant (which is very improbable), or to assume that it develops according to the most probable income growth forecast (which is inconsistent with treating income as a stochastic variable later on).

It will be impossible or very impractical to treat the mechanisms of the transport model, the methodological issues surrounding the CBA and the decisions of the national government as stochastic variables in the same way. Uncertainty about such issues must be dealt with by sensitivity analysis and general comments. But with some effort, it might be possible to establish demographic growth, income growth, fuel efficiency and international prices as stochastic variables, and to use expected values for these variables and the variables that depend on them as a starting point for assessing risks.

The relevant risk is systematic risk

Next, we should be clear about what risks are relevant to the evaluation of a public sector project. In a famous article, Arrow and Lind (1970) finds that as the risk of a public project is

spread out among very many taxpayers, it becomes negligible for each and every one of them, and therefore for society. Consequently, governments should be risk neutral. But this result depends on the assumption that the returns of the project have zero covariance with national income. Thus it does not really contradict the findings in another strand of literature, which shows that the relevant measure of risk in public projects is based on the covariance of the project returns with national income. To the extent that this covariance is not zero, governments should be risk-averse, like the individual members of society. In fact, to achieve efficiency in production, governments should price risk in the same way as private firms do (Sandmo 1972). Otherwise, public sector projects might crowd out more profitable private projects.

The principle that what matters is not the riskiness of the project as seen in isolation (the standard deviation of returns) but the contribution it gives to the riskiness of all sources of income (the covariance) is known as the portfolio principle. In finance theory, private investors are assumed to follow this principle. They have different levels of risk aversion, but taking the view that what matters is the volatility of their total portfolios, not the single assets, they will want to diversify their holdings to get rid of *unsystematic risk*, i.e. risk that the market will not pay them to hold. Under rather strict assumptions, an equilibrium can then be shown to occur, where all investors hold a mix of a risk-free asset and the market portfolio, i.e. a portfolio consisting of all risky assets in proportions equal to their respective total market value at equilibrium. Since the market at equilibrium does not pay investors for risks that can be diversified away, but only the risk remaining after optimal diversification (*systematic risk*), their different attitudes to risk is not reflected in different compositions of their portfolio of risky assets, but in the mix of risky assets (the market portfolio) and risk-free assets.

This is the so-called Capital Asset Pricing Model (CAPM). According to CAPM, the expected rate of return on an asset at equilibrium consists of a risk-free rate plus the risk premium of the market portfolio times a variable called *beta*, which is the covariance of the asset with the returns on the market asset, normalised by dividing by the standard deviation of the market portfolio. Let $E(R_i)$ be the expected rate of return on asset i, R_f be the return on the risk-free asset and R_m the expected return on the market portfolio. Furthermore, let σ_{im} be the covariance of asset i and the market portfolio and σ_m^2 be the variance of the market portfolio. The CAPM equation is then

$$\begin{aligned} E[R_i] &= R_f + \beta_i \cdot \left(E[R_m] - R_f \right) \\ &= R_f + \frac{\sigma_{im}}{\sigma_m^2} \left(E[R_m] - R_f \right) \end{aligned} \qquad (1)$$

In (1), the second line defines beta. The whole of the second term on the right-hand side is called a risk premium, while $E[R_m] - R_f$ is called the market premium or the market price of risk. Beta is then a measure of the amount of risk connected with investments in object i.

The price of the asset at equilibrium will be such that investors will require an expected return like (1) to hold the asset, and they will all actually hold the asset as part of their portfolios and expect this return to materialise. Thus according to CAPM, the relevant risk for pricing a risky asset is the systematic risk as expressed by the beta, and the cost of providing more equity capital to the firm i will be $E[R_i]$.

The relevance for public project evaluation

If efficiency in production is as important for public project evaluation as suggested by Diamond and Mirrlees (1971) or Sandmo (1972), we should use any information on how the private sector prices risk to price it the same way in our CBA. Actually, the private sector uses the CAPM betas to assess the risks of assets, and betas, computed from day-to-day stock price variations, are often reported in the financial papers. Furthermore, a theoretical result in Sandmo (1972) suggests that is will always be possible to construct a portfolio of privately held assets that exactly reproduces the volatility of a public project.

However, the task is more difficult than it seems. The expected risk premium of the market portfolio is changing over time and might not be the same in the future as in the past. Lately, the private "price of risk" has gone up due to the uncertainties surrounding the whole financial system. What we want for project evaluation is a long-term price of risk, which by the nature of things must be derived from long-term data. We will also have to modify the market risk premium of the stock exchange to take account of the fact that the average private project uses a mix of equity and loan capital, and to take account of tax rules.

Next, it will not do to just take over the betas of construction firms and apply them to the cost of construction, or the betas of firms active in the transport business and apply them to the annual benefits of projects in the different sectors of transport. Construction firms share risks in the construction phase with their principals, the transport authorities, and therefore have a different beta than the one facing society as a whole. The observed betas depend on the regulatory regime (Alexander et al 1999), while the beta we seek should cover risks irrespective of who bear them. Likewise, transport firms, or even private infrastructure holders, if they exist, capture only parts of the annual benefits and bear only parts of the annual costs for society as a whole. It may even be the wrong parts, as when a private airport gets most of its income from renting out shopping space.

At a deeper level, the public projects we are evaluating will not produce returns (in the form of annual net benefits) that are tradable or can easily be captured by any one agent. Roads and other infrastructure are not traded in markets, and even if they were, only a part of the benefits and costs would flow to the infrastructure owner, while the most important parts, the consumer surpluses, will remain in the hands of the same old users. The problem is that diversification on the part of those who reap the benefits and bear the costs is largely impossible. The analogy with stocks and other tradable commodities is a weak one.

Infrastructure is not the only type of assets that cannot be traded. For applications to society as a whole, perhaps the biggest problem is human capital. The stream of wage income that an individual can expect to earn over a lifetime cannot be sold to others, since slavery is forbidden. The most important source of income to most individuals, and consequently to society, is a non-tradable asset. Most individuals will *not* be able to choose a mix of risk-free assets and the market portfolio, and therefore society as a whole retains considerable unsystematic risk.

Enough is said by way of introduction to see that

- We need a more systematic approach to uncertainty and risk in CBA,
- We should use expected values as far as possible, single out systematic risk for quantification and valuation and let the law of large numbers take care of unsystematic risk,

- Sensitivity analysis and scenario analysis are of help in identifying key sources of uncertainty and risk, but since they do not attach probabilities to outcomes and do not consider covariances with other sources of national income and welfare, they cannot be used to quantify and value systematic risk,
- Including a risk premium in the discount rate and using the CAPM as a framework for quantifying and valuing the risk is a possibility, but it is not without problems and weaknesses,
- Nevertheless, some form of covariance measure of systematic risk is obviously called for.

The rest of the paper is organised as follows: Section 2 sketches the approach to risk in Norwegian CBA, as set out in official guidance from the Ministry of Finance. Section 3 develops the model used to determine risk premia in Norwegian transport CBA, and Section 4 treats estimation of the model. Finally, Section 5 concludes with some practical advice.

2 The Norwegian approach to risks in CBA

In September 2005, the Ministry of Finance in Norway issued an official guidebook for cost benefit analysis (Finansdepartementet 2005) and a two-page regulation (Rundskriv R-109/2005) outlining the few things that they wanted all analyses to comply with. The regulation treats the discount rate and calculation prices. The discount rate is to consist of a risk-free rate and a risk premium. Alternatively, risk can be handled by adjusting annual net benefits down to the safety equivalents and discounting by the risk-free rate.

The latter method is to be used for large projects and projects where there are clear milestones, i.e. points where risk is dissolved and the risk profile takes on a new form. According to the guidebook, the method of safety equivalents gives a clearer view of how much (of the discounting) is due to risk and how much is due to time. But also according to the guidebook, the two methods coincide if the risk is the same in every period. In fact, their proposed method of computing safety equivalents is to reduce annual net benefits by the factor

$$a_t = \left(\frac{R_f}{E\left[R_j^t \right]} \right)^t$$

there t is time and the other variables are defined in connection with equation (1). This can only produce a result different from discounting with $E[R_j]$ if R_j^t changes with time. If this is all there is to computing safety equivalents, we might as well include the risk premium in the discount rate. For the few instances where there is a milestone, we simple discount the years before and after the milestone with different risk premia in the discount rate.

The regulation defines the risk-free rate as the long-term risk-free real interest rate before tax, and sets it at 2 per cent. The Ministry reserves the right to adjust this rate if evidence shows that it is changing.

The risk premium should depend on systematic risk, the regulation states. It proceeds by indicating the appropriate level of the risk premium for different types of project. The normal public project is thought to be only mildly dependent on the ups and downs of the economy, and the risk premium is set at 2 per cent (a discount rate of 4 per cent in all). Public sector companies that compete with private production are to use the same risk premium as their

competitors. This clearly reflects a concern for efficiency in production, and is consistent with the general principle that market prices should be used as far as possible in a CBA. If there is considerable systematic risk, the regulation suggests to use a discount rate of 6 per cent, which is consistent with their estimate of the price of risk for the average Norwegian firm (4 per cent).

Finally, the regulation states that groups of projects with similar risk could set their risk premium based on a special study of their circumstances. This is an opening for the various sector authorities to set their own risk premium. It is indicated that the CAPM might be a useful tool for such studies.

It should be noted that a similar regulatory note was issued as early as 1999, but without a guidebook. To the extent that practitioners took notice at all, a variety of different interpretations and practices developed.

Commissioned by the Ministry of Transport and Communications in 2005, the Institute of Transport Economics studied how to apply these regulations and guidelines in the transport sector. It was decided not to use the CAPM, both because it seemed impossible to identify private sector companies with a risk profile similar to the average transport project, and because infrastructure is a non-tradable asset, with little similarity to stocks that can be bought and sold all the time. Instead, a special model was developed (Minken 2005). That model is the subject of the next section.

3 A model to determine risk premia in transport sector CBA

Our model builds on a similar model in Lund (1987, 1993). A few additional assumptions and formal changes have been made. There is a single decision-maker who at the start of period 0 already finds himself with a certain amount of two risky non-tradable assets, infrastructure and human capital. These assets he cannot sell or get rid of in any way. He also owns given initial amounts of three tradable assets, one of them risk-free and the two others risky. He works and consumes in period 0 and 1, and in period 0, he has the opportunity to buy or sell the three tradable assets. In addition, he faces the possibility of acquiring *more* of the two non-tradables, i.e. to use some time in period 0 to go to school or to build more infrastructure for himself. The latter option will be used if the price is low enough and the returns high enough. The point of the model is to derive a criterion for when it pays to acquire more of these two assets.

Let object 0 be the risk-free object, objects 1 and 2 stocks or other risky paper, object 3 human capital and object 4 infrastructure. We measure the amount of investments in physical units, i.e. the number of government bills (the risk-free object), the number of stocks, the number of hours of education and the number of minutes saved on an average trip. Note especially that infrastructure is measured by its effect.[11]

It might reasonably be asked how it is possible for a private person like our decision-maker to invest in infrastructure. A possible answer might be that there is a government who makes this decision and sends the bill to consumers. The government makes it decision with full

[11] There is no uncertainly with respect to the effects of new infrastructure, but only with respect to how it is valued by the user and how many times he will use it. Or if there is uncertainly about the effect, it is unsystematic.

information about the willingness of consumers to pay for infrastructure, and the consumers are fully aware of the governments's intentions and adapt accordingly.

All objects except infrastructure have given market prices. The infrastructure case is a little bit more complicated. Presumably, there exist several plans for infrastructure improvement, each with their particular price as measured by the investment cost necessary to achieve one minute of travel time savings. We must assume that all the cheapest ways of saving a minute has been realised before period 0, but that there remains a set of investment opportunities with approximately the same cost of saving a minute of travel time. The will also be more expensive investment opportunities – these can be disregarded. The price of object *j* in period 0 is P_{j0}.[12] The amount invested in object *j* in period 0 is X_j.

Let $\overline{X}_j$ denote the initial stock of object no. *j*. While amounts to be invested in the risk-free object X_0 and the risky objects X_1 and X_2 can be chosen freely, irrespective of initial stocks, the decision-maker is stuck with the initial stocks of human capital and infrastructure, i.e. $X_3 \geq 0$ and $X_4 \geq 0$. No such restriction exists with respect to X_0, X_1 and X_2.

The consumer has non-wage income *M* after investing and a wage income of *wL*, where *w* is the hourly wage (we ignore taxes) and *L* the number of working hours in the period. He spends his income on *n* different goods and services plus a number of kilometres of transport. Let p_i be the price and y_i the amount of good no. *i*, and let *p* without subscript be the transport cost per kilometre. The number of kilometres is *a*. Thus his expenditure is:

$$\sum_i p_i y_i + pa = M + wL \tag{2}$$

If total hours in the period is $\overline{t}$, leisure time is *z* hours and the time per kilometre in transport (the inverse of speed) is *t*, the time budget becomes

$$z + L + ta = \overline{t} \tag{3}$$

Using (3) to substitute for *L* in (2), we get:

$$\sum_i p_i y_i + (p + wt)a + wz = M + w\overline{t} = C \tag{4}$$

Note that full time income (if all of the leisure time was used for work) is called *C* in (4). Based on (4) we can write the indirect utility function of the consumer as $V(\mathbf{p}, p + wt, w, M + w\overline{t})$, where $\mathbf{p} = (p_1, ..., p_n)$.

Subscripts 0 and 1 on the variables denote the period. Consider first his utility in period 0. We assume that at the start of that period he has a wealth W_0 which he uses for investing, so that what remains for consumption is C_0. The investment in human capital takes the special form allocating hours to education, and therefore of a modification of the time budget (3). Consequently,

$$C_0 = W_0 - \sum_{j=0,1,2,4} P_{j0} X_j + w_0 (\overline{t} - X_3) \tag{5}$$

and

[12] P_{30} will be called w_0.

(6) $$V_0 = V\left(\mathbf{p}, p + w_0 t_0, w_0, W_0 - \sum_{j=0,1,2,4} P_{j0} X_j + w_0\left(\overline{t} - X_3\right)\right)$$

In period 1, which is the last one, the consumer lives off his wealth, but continues to work and to travel as before. The price of the tradable objects in period 1 is P_{j1}, $j = 0,1,2$. Apart from the risk-free object, this price is stochastic. Thus he sells all his tradable assets and receives the sales proceeds. At the same time, he reaps the returns from his education in the form of a higher wage rate, the returns on the infrastructure investment in the form of reduced travel time. His wage rate in period 1 is w_1:

(7) $$w_1 = w_0 + bX_3$$

Travel time is t_1:

(8) $$t_1 = t_0 - cX_4$$

b is a stochastic variable, while c is a constant. It is the value of c to the consumer – the value of time – that is uncertain, not c itself. The uncertainty of the value-of-time stems from the uncertainty of the period 1 wage rate, which depends on the uncertain returns on investment in human capital (see (7)). Now we have:

(9) $$C_1 = \sum_{j=0}^{2} P_{j1}\left(X_j + \overline{X}_j\right) + \left(w_0 + bX_3\right)\overline{t}$$

and

(10) $$V_1 = V\left(\mathbf{p}, p + \left(w_0 + bX_3\right)\left(t_0 - cX_4\right), w_0 + bX_3, \sum_{j=0}^{2} P_{j1}\left(X_j + \overline{X}_j\right) + \left(w_0 + bX_3\right)\overline{t}\right)$$

The investment problem

The consumer maximises expected utility over the two periods. Let θ be a given utility discount factor. The problem is:

$$\underset{X_0, X_1, X_2, X_3, X_4}{Max}\ U = V_0 + \theta E\left[V_1\right] \quad \text{gitt } X_3 \geq 0, X_4 \geq 0$$

where V_0 and V_1 are given by (6) and (10).

Optimality conditions for the tradable assets

For j = 0, 1, 2 the first-order conditions for a maximum are

(11) $$\frac{\partial V_0}{\partial C_0}\left(-P_{j0}\right) + \theta E\left[\frac{\partial V_1}{\partial C_1} P_{j1}\right] = 0$$

Define the returns on object j, R_j, by $R_j \equiv P_{j1}/P_{j0}$. By the rule E[XY] = cov(X,Y) + E[X]E[Y], (11) may then be transformed to

(12) $$\frac{\partial V_0}{\partial C_0} = \theta \operatorname{cov}\left(\frac{\partial V_1}{\partial C_1}, R_j\right) + \theta E\left[\frac{\partial V_1}{\partial C_1}\right] E\left[R_j\right]$$

For object 0, the covariance of (12) is 0 and $E\left[R_0\right] = R_0$, which gives us:

(13) $$R_0 = \frac{\partial V_0 / \partial C_0}{\theta E\left[\partial V_1 / \partial C_1\right]}$$

Since we assume unrestricted lending and investment in the market for risk-free assets, the subjective substitution rate R_0 will necessarily equal 1 + the market interest rate for risk-free assets for all consumers, including the one we consider. Thus we interpret R_0 as 1 + the risk-free interest rate in the market.

Using (13) in (12) and rearranging, we get for j = 1, 2:

(14) $$E\left[R_j\right] - R_0 = -\frac{\operatorname{cov}\left(\frac{\partial V_1}{\partial C_1}, R_j\right)}{E\left[\frac{\partial V_1}{\partial C_1}\right]}$$

The consumer experiences the left-hand side as exogenously given and adapts the right-hand side accordingly. Thus investments in object j should be undertaken until the difference between expected returns and the returns on the risk-free object equals the right-hand side. Assuming C_1 and R_j are both normally distributed, we may use Stein's lemma (see for instance Lund 1993) to give the right-hand side a more suitable form. Stein's lemma says that provided X and Y are both normally distributed and the function $g(.)$ is bounded,

(15) $$\operatorname{cov}\left(g\left(X\right), Y\right) = E\left[g'\left(X\right)\right]\operatorname{cov}\left(X, Y\right)$$

By Stein's lemma, then:

(16) $$E\left[R_j\right] - R_0 = -\frac{E\left[\frac{\partial^2 V_1}{\partial C_1^{\,2}}\right]}{E\left[\frac{\partial V_1}{\partial C_1}\right]}\operatorname{cov}\left(C_1, R_j\right)$$

The first factor on the right-hand side of (16) – everything but the covariance – is the so-called absolute measure of risk aversion. If it is zero, we say that the consumer is risk neutral. In that case, he will not require any extra returns on risky objects. The required return is higher the more risk averse the consumer is. Furthermore, we see that the required return is higher if the covariance between the returns on the object and the total expenditure budget is large. *It is the covariance with total disposable income, not the variance, that matters.*

Aggregating over all risky tradable objects (here: objects 1 and 2) we may get from (16) to a generalised CAPM model for tradable assets in the case where there also exist non-tradable assets (see Lund 1993). Here, we will only perform the first step in such a process, since our interest is in the non-tradable objects. We define R_m as any weighted average of R_1 and R_2:

(17) $$R_m \equiv \alpha_1 R_1 + \alpha_2 R_2, \quad \alpha_1 + \alpha_2 = 1$$

Summing (16) over j = 1,2 with the use of these weights:

(18) $$\sum_{j=1}^{2} \alpha_j \left(E\left[R_j\right] - R_0\right) = -\frac{E\left[\frac{\partial^2 V_1}{\partial C_1^{\,2}}\right]}{E\left[\frac{\partial V_1}{\partial C_1}\right]}\sum_{j=1}^{2} \alpha_j \operatorname{cov}\left(C_1, R_j\right)$$

If we use the rules for computing expectations on the left-hand side and the rules for covariances on the right-hand side, (18) becomes

$$E\left[R_m\right]-R_0=-\frac{E\left[\frac{\partial^2 V_1}{\partial C_1^{\,2}}\right]}{E\left[\frac{\partial V_1}{\partial C_1}\right]}\operatorname{cov}\left(C_1,R_m\right) \tag{19}$$

(19) gives us an opportunity to replace the absolute risk aversion coefficient in (16) with the left-hand side of (19) divided by the covariance on the right. Doing this, (16) becomes:

$$E\left[R_j\right]-R_0=\frac{\operatorname{cov}\left(C_1,R_j\right)}{\operatorname{cov}\left(C_1,R_m\right)}\left(E\left[R_m\right]-R_0\right) \tag{20}$$

Let us stop to consider what we have done so far. (20) is nothing but a transformed version of the optimality condition (11), which is one of the conditions that must be met if the consumer is to get as much as he can out of his investments. The transformation is valid if Stein's lemma can be used and when R_0 and R_m are defined as we have done. We argued that R_0 as defined by us may be understood as the market's rate of interest on risk-free objects. We may now also reinterpret our definition of R_m as the rate of return on the market portfolio or the rate of return on the stock exchange index. Just choose the weights α_1 and α_2 such that the weight of each asset corresponds to the share this asset has in the total value of stocks on the exchange.[13]

Formally, (20) resembles very much the CAPM equation (1), but instead of the beta of CAPM, defined as the covariance between R_j and R_m divided by the variance of R_m, we have the proportion between two covariances, each involving consumption opportunities in period 1, C_1. In (20), the left-hand side is still exogenous, while the CAPM left-hand side is the result of a market equilibrium. What CAPM does is to show what will be the result in the market when many consumers all adapt according to (20) and no income from non-tradable objects enters C_1. The CAPM equation is not valid for non-tradable objects. As already stated, we are not going to derive the generalised CAPM, but will stay content with (20). Our purpose is to show that something weaker, namely (20) with inequality, can be derived for the non-tradable objects, and use that to estimate "transport betas".

Optimality conditions for the non-tradable objects

In equation (11) we gave the three first order conditions for optimum with respect to X_0, X_1 and X_2. We continue with the Kuhn-Tucker conditions for X_3 og X_4. Since they must be positive, these conditions have the form of unequalities.

[13] For R_m to be interpreted as the returns on the market portfolio, all papers on the stock exchange must be included. In our model, we assume that the stock exchange deals in only two risky assets, but that could obviously have been expanded to any number.

$$\frac{\partial U}{\partial X_3} = \frac{\partial V_0}{\partial C_0}\left(-w_0\right) + \theta E\left[-\frac{\partial V_1}{\partial C_1}\left(bt_1a_1 + bz_1 - b\bar{t}\right)\right]$$
$$= -w_0\frac{\partial V_0}{\partial C_0} + \theta E\left[\frac{\partial V_1}{\partial C_1}bL_1\right] \le 0 \quad (= 0 \text{ for } X_3 > 0) \tag{21}$$

Roy's identity has been used in the first line. The equality sign in the second line follows from (3). Now define

$$R_3 \equiv \frac{bL_1}{w_0} \tag{22}$$

The interpretation of R_3 is natural – we see that R_3 is the relative increase in labour income in period 1 per hour of education in period 0. With this definition, (21) gets the same form as (11). Except for the equality sign, the further derivation in this case follows the same steps as the derivation from (11) to (20). We end up with

$$E\left[R_3\right] - R_0 \le \frac{\operatorname{cov}\left(C_1, R_3\right)}{\operatorname{cov}\left(C_1, R_m\right)}\left(E\left[R_m\right] - R_0\right) \tag{23}$$

Note that if the right-hand side is too large in optimum, it does not pay to invest in human capital.

Finally, we treat the case of infrastructure investment. The Kuhn-Tucker condition is

$$\frac{\partial U}{\partial X_4} = \frac{\partial V_0}{\partial C_0}\left(-P_{40}\right) + \theta E\left[-\frac{\partial V_1}{\partial C_1}a_1\left(-w_1c\right)\right]$$
$$= -P_{40}\frac{\partial V_0}{\partial C_0} + \theta\left\{\operatorname{cov}\left(\frac{\partial V_1}{\partial C_1}, w_1a_1c\right) + E\left[\frac{\partial V_1}{\partial C_1}\right]E\left[w_1a_1c\right]\right\} \le 0 \quad (= 0 \text{ for } X_4 > 0) \tag{24}$$

In the second line of (24), the rule for computing covariances has already been applied. In the same way as before, we define the returns on infrastructure investment, R_4:

$$R_4 \equiv \frac{w_1a_1c}{P_{40}} \tag{25}$$

Here, both the level of travel activity a and the period 1 value-of-time w_1 are stochastic variables, while c/P_{40} is a constant that is difficult to estimate.

R_4 is the returns on infrastructure investment, but at the same time, it can also be seen as a benefit-cost ratio (BCR), with travel time savings in the nominator and the investment cost in the denominator. The only difference is that except that the saved travel time in the nominator is not discounted. Anyhow, the same derivation from (11) to (20) again gives us the resulting inequality:

$$E\left[R_4\right] - R_0 \le \frac{\operatorname{cov}\left(C_1, R_4\right)}{\operatorname{cov}\left(C_1, R_m\right)}\left(E\left[R_m\right] - R_0\right) \tag{26}$$

Thus there will be no investment in infrastructure unless expected returns from the best available projects are large enough to make (26) an equality at optimum.

An investment criterion

We might call (26) with strict inequality a non-invest criterion. It is natural to believe that if we reverse the inequality sign, we have an investment criterion. Still following Lund (1993), we will show this to be the case.

Suppose the consumer has solved the investment problem with the result that there is not going to be any infrastructure investment, i.e. $X_4 = 0$. but subsequently, a new investment opportunity appears, this time with a price lower than P_{40}. This very realistic in the case of transport, where new project proposals are developed all the time. Some of them will be more efficient in saving time than the estimated P_{40} euros per minute of time savings.

The new project will have to be financed by a reduction in W_0 of the first period (see equation (5)). Suppose the new cost per unit of time savings is I, so that the reduction in W_0 amounts to IX_4. the project is profitable provided the loss of utility of the reduction in W_0 is compensated by a gain in utility in period 1, or

$$-I\frac{\partial V_0^*}{\partial W_0}+\theta E\left[\frac{\partial V_1^*}{\partial X_4}\right]\geq 0 \tag{27}$$

Using (6), (10) and Roy's identity, we see that (27) can be written

$$-I\frac{\partial V_0^*}{\partial C_0}+\theta E\left[\frac{\partial V_1^*}{\partial C_1}w_1a_1c\right]\geq 0 \tag{28}$$

Comparing (28) with the first line in (24), we see that this expression can be further transformed in the same way we transformed the Kuhn-Tucker conditions in the last section. Define R_I, the returns on a project with the price I, in the same way as before:

$$R_I\equiv\frac{w_1a_1c}{I} \tag{29}$$

and insert:

$$\begin{aligned}E\left[R_I\right]-R_0&\geq\frac{\operatorname{cov}\left(C_1,R_I\right)}{\operatorname{cov}\left(C_1,R_m\right)}\left(E\left[R_m\right]-R_0\right)\\&=\frac{c}{I}\cdot\frac{\operatorname{cov}\left(C_1,w_1a\right)}{\operatorname{cov}\left(C_1,R_m\right)}\left(E\left[R_m\right]-R_0\right)\end{aligned} \tag{30}$$

Equation (30) may be interpreted to show that projects with a higher internal interest rate than the risk-free rate plus the right-hand side risk premium, are profitable and should be implemented. The problem with applying this criterion is that it seems impossible to get a good estimate of c/I. This problem is solved as follows[14]:

Consider a marginally profitable project, i.e., one with equality in (30). For such a project, the benefit cost ratio BCR must by definition be 1 when we discount period 1 benefits with the correct discount rate and use expected values for the uncertain variables in the formula. Call the correct discount rate q. Marginally profitable projects will have

[14] Here we leave Lund (1993).

(31) $$BCR = \frac{1}{1+q}\frac{E\left[w_1 a_1\right]c}{I} = 1$$

From (31) follows $c/I = (1+q)\left(E\left[w_1 a\right]\right)^{-1}$. At the same time, we have in this case that $E\left[R_I\right] = 1+q$, or else q would not have been the correct discount rate. And we have equality in (30). From this we can compute $1+q$, which can then be inserted in (30). We use a rule for computing covariances to insert $E[w_1 a]$ in the covariance of (30) and arrive at our final investment criterion:

(32) $$E\left[R_I\right] - R_0 \geq \beta\left(E\left[R_m\right] - R_0\right)$$

where

(33) $$\beta = R_0 \frac{\dfrac{\operatorname{cov}\left(C_1, \dfrac{w_1 a_1}{E\left[w_1 a_1\right]}\right)}{\operatorname{cov}\left(C_1, R_m\right)}}{1 - \dfrac{\operatorname{cov}\left(C_1, \dfrac{w_1 a_1}{E\left[w_1 a_1\right]}\right)}{\operatorname{cov}\left(C_1, R_m\right)}\left(E\left[R_m\right] - R_0\right)}$$

Equation (32) is the final result from the theoretical model as far as infrastructure investment is concerned, and (33) is the beta to be estimated. Obviously, the two covariances must be estimated separately and inserted in the formula. This does not at all seem impossible, even if the formula itself looks a bit frightening. Anyway, we will be estimating something that seems theoretically perfectly sound, instead of making do with market analogies. Formally, (32) is equal to the CAPM; the whole difference lies in beta. Vi see that the stochastic returns should be normalised, and we may also normalise the stochastic variable C_1, since it appear in both the covariance of the nominator and the denominator. For Norwegian applications, we are happy to accept the values $R_0 = 1.02$ and $E\left[R_m\right] - R_0 = 0.04$ as set by the Ministry of Finance.

Weaknesses of the model

The model applies to a single individual. That individual might have been a representative consumer if it were not for the fact that this interpretation is inconsistent with risk aversion. For a representative consumer to exist, all individuals must have preferences of the Gorman Polar Form, which means that the second derivative of income is zero. The somewhat disturbing implication is that if we want to use (32) and (33) as an investment criterion on the level of society (which we do), we may do so, but we cannot deny anybody the right to disagree with it.

The assumption that the private consumer produces and uses his own infrastructure is in fact less disturbing. If this should ever be an actual possibility, efficiency in production would require the government to use the same investment criterion. The omission of taxes is perhaps more of a weakness.

The relationship between the transport models and models like the one we have developed here is not an easy one. This is partly because the demand for transport in the transport models usually is a special form of the Gorman Polar Form (income does not enter the

demand functions), thus contradicting the need to consider systematic risk, and partly because it is cumbersome to take account of probability distributions of key variables by way of repeated runs of the transport model. Our partial solution to the latter problem has been to insist that the *input* to the transport model and the CBA should be expected values wherever possible, and to let the covariance between income and the uncertain *product* of demand, time savings per trip and the value of time be a key determinant of the risk premium.

4 Estimation

To estimate the covariances of (33), we need data on C and R_m and on wa for car, bus, boat, air and rail. Annual aggregate national data are used. The data cover the period 1985-2004. There are two reasons why we start at 1985. First, the Norwegian stock exchange was not very well developed until 1985, and second, it was not until that time that car-ownership achieved what we may call maturity.

The covariances of (33) are covariance between the variables over different states of the world. A state of the world is defined by the oil price and the international economic climate (boom, recession). The states of the world are of course not chronologically ordered, so we are not estimating a time trend. On the contrary, everything possible must be done to detrend the data, so that the years will be comparable to each other in every respect except the state of the world and the level of our detrended variables. Put otherwise, any year in the data represents a realisation of period 1 in our model, and any systematic difference between them ought to stem solely from a difference in the state of the world. Of course, this ideal cannot be fully realised. Anyhow,

- We must correct for price increases and express all variables in fixed prices
- We must detrend the data, i.e. remove the influence of everything that causes growth over time
- We must see to it that all possible states of the world are represented in the data and that each state appears with the same frequency that can be expected for the future.

The period 1985-2004 contains recessions and booms in a mix that we have no reason to believe to be different in the future. On the other hand, there are many more years of low oil price than can reasonably be expected in the future. Originally, we did not correct for that. As it turned out that the beta estimates were very sensitive to the number of years of high oil price, we also tried to put more weight on such years, which produced lower betas. Another problem for the estimation was a string of years when the stock exchange and the oil price moved in opposite directions, contrary to what we expected to be normal (and certainly contrary to the following years 2005-2007).

C was defined as household consumption at fixed prices. w was defined as hourly wage after tax for passenger transport, and as wage cost including social costs for freight. The variable a stems from an annual statistical publication published by the Institute of Transport Economics. The variables C and a were computed per capita, thus removing the influence of population growth. All variables entering the covariances (C, R_m and wa) were normalised by dividing with the mean.

Different models of detrending were tried, and we found the DSP (difference stationary process) model to work best. It was applied to C and wa. No apparent trend was found for R_m.

Further details of the estimation process and the data are given in Minken (2005).

Results

Betas between 0.23 and 0.74 were found for the various modes of passenger transport, producing discount rates ranging from 3 to 5 per cent. All modes except sea were in the fairly narrow range of 0.58-0.74, with discount rates around 4.5 per cent. For freight, betas between 0.65 and 1.42 were found, with rail as the only value above 1 at 1.42. This produces discount rates around 5 per cent, with 7.5 per cent for rail. Pooled data for all modes with a higher proportion of years with high oil price gave discount rates for passengers and freight around 3.5.

Taking the sensitivity of the results to small changes in the data into account, we proposed to use a discount rate of 4.5 per cent for all modes except air and sea freight, where a rate of 5 per cent was proposed. The Ministry eventually set the discount rate at 4.5 per cent for all modes without exception. In the light of economic events in the years after 2004, there might be a case for reconsidering both the risk-free rate and the risk premium.

5 Practical advice

The principles governing the official discount rate(s) needs to be accepted by decision-makers and the public at large, or else there is a danger that CBA methods fall into disrepute. This happened in Norway in the years 1999-2005, when pretty much the same framework of a risk-free rate plus a risk premium was used, but without strict regulation and guidance. In particular, the use of different rates for different modes (a result of faulty estimation) aroused much controversy. In that light, the 2006 decision of the Ministry of Transport and Communications to use the same rate for all modes was a wise one. Nobody discusses the discount rate any more.

The main message from this paper is not formula (33) but the simple general principles:

- Use expected values as far as possible in CBA,
- Take risk into account by reducing annual benefits to safety equivalents or by including a risk premium in the discount rate,
- Whether simple comparisons with the risk of an average company on the stock exchange or more sophisticated methods like formula (33) is used, the thing that matters is systematic risk, i.e. whether annual benefits are high in economic upturns and low in depressions or vice versa,
- Make it operational and simple, but without sacrificing the general principle of systematic risk. Try to get this principle across to decision-makers and the public.

So far, we have said nothing about the risks in construction. The same principles should be applied there, and it seems that the best consultants in the field are indeed thinking in terms of adding a reserve based on systematic risk when they assess construction costs. They would achieve exactly the same if they added a risk premium to the interest rate. This risk premium will however not be the same as the one to be used for the annual costs. The covariance of construction costs with national income is not the same as the covariance of annual transport benefits with national income. The completion of construction is a milestone: one type of risk is totally dissolved, and another takes over.

References

Alexander, I., A. Estache and A. Oliveri (1999) A few things regulators should know about risk and the cost of capital. World Bank Policy Working Paper

Arrow, K.J. and R.C Lind (1970) Uncertainty and the evaluation of public investment decisions. *American Economic Review* **60**, 364-378.

Diamond, P.A and J.A. Mirrlees (1971) Optimal Taxation and Public Production, I: Production Efficiency" and "II: Tax Rules", *American Economic Review* LXI (1), 8-27 and LXI (3), 261-278.

Finansdepartementet (2005) Veileder i samfunnsøkonomiske analyser.

Lund, D. (1987) *Investing in non-marketable assets*. Memorandum no 2, February 19th 1987, from Department of Economics, University of Oslo.

Lund, D. (1993b) Usikre investeringer under begrenset diversifisering. *Beta* 2/1993.

Minken, H. (2005) Nyttekostnadsanalyse i samferdselssektoren: risikotillegget i kalkulasjonsrenta. TØI-rapport 796/2005.

Sandmo, A. (1972) Discount rates for public investment under uncertainty. *International Economic Review* **13**, 287-302.

2.5 Appraising the sustainability of urban land use and transport strategies

Appraising the sustainability of urban transport and land use strategies[15]

Harald Minken

Institute of Transport Economics

Contents

[15] Unpublished (submitted somewhere but not accepted), June 2003.

ABSTRACT

In the appraisal of land use and transport strategies with irreversible long-term effects, economic efficiency as measured by a cost-benefit analysis will not capture all our concerns. The concept of sustainability is broader than the concept of economic efficiency, since it involves a concern for the very distant future (intergenerational equity) as well as a concern for the preservation of stocks of natural and cultural resources – a concern that goes beyond the trade-off between consuming them now or later. Arguably, sustainability should be the basic objective of urban transport and land use strategies. Drawing upon recent work in welfare economics and resource economics, we devise a framework for the appraisal of such strategies with respect to sustainability. The framework leads on to a set of indicators that can be computed from integrated transport/land use models. The decision maker may choose to set targets for some of the indicators while combining others (the money metric indicators) into a sustainability objective function. The ensuing constrained optimisation problem is solvable by repeated runs of a model system, as shown in the PROSPECTS project.

Key words: Land use/transport planning, appraisal, sustainability.

1 Background

The private car has been a major force in shaping city areas. Car commuting has given firms access to a wider labour force, furthering specialisation and boosting productivity. Car shopping has given retailers access to a wider customer base, furthering competition and perhaps specialisation and product diversity. The need for workers to settle near their workplaces has been removed, making them freer to choose employment and residence. But the resulting urban sprawl has also undermined public transport, walking and cycling and led to car dependency in a self-enforcing process (Newman and Kenworthy 1999). This may lead to social exclusion of those without a car. Increasing levels of car ownership and car use pose problems of congestion, air pollution, car accidents and noise. New infrastructure building to alleviate congestion may pose a threat to green areas within the city, to the old city centre and to cultural heritage sites. Recently, global warming has emerged as a major challenge, making it imperative to reduce fossil fuel consumption in transport and housing in the cities.

It is the task of strategic land use/transport planning to address all of the problems brought about by the car while not undermining the benefits. Clearly, this is an undertaking with multiple and conflicting objectives. Two main tools are needed – a model system to predict the impacts of a strategy and an appraisal framework. Appraisal needs to integrate all objectives in a consistent framework. The model system must be able to give a good representation of the complex interactions between transport and land use, the environment and the rest of the economy.

In this paper we show how the concept of sustainable transport and land use may be used to integrate all the objectives into a single appraisal framework. The framework consists of an intergenerational welfare function derived by Chichilnisky (1996) and target values set for other (non-welfare) objectives. Strategic planning takes the form of maximising the welfare function subject to the other targets being reached. This approach was taken in the EU Fifth Framework project PROSPECTS (Minken et al. 2003, May 2003). The paper summarisies the PROSPECTS approach with special emphasis on explaining the rationale behind the Chichilnisky function.

Section 2 presents the PROSPECTS approach to strategic planning. Section 3 outlines appraisal within this approach. In Section 4, the contribution of Chichilnisky (1996) and Heal (2000) to planning for sustainability are presented and applied to urban transport/land use planning, giving rise to the particular form of the sustainability planning problem set out in Section 5. Section 6 concludes.

2 The PROSPECTS approach to transport/land use planning

A land use plan, such as the allocation of new land to housing, induces behavioural responses. Developers may decide to build, building induces relocation of households, which in turn induces some relocation of businesses, etc. All of this in turn affects the transport system in the form of changes in car ownership, destination choices, mode choices, etc. These changes have environmental impacts as well as welfare and equity impacts, as the system moves towards new equilibrium land prices, rents and travel costs.

In the same way, a transport plan may bring about relocation. Again, the changes have environmental impacts as well as welfare impacts through new transport costs and rents.

Thus transport, land use and the environment are linked together, and planning needs to take this into account. On the other hand, the links to other markets and activities in the city may be so much weaker that we can ignore them and treat the volume of production, the income level etc. as exogenously given. With this, we have defined the system that we are planning for. We consider it meaningful to talk about a sustainable land use and transport system, without considering other consumption and production in the city.

This system is complex. It is impossible to grasp intuitively how different strategies impact on the various parts of the system. This is why we need a mathematically formulated model system (an integrated land use/transport, or LUTI, model) to predict the impacts. If based on sound economic principles, it can capture the impacts of transport and land use plans and strategies and provide a basis for appraisal of such strategies. The number of markets and relationships accounted for in such models may vary. The PROSPECTS approach should in principle be applicable with models that cover from a few to all aspects of consumption on the city. But it cannot be applied to pass judgement on the sustainability of the production sectors in a city, since they will often be parts of a worldwide system of production and trade and must be judged in that context.

The subject matter of PROSPECTS is planning for sustainable urban land use and transport, which is a meaningful concept if and only if the modelled sub-systems interact only weakly with other systems of the city. The objective – sustainability – is very general, the scope is the entire urban land use/transport system, the time horizon is fairly long and whole packages of policy instruments are considered, so this planning is strategic in nature.

Planning takes place within a certain given context, consisting of the institutional framework, demographic forecasts and assumptions about income growth, national policy such as car and fuel taxation, available technologies, given constraints on land use etc. This context, as it develops over time, is called a *scenario*. Within this, or across different scenarios, *strategies* consisting of a set of available policy instruments, each at their particular level of use, are tested by implementing them in the model system. The policy instruments include pricing instruments, infrastructure provision, public transport policies, restrictions etc. *Barriers* on the use of the policy instruments and on available finance will have to be considered.

The appraisal framework is used to select or recommend a best strategy, rank strategies, discard useless and unacceptable strategies or select a set for further study. It starts from a definition of a sustainable urban transport/land use system, derives a set of objectives that legitimately belong under this definition, and devises indicators covering the whole range of objectives. The indicators may be combined to form an objective function, or targets, reflecting sustainable levels or milestones on the way to sustainability, may be set for them individually. The indicators must be computable from model system output, since we are engaged in planning for the future and not in measuring progress at the present.

A particular feature of the PROSPECTS approach is the use of optimisation. Algorithms involving repeated runs of the modelling system have been designed to find the strategy that maximises the objective function subject to other indicators reaching their target levels and other constraints. To the extent that policy instruments can change over time, this amounts to solving a dynamic optimisation problem subject to constraints. The first tests of dynamic optimisation were made in PROSPECTS.

Even if we are not prepared immediately to accept the "optimal" strategy, this procedure is useful. It is able to produce strategies previously not thought of and to produce new knowledge about the complex ways policy instruments interact.

The approach requires the participation of decision-makers at various points in the planning process: defining objectives and their priorities, choosing possible policy instruments, setting targets and possibly using the results to reconsider priorities and targets.

3 The PROSPECTS appraisal framework

3.1 A hierarchy of objectives and indicators

Starting from a general conception of what sustainability is, the PROSPECTS project proceeded to derive sub-objectives that should all legitimately belong under the main objective of sustainability. These were further specified until it was possible to construct an indicator for each sub-objective or sub-sub-objective. The indicators would have to be measurable from model output.

The PROSPECTS definition of a sustainable land use/transport system is:

A sustainable urban transport and land use system

- *provides access to goods and services in an efficient way for all inhabitants of the urban area*
- *protects the environment, cultural heritage and ecosystems for the present generation, and*
- *does not endanger the opportunities of future generations to reach at least the same welfare level as those living now, including the welfare they derive from their natural environment and cultural heritage.*

Note that welfare in the form of efficient provision of goods and services is a legitimate sub-objective of sustainability – at least as long as it does not hinder the attainment of environmental objectives. Next, there will obviously be a place for environmental sub-objectives as well. Also, please note the little words "all inhabitants". They imply the objectives of fair distribution, equity and social inclusion. All such objectives should legitimately be sub-objectives of sustainability. This is often expressed by saying that we require economic sustainability, environmental sustainability and social sustainability.

A survey of 54 European cities confirmed that the following six objectives all belong as aspects of the overarching objective of urban sustainability:

1 economic efficiency
2 liveable streets and neighbourhoods
3 protection of the environment
4 equity and social inclusion
5 safety; and
6 contribution to economic growth.

Some of the objectives, such as protection of the environment, obviously need to be specified further. See Minken et al. (2003) for this. At any point in time, such as for instance the year 2010 or the year 2100, there will presumably be a concern for each of these six objectives. To take account of these objectives in a way that brings about sustainability, however, we need an objective that does not concern any single year. Rather, it concerns how we trade off the achievements in the various years against each other. So we require

7 intergenerational equity.

There might be a case for including *health concerns* as an additional separate objective. However, health objectives are taken care of by the (traffic) safety objective, by sub-objectives such as air pollution objectives under protection of the environment and by walking and cycling benefits under the economic efficiency objective.

Missing from this list is an objective that does not concern the outcome from a strategy under a given scenario, but rather the ability of a strategy to perform well under different scenarios. This objective might be called

8 robustness.

Since the future is inherently uncertain, the robustness of a strategy is a valuable property. Its evaluation requires that strategies be tested in a variety of scenarios.

Finally, cities will also have objectives that concern the planning process itself. These can be summed up as

9 a democratic planning process.

Since this objective does not concern the outcome of particular strategies, it cannot be fitted into our appraisal framework. This objective has to be taken care of by public participation, institutional reform, simple and clear forms of presentation of the results etc.

Having established the objectives, we turn to their measurement. Economic efficiency is measurable by standard methods of cost-benefit analysis, including the costs of air pollution, noise and accidents. Some comments will be made in Section 4 below on how to assess the value of unused land and the costs of CO_2. The latter subject is developed further in Minken et al. (2003). Likewise, equity indicators have been devised in PROSPECTS. Safety may be measured by accident costs or the number of accidents. The problem areas, then, concern liveable streets and neighbourhoods and economic growth. In PROSPECTS, it was suggested that the cost of accidents involving a pedestrian or cyclist and a car could be used as an indicator of liveable streets and neighbourhoods. It remains to be seen if it works. In spite of the work of SACTRA (1999), economic growth remains a real problem area. It was suggested in PROSPECTS that the net benefit to households, firms and the government may be used as showing the growth potential of a strategy.

3.2 Use of the indicators

Obviously, the simplest form of appraisal consist in computing the indicators for each of the tested strategies, and leave it to the participants in the decision making process to work out their decisions based on this information and any other information they might have. Since no formal criterion is used to produce a ranking of strategies or to partition the set of strategies in recommended and discarded strategies etc., this form of appraisal might be called informal. However, it is still based on a systematic and comprehensive framework of quantified indicators, covering the whole range of objectives.

Among the formal forms of appraisal, we make a distinction between those that result in a complete ordering of strategies and those that do not. *Setting targets* is the basis for appraisal that does not produce a complete ordering of strategies. *Forming an objective function* is the basis for appraisal that does produce a complete ordering. An objective function is a (linear) function of a sub-set of the indicators, to be used for (partial or comprehensive) appraisal of strategies or for optimisation. In Section 5, we establish the general form of a sustainability objective function that includes the intergenerational equity objective, the economic efficiency objective and a least some of the environmental objectives – international objectives with respect to CO_2 emission, and local objectives with respect to land use, air pollution and noise.

The PROSPECTS appraisal framework is flexible, and it is by no means mandatory to use this objective function. But assuming that we do, we have to consider what to do with the objectives not covered by this objective function. At least, this concerns the (intragenerational) equity objectives. Three options exist: (a) to add them to the objective function, effectively forming a multi-criteria objective function, (b) to set targets and include these as constraints in the optimisation problem, and (c) to compute and present their indicators as complementary information to the decision makers.

Targets are defined as the level of the indicators that is necessary to bring about a sustainable urban land use and transport system. A lot of subjective judgement is required to set the targets, and there are bound to be different opinions about them. Since the farther we look into the future, the less we know, it does not make much sense to make detailed model predictions and compute indicators beyond 2020 or 2030, say. But by that time we should not expect the urban system to be sustainable in the full sense. So some judgement must be made as to what targets are the most important and what their levels should be to secure that the 2020 state would evolve to be fully sustainable.

If we allow targets not to be fully reached, we will speak of them as *goals* rather than targets. Roughly, goals express the ideal or final state that we aim for, while targets express the necessary minimum levels that we do not want to fall below at any cost. *The level of goal achievement* with respect to a goal is 0 in the present state and 1 if the goal is exactly reached. For intermediate states we define it as the difference between the achieved level and the level of the present state, divided by the difference between the goal and the level of the present state. Using this metric, constraints to the optimisation problem can be expressed by requiring the indicators to be above a certain number between 0 and 1 in a certain year. Obviously, the last modelled year is the most important in this respect.

Targets may not only be set for the indicators that are excluded from the objective function. For instance, even if accident costs are included in the objective function, it may be that the city has a target with respect to accident reduction. If this turns out to be a binding constraint in optimisation, it means that the decision makers value accidents higher than implied by the usual unit cost.

4 Measuring sustainability

The purpose of this section is to establish the general form of a sustainability objective function that includes the intergenerational equity objective, the economic efficiency objective and a least some of the environmental objectives. To extract the principles that should be reflected in such a sustainability objective function, we turn to economics.

4.1 Welfare measures

An *allocation* is an assignment of a bundle of goods to each individual of society. A *welfare criterion* is some kind of rule that can be used to make such judgements about allocations, at least in some of the cases. Surely, each person may have its own, but economists are on the outlook for a rule that embodies as little value judgement as possible, and such a rule has been found in the Kaldor-Hicks-Samuelson (KHS) welfare criterion (Chipman and Moore 1994). An underlying assumption is that every individual is able to rank all bundles of goods given to her – that is, she has a utility function. Obviously, each allocation involves definite total quantities of all the goods. Consider the allocation A. We may speak of a redistribution of A as any allocation that keeps within the same bounds on total quantities as A. According to the Kaldor-Hicks-Samuelson criterion, then, allocation A is better than allocation B if for any possible redistribution B' of the goods of allocation B there is a redistribution A' of the allocation A such every individual weakly prefers their bundle of goods in A' to their lot in B'.

Note that the KHS criterion is based solely on individual preferences and the relatively uncontroversial social value judgement that if any individual, regardless who, experiences an improvement and nobody else gets worse off, this is an improvement to society (the Pareto property). However, for A to be judged better than B, the redistribution that brings about a Pareto improvement need not actually be carried out. This is the weak spot of KHS. If instead we required an actual Pareto improvement we would not be able to decide many cases, though.

The question arises if KHS is able to decide all cases. The answer is yes, provided the individual utility functions have a particular form, the Gorman polar form, and individuals maximise utility subject to an ordinary budget constraint. We mention this because the agents of our models often have indirect utility functions of this form. In this case, then, a *welfare measure* can be constructed, such that if allocation A is better than B according to KHS, it gets a higher value of the welfare measure. This is because in this case, a representative consumer exists. The welfare measure is the indirect utility function, the expenditure function or the money metric indirect utility function of the representative consumer.[16]

Regardless of the form of the utility functions, we might of course construct a welfare measure if we are willing to make somewhat stronger normative assumptions on behalf of society. Any function defined on allocations would do, but economists will usually require that it respects individual preferences. A (social) welfare function is a welfare measure that is defined over individual utilities instead of directly over allocations and is strictly increasing in each argument. The social welfare function that is simply the sum of individual utilities is called a *utilitarian welfare function*. Cost-benefit analysis is carried out either by assuming a utilitarian welfare function where all utilities are in the same money metric or by the representative consumer approach. If there is a representative consumer in the underlying model, the two approaches are identical. If society consists of groups, each with their representative consumer, they may be combined.

Social theorists object to cost-benefit analysis on the grounds that society cannot be reduced to its individual members or that individuals cannot be reduced to utility maximisers. Environmentalists object that the environment should not be valued by individuals' willingness to pay for it. Many academic economists find it all too simplistic. Could a valid measure of sustainability be constructed from this fundament? We shall see.

[16] See Varian (1992) or any other textbook for definitions.

4.2 Chichilnisky's theorems

Assume an infinite series of generations, $g = 1, 2, \ldots, \infty$. An intertemporal welfare function will have to be defined over the utilities of all these generations. For short, it must be defined over infinite utility streams. It must be strictly increasing in the utility of any generation (the Pareto property), and it must ascribe a real number – the social welfare level – to any allocation of goods to the generations. Here generations play the role of individuals in the preceding section, but the fact that the generations are ordered in time and go on to eternity makes a difference. Without some form of weighting of the individual utilities, intertemporal welfare will easily be infinite in many cases. But this makes comparison impossible and is ruled out by the very definition of a welfare function.

Why then do we introduce infinity? Why not set a cut-off date for the society we consider? First, because such a date would be entirely arbitrary. From a sustainability point of view we also take an interest in what happens after an arbitrary cut-off date. Second, because conceptually, it will be better to treat the end of society, if it happens, as an event that happens inside our infinite perspective. After all, we might be able to influence it or postpone it.

The form of weighting of the utilities of individual generations that is commonly used is of course discounting. It need not be discounting at a constant discount rate: any set of weights on the utility of individual generations that makes the intertemporal welfare function converge will do. Following Chichilnisky (1996), we assume that the utility of individual generations is bounded below and above. Then any function of g, $\Delta(g)$, that satisfies $\Delta(g) \geq 0$ for all g and $\sum_g \Delta(g) < \infty$ will do the trick and is called a discount factor.

The utility of generation g is called U_g. A *utilitarian intertemporal* welfare function W_u can now be defined as

$$W_u = \sum_{g=1}^{\infty} U_g \, \Delta(g) \qquad (1)$$

Except for the facts that (a) time goes to infinity instead of an arbitrary cut-off date and is measured in generations instead of years, and (b) the discount rate is more general than what we usually use, W_u is what we measure in a perfectly normal cost-benefit analysis. W_u is a complete ordering of utility streams and satisfies the Pareto property.

Is W_u a good measure of sustainability? We might for instance specify the discount factor as involving not a constant discount rate, but a discount rate that declines over time. There is much interest in such declining discount rates as a means to evaluate climate policy (IPCC 2001). But according to Chichilnisky (1996), W_u is useless as a measure of sustainability.

Next, consider an entirely different option. By the *sustainable utility SU* of a stream of utilities $\{U_g\}_{g = 1,2,\ldots}$ we shall mean the utility level as time goes to infinity:

$$SU = \theta \lim_{g \to \infty} U_g \qquad (2)$$

Here θ is an arbitrary constant. SU is not a *welfare function*, since it is not strictly increasing in all U_g and since it is not defined for utility streams without a limit. Nevertheless, it may be the basis for a *welfare criterion*, namely that of two utility streams, society prefers the one that is better in the long-run future (provided that can be determined). One possible interpretation of the definition of sustainability given by the Brundtland Commission is that from two feasible policies, each with their own stream of utility and with U_g for the generations living now sufficiently high in both, we should adopt the one with the highest *SU*. To choose by the *SU*

alone would be even more future-oriented. It would mean that we should make any sacrifice now if it would improve the long-run future.

Still another welfare criterion would be the Rawlsian criterion: of two utility streams, chose the one where the *least* U_g is the largest.

Chichilnisky (1996) proposes two axioms that she wants an intertemporal welfare criterion to obey – she calls them "no dictatorship of the present" and "no dictatorship of the future". Suppose we have two utility streams $U = \{U_1, U_2, \ldots\}$ and $V = \{V_1, V_2, \ldots\}$ and we change them to U^k and V^k by putting in minimum values for the U_g and V_g after some point in time k. (Remember that is was assumed that there is a minimum and a maximum attainable utility level). Now if we apply a welfare criterion to the new utility streams U^k and V^k, we find for example that U^k is better. Now go back to our original utility streams U and V and apply the welfare criterion to them. If for some k, we could *always* judge the matter of who is best of U and V by looking only at who is best of U^k and V^k, then the welfare criterion we apply is a dictatorship of the present.

With a dictatorship of the present, whatever happens after a certain point in time can never make a difference to the ranking. That point in time may be long into the future if the difference in the "present" (before k) utilities is small, but it exists. Conversely, in a dictatorship of the future, whatever happens before a certain point in time can never make a difference to the ranking.

The "no dictatorship of the present" axiom embodies our concern for the distant future in matters such as global warming, nuclear waste management, biodiversity etc. Many people will find that it corresponds to their notion of what sustainability is.

Theorem 1 in Chichilnisky (1996) states that the welfare function W defined by

$$W = W_u + SU \qquad (3)$$

satisfies the two axioms and is neither a dictatorship of the present nor a dictatorship of the future. It also states that W_u *is* a dictatorship of the present, regardless of the discount factor chosen. Obviously, SU is a dictatorship of the future. The combination of two deficient criteria, however, makes a satisfactory criterion.[17] W also has the Pareto property and ranks all utility streams, so may be called a welfare function.

Chichilnisky's Theorem 2 is astonishing. It states that if we require the intertemporal welfare function to satisfy the Pareto property and the two dictatorship axioms and to be continuous and linear in the utility of generations, then W as defined here (see also footnote 2) is the *only possible candidate*.[18]

Roughly speaking, Theorem 3 says that in general, one cannot approximate the solution to a dynamic optimisation problem with the sustainability function W as objective function by using a suitably modified W_u instead. "Sustainable optima and discounted optima can be far apart". Correspondingly, it will not always be possible to specify dated prices so that if you maximise the discounted value of consumption at these prices, you get the sustainable optimum.

---

[17] In fact, the SU as defined here is only one of a class of functions that produces a W that satisfies the criteria. All have to do with how the utility stream behaves in the limit.

[18] Again, the theorem holds under slightly broader conditions: the linearity may be replaced by a condition called independency.

In general, markets cannot solve sustainability problems. Not even planners doing cost-benefit analysis can.

4.3 Discussion

Assuming that we accept her assumptions, Chichilnisky's work has philosophical and practical implications.

At a philosophical level, we might ask (as Chichilnisky does herself) whose preferences it is that is reflected in *W*? Assuming that agents are utility maximisers without altruistic elements, no single agent has preferences of this form. Even if environmental qualities are included in their utility functions, we will probably not be able to bring out the true value of these qualities to society by asking them for their willingness to pay for them. We did not ask all generations. It seems that society as something distinct from its individual members re-emerges in welfare economics through this work.

Obviously, society will also have practical tasks to perform in planning for sustainability, deriving the right shadow prices (social cost) of goods and resources that affect the long-term utility and implementing the sustainability plan through Pigouvian taxes and regulation. Following Coase (1960), society might alternatively internalise externalities through defining property rights and establishing the missing markets, but since future generations cannot take part in these markets and no utility maximising agent can represent their interests fully, this solution will sometimes not do.

But not all of our actions now affect the long-term future. If we use the Chichilnisky criterion to pass judgement on actions that only makes a difference in the short run, it reduces to an ordinary cost-benefit analysis.

Chichilnisky's work has gone some way towards resolving the long-standing dispute about what discount rate to use in cost-benefit analysis of plans with long-term impacts (see Portney and Weyant 1999, Weitzman 1998 and IPCC 1996 and 2001 for introductions to this debate). In the sustainability welfare function, the discount rate is relieved from the task of reflecting concerns about the distant future and is free to reflect the intertemporal preferences of individual utility maximisers. The discussion can still go on about whether to use a constant discount rate or a discount rate that decreases with time. The first produces time consistency in the individual preferences, while the other produces time inconsistency[19], but seems to reflect people's actual choices better (Heal 2000).

However, Chichilnisky's solution has so far had little practical impact on the discussion about discounting and is not, for instance, mentioned in IPCC (2001).

Time inconsistency is not an issue with respect to the sustainability objective function *W*, since we do not pretend that it reflects the preferences of an infinitely long-lived utility-maximising individual. Instead, it is better seen as a function that reflects our views about intergenerational equity as well as efficiency. This feature of *W* is reflected in the arbitrary constant θ that we included in SU. In fact it is not arbitrary, but reflects our concerns about the future welfare relative to our concerns about the present.

[19] Suppose you solve a dynamic optimisation problem and start to implement the optimal solution. If at some later date you solve the problem again and find that you want to depart from the optimal path found the first time, then your objective function (intertemporal utility function) exhibits time inconsistency.

4.4 Valuing stocks of natural resources

Conservationist concerns are only implicitly present in Chichilnisky's axioms of "no dictatorship of the present" and "no dictatorship of the future". Heal (2000) suggests that the essence of sustainability lies in three axioms:

(1) "A treatment of the present and future that places a positive value on the very long run".

(2) "Recognition of all the ways in which environmental assets contribute to economic well-being".

(3) "Recognition of the constraints implied by the dynamics of environmental assets".

The first point was covered above. A rough first approach to the third point is to divide natural resources into exhaustible and renewable resources. The real news is in the second point.

The second point captures the conservationist concerns inherent in the concept of sustainability. Natural resources should be valued not only as something that may be consumed (in production or consumption), but also as stocks that benefit us even when not being consumed. The fundamental reason for this is that we are dependent on some basic qualities of our surrounding ecosystems for our quality of life and indeed to continue to exist.

To take care of this point, Heal introduces stocks as arguments in the utility function, i.e. $u = u(c, s)$, where c is consumption and s is the level of stocks (c and s might be scalars or vectors). This leads to a greener optimal policy even in the discounted utilitarian framework. But stocks giving utility can of course also be introduced in Chichilnisky's welfare function.

To arrive at Heal's formulation of this, we note that if annual utilities can be summed across the generations living in that year, and if the utility of a generation is linear in years, Chichilnisky's function can be recast as involving annual utility instead of the utility of generations. By assumption, the form of the annual utility function is the same for all years. Annual utility must now be interpreted as society's welfare function for that year. Making time continuous instead of discrete, and multiplying by the constant $\alpha = (\theta + 1)^{-1}$ we have:

$$W = \alpha \int_0^\infty u(c(t), s(t))\Delta(t)dt + (1-\alpha)\lim_{t\to\infty} u(c(t), s(t)) \qquad (4)$$

Here, $c(t)$ is consumption at time t (possibly a vector), $s(t)$ is the stock of natural resources at time t (also possibly a vector) and $\Delta(t)$, normalised without loss of generality so that $\int_0^\infty \Delta(t)dt = 1$, is the discount factor. The constant α determines how much weight is given to the long-term future relative to the "present" and may be called the intergenerational equity parameter.

If we assume that the discount rate is a constant, the discount factor in the case of continuous time can be written as e^{-it}, where i is the constant discount rate. But we might also use a discount rate that decreases with time. Indeed Heal (2000) argues repeatedly for logarithmic discounting, that is, using $e^{-i\ln t}$ instead.

W as given here satisfies Chichilnisky's two axioms and the requirements of a utilitarian welfare function, and by Chichilnisky's Theorem 2 is the only function to do so.

A sustainability problem – or a Chichilnisky problem, as we shall call it – can now be written as

$$\text{Max } W = \alpha \int_0^{\infty} u(c(t), s(t))\Delta(t)dt + (1-\alpha)\lim_{t\to\infty} u(c(t), s(t)) \qquad (5)$$
$$\text{subject to } s'(t) = r(s(t)) - c(t), \qquad 0 < \alpha < 1.$$

Here, *r(s(t))* is the rate of regeneration of the stock, which is a function of the size of it. For exhaustible resources, the rate of regeneration is always zero, so such natural resources have very simple dynamics: the stock diminishes by what is consumed.

Heal studies solutions to this problem under different assumptions. For our purposes, we only need to mention some of the results.

For exhaustible resources, it is well known that when stocks do not give utility, the optimal path of depletion involves to let the level of stocks approach zero in the limit (Hotelling 1931). When stocks do give utility, some of the resource may be retained forever. The optimal path of depletion in the Chichilnisky problem involves keeping at least as much of the stock forever as in the utilitarian framework with stocks giving utility. So the combination of points (1) and (2) above does lead to a greener policy.

For renewable resources, there is no solution to the Chichilnisky problem unless the discount rate approaches zero in the limit. If it does, the solution is identical to the utilitarian problem with the same discount rate.

4.5 Applying the Chichilnisky/Heal framework: The problem of infinity

Infinity is central to the arguments leading to *W* and the proof of the theorems. However, in practical applications, we are confined to a finite horizon. Beyond 30 years, say, we will obviously not be able to predict consequences with anything resembling accuracy and we will have next to nothing to say about *SU*. Does this make the whole construction superfluous and lead us back to an ordinary cost-benefit analysis? We do not think so. We *want* to think about long-term consequences, and in a way it is this thought that matters. What we need to do is to use the conditions in the last modelled year – year 30, say – as indicating the long-term consequences. For instance, green land that is built down 30 years ahead may be considered lost forever. By applying the *W* with year 30 as an indicator of *SU*, some of the long-run costs (and benefits) of this may be captured. Also, targets can be set for year 30 with respect to environmental qualities to make this year resemble more closely what we think about the long-term sustainable situation. A public sector deficit should not be allowed for year 30, since such deficits cannot be sustained indefinitely.

If we use subjective judgement to set the levels of exhaustible resources that will be kept forever from year 30 on, the question arises if we can then do without the *SU* term and the intergenerational equity parameter α. Generally, the answer must be no, because we are interested in making the long-term welfare level as large as possible within the limits sets by the natural resources, and we may have enough policy instruments at our disposal to influence it even after we have achieved the environmental targets. Note however that if the year 30 targets are too many or are set too high, there might be little scope to influence the sustainable utility *SU*, and thus the choice of α becomes unimportant for the result.[20] The whole exercise

[20] This observation accords well with Proposition 16 of Heal (2000), which says that if we are able to find the proportion of the stock that should be kept forever in the Chichilnisky problem, the problem can be converted to the utilitarian problem to maximise W_u subject to consumption keeping within this bound.

in this case becomes one of backcasting. Sensitivity tests might reveal if one or two of the normative choices have determined the result.

It might not be optimal to force the path towards sustainability to take only 30 years, and so some compromises may have to be made in setting the targets. Nevertheless, if we cannot predict more than 30 years ahead, then if we want to apply *W* at all, we have to identify *SU* and year 30.

4.6 Further issues of application

Hopefully, the preceding sections have shown that the strategic transport/land use problem set out in Section 3 can be cast as a Chichilnisky problem with some additional constraints. The main argument in favour of this is obviously that it takes care of our concerns for the objectives of economic efficiency, intergenerational equity and conservation of the environment in a simple formulation that has a good theoretical foundation in welfare economics. Other objectives may be cast as additional constraints.

This approach fits in well with the optimisation approach to strategic planning taken in the PROSPECTS project. It also fits in with the way ordinary cost-benefit analysis is performed in multimodal transport planning. Interpret $u(c,s)$ as the annual net benefit of a strategy, consisting of benefits to households and firms, government financial surplus and external cost savings. The benefits of *consuming* natural resources are captured by the consumer surplus of the households and the producer surplus of firms. In case the best alternative use of the resources is to keep them as stocks, the true social cost of resources equals the marginal utility of keeping them as stocks, measured in monetary terms. This marginal utility – the willingness to pay to save the resources – can be assessed through stated preference surveys in the case of local resources. Once had, the correction from perceived costs to true social costs is achieved by making the right entries in the government surplus column and the external cost column of the cost-benefit accounting table (Minken et al. 2003, Chapter 10). What this amounts to, is that the annual net benefit can be written as $u(c,s) = u_1(c) - p(s)s$, where

- $u_1(c)$ is net user benefits and producer surplus (including the perceived costs of consuming the resource, i.e. the market price including tax), plus government financial surplus, including revenue from taxes on the resource,
- $p(s)$ is the willingness to pay to keep the stock, so that $p(s)s$ are the environmental costs.

The social cost $p(s)$ may be a function of the level of stocks, and as a rule we must assume that it is rising as long as the stock diminishes. Thus a correctly conceived cost-benefit analysis does put a value on stocks, and results in an annual net benefit that is separable in the utility of consumption of resources and the utility of saving the resources.

Two main *renewable* resources in urban land use/transport planning are local air quality and silence. They are dealt with by including air pollution costs and noise costs in the utility function $u(c,s) = u_1(c) - p(s)s$ and/or by political targets. In general, we are only implicitly considering the dynamics of renewable resources in strategic land use/transport planning, since the time scale of such dynamics is much shorter than our planning horizon. Thus our planning problem is basically concerned with exhaustible resources.

The local exhaustible resources are various types of “unused” or “underutilised” land, be it farmland, woodland, parks, or even cultural sites and old factory sites. The terms ‘white land’, ‘green land’ and ‘brown land’ are sometimes used to classify the types. We need to make an initial guess at the amount of these resources at each location that we think should be kept forever, taking the interests of future generations into account. This is included as part of the

scenario assumptions. The difference between the currently used areas and the areas that are excluded from use is the resource that may be consumed or kept. The value of keeping this resource is determined through stated preference analysis. It might be adjusted somewhat upwards as the resource is depleted.

The non-local exhaustible resource that we need to include is 'atmosphere', which is depleted through CO_2 emissions leading to global warming. Here the matter is rather different than with respect to land. The CO_2 cost path is the outcome of quite another optimisation problem, namely a discounted utilitarian problem or a Chichilnisky problem at the level of the world. The shadow cost resulting from that problem is to be inserted as a given datum in our problem. A local CO_2 target is probably not a wise policy, since it produces a shadow cost of CO_2 that differs from city to city, giving rise to cost inefficient strategies of CO_2 reduction.

Ideally, one would want the CO_2 costs found in for instance IPCC (2001) and other utilitarian frameworks to be adjusted upwards to take account of the long-term future. To see this, consider the Chichilnisky problem with a separable utility function, $u(c,s) = u_1(c) + u_2(s)$. The resource is assumed to be exhaustible. Heal (2000) shows that the shadow price $\hat{\lambda}$ of the exhaustible resource at the stationary state, i.e. when consumption of the resource has stopped, is:

$$\hat{\lambda} = u_2'(\hat{s})\left(\frac{1}{i} + \frac{1-\alpha}{\alpha}\right)$$

Here, $\hat{s}$ is the level of stock that is kept forever. u_2' is the marginal utility of the stock (or if you will the willingness to pay to acquire one more unit of it). $u_2'(1/i)$ is the net present value of a perpetual stream of constant benefits to the amount of u_2'. Consequently, the value to society of one unit of the stock remaining at the stationary state is this net present value or capital value as perceived by the individuals living in sustainable conditions, increased to take account of the intergenerational equity issue.[21]

The problem with including such an adjusted CO_2 cost in our objective function is of course that no α at the world level is agreed. If we use a locally agreed α, our CO_2 reduction policy will be cost inefficient, and if we value CO_2 as everybody else does, we ignore the really long-term effects. No good solution exists to this problem, but we prefer the "local" solution, since this may induce other to value CO_2 similarly, which is what is needed on a global scale. Thus we propose to use a CO_2 cost broadly derivable from IPCC (2001) and insert it in our objective function as any other cost, which means that in our local planning problem, CO_2 will implicitly get a shadow cost resembling Heal's formula as time approaches year 30. (Due to the cut-off date, the difference will be in the first of the two terms in the parenthesis, not in the second).

4.7 Intragenerational equity

One aspect of the concept of sustainable development that is obviously not covered by Heal's definition of section 4.4 or in the Chichilnisky problem is intragenerational equity. A fair distribution across members of society or across countries is an objective in its own right as well as a precondition to achieve sustainable use of resources. We do not propose to tackle world poverty through land use/transport planning in urban areas of the industrialised world, but nevertheless there are important equity aspects of our strategies that need to be addressed.

---

[21] The shadow price at points in time before the stationary state, or the amount of the stock that is available for consumption until the stationary state, cannot be determined by such simple formulas.

These are concerned with the geographical distribution of benefits in the strategy, with the capability of strategies to counteract an unjust income distribution, with equity between those that have access to a car and those who have not, and with the dissipation of benefits to the wider regional or national level, leaving the inhabitant's of the city to experience the costs only.

Obviously, all of these aspects of equity might be of crucial importance to the implementation of a strategy. However, there are ways of redistributing the monetary benefits to achieve better outcomes with respect to equity in any strategy, and these redistribution measures should be thought of as forming an element in the strategy. Interestingly, one way of keeping the benefits of road pricing within the urban area is to earmark the revenue for use in the local transport system. Even if this might not be as efficient as other uses of the revenue, it is rapidly becoming the standard requirement when road pricing schemes are introduced.

Minken et al. (2003) gives a range of equity indicators covering all the aspects of equity mentioned here. Targets with respect to the most important of these indicators in any particular case – measured after redistribution measures have been taken! – could be used as constraints in the optimisation problem.

Potentially, there is a sharp conflict between the equity and efficiency of transport/land use plans. It comes to light when we recognise that the most important forms of taxation have distortionary effects in the total economy. Therefore, if revenue collected in the transport and land use system is used to cut back such taxes, there is an efficiency gain. The problem of using the revenue to achieve equity objectives is that usually, this means that the distortionary effects of the tax system are not reduced (Fridstrøm et al. 1999, Parry and Bento 1999). This issue should be addressed when the redistribution schemes in a strategy are devised.

5 The PROSPECTS objective function

We are now able to summarise our discussion in the preceding sections – in particular, Section 3 and Section 4. We propose to adopt an objective function OF to appraise the sustainability of strategies and to be optimised and find optimal strategies. The general form of OF is:

$$OF = \sum_t \alpha_t \left(b_t - c_t - I_t - \gamma_t g_t \right) + \sum_{it} \mu_{it} y_{it} \qquad (6)$$

where

$\alpha_t = \alpha \frac{1}{(1+r)^t}$ for all years between 0 and 30 except year t^*, the last modelled year, r is a discount rate and α, the intergenerational equity constant, is a constant between 0 and 1, reflecting the relative importance of welfare at present as opposed to the welfare of future generations.

Also,

$$\alpha_{t^*} = \alpha \frac{1}{(1+r)^{t^*}} + (1-\alpha)$$

b_t and c_t are benefits and costs in year t, including user benefits, producer surpluses, benefits to the government, and external costs. Investment I_t has been singled out as a special type of cost.

γ_t is the shadow cost of CO_2 emission, reflecting national CO_2 targets for year *t*,

g_t is the amount of CO_2 emissions in year *t*,

μ_{it} is the shadow cost of reaching the year *t* target for sub-objective *i*,

y_{it} is the level of indicator *i* in the year *t*.

Many of these variables are of course specific for a particular strategy – a subscript denoting strategies is however omitted here.

The Sustainability Objective Function *OF* is in accordance with the definition of sustainability, because it involves the weighted sum of a CBA and the welfare of an undiscounted year (this is the first summed terms) plus penalties to assume that this last year stays within environmentally sustainable limits (these are the last summed terms). Optimisation may be carried out with this function or with the first part of it (the Chichilnisky function), making the y indicators into constraints. The most probable indicators to be used as constraints are equity indicators, accident indicators, local air emission indicators and financial constraints. A financial constraint securing public sector surplus in the final year should always be included.

If constraints have been set in such a way that we are pretty sure that the last year of the appraisal period is sustainable, we might modify *OF* by including the perpetual benefits of that year in it. That is, instead of α_{t^*} as given above, use

$$\alpha_{t^*} = \alpha \frac{1}{r(1+r)^{t^*}} + (1-\alpha).$$

However, unless the optimal policy produces annual benefits that change very much towards the end of the appraisal period, the same effect can be accomplished by increasing α slightly.

To apply such a combined framework of a sustainability objective function and targets to the appraisal of strategies, the decision makers must be able to make normative decisions at three points: First, they will have to discuss the objectives and reach a clear understanding of their priorities with respect to them. This will determine what indicators to use. Second, they must be able to decide on what α to use. And third, they must be able to decide on targets.

Of course, it is not to be expected that decision-makers will be able to set the level of α directly. One possibility is to solve the optimisation problem for different alphas and point out the trade-offs in terms of the level of achievement with respect to the different objectives. Preferably, the decision makers should be given the option to reconsider all of their normative choices when they see the results, and this is an important reason why the modelling system should be fast to run.

The PROSPECTS appraisal framework requires that the model system is able to compute environmental effects and safety effects. Assuming that we *are* able to compute user benefits and other elements of welfare for different groups of agents in the model, the computation of equity effects poses no additional problems.

6 Conclusion

Planning for a sustainable land use/transport system is a multi-objective task. Recent work of Chichilnisky (1996) and Heal (2000) has shown that three of the objectives – intergenerational equity, economic efficiency, and conservation of natural resources – can be fully and

consistently accounted for in the framework of welfare economics. Thus the problem to find an optimal sustainable transport/land use plan should be cast as the dynamic optimisation problem to maximise a Chichilnisky welfare function subject to the appropriate constraints.

The most important exhaustible resources involved are “land” and “atmosphere”. Targets reflecting our best guesses about the stocks of different types of “unused” land that we should retain forever should be included as constraints in the optimisation problem. With respect to land use that keeps within these constraints, we need estimates of the willingness to pay to save this land from being built down. The social cost of CO_2 emission at any point in time should be derived from international or national political targets thought to apply at that time and be included in the objective function in the same way as any other cost.

Indicators of accident costs, noise costs and air emission costs are becoming standard in transport appraisal and pricing, and may be included in the economic efficiency indicator. At the same time, targets with respect to accidents, noise and local air quality may also be used as constraints in the optimisation problem.

The remaining sustainability objectives concern intragenerational equity and social exclusion, the liveability of streets and neighbourhoods, and economic growth. Along with financial constraints, equity concerns are vital for the feasibility and implementation of transport/land use strategies, and they need to be included as constraints in the problem. The PROSPECTS project developed a wide range of equity indicators for possible use in planning for sustainability. More work needs to be done to develop indicators of the liveability of streets and neighbourhoods at the level of strategic planning, and economic growth impacts are still inherently difficult to forecast.

Appraisal with respect to sustainability in this framework is different from ordinary cost benefit analysis in the following respects:

1. We use a parameter for equity between generations. It may be seen as a special form of discounting with a decreasing discount rate. Its level is a normative choice.

2. The shadow prices of unused land and ‘atmosphere’ are set to reflect their value as stocks and are assumed to increase with depletion.

3. Strategies that do not meet preset environmental targets, financial constraints and other political targets by the end of the appraisal period are discarded.

This approach “dissolves” the problem of what discount rate to use in problems involving long-term effects, and avoids the pitfall of using “sustainability” as an excuse to reduce the requirements on any project that may be termed long-term. For instance, investments will only be accepted if they reduce CO_2 or land use in the long term or contributes appreciably to other legitimate sustainability objectives.

The PROSPECTS framework for the appraisal of strategic transport/land use plans with respect to sustainability may be used to test selected strategies or to actually carry out optimisation with respect to the available policy instruments. For optimisation, the model systems need to be fairly simple, while still taking account of the whole range of behavioural responses to the strategies and reflecting accidents and environmental impacts in sufficient detail. Experience has shown that the framework is able to appraise dynamic strategies in models that take account of the different timescales of the adjustment processes in the land use/transport system. New general rules of action may perhaps be derived from such tests.

References

Brundtland Commission (World Commission on Environment and Development), 1987. Our Common Future. Oxford University Press, UK.

Chichilnisky, G., 1996. An axiomatic approach to sustainable development. Social Choice and Welfare 13(2), 231-257.

Chipman, J.S. and Moore, J.C., 1994. The Measurement of Aggregate Welfare. In: Eichhorn, W. (Ed.), Models and Measurement of Welfare and Inequality. Springer-Verlag, Berlin.

Coase, R.H., 1960. The problem of social cost. The Journal of Law and Economics 3, October issue, 1-44.

Fridstrøm, L., Minken, H., Moilanen, P., Shepherd, S.P. and Vold, A., 2000. Economic and equity impacts of marginal cost pricing in transport. Case study from three European cities, VATT Research Reports 71, Helsinki.

Heal, G., 2000. Valuing the Future. Economic Theory and Sustainability. Economics for sustainable earth series, Columbia University Press, New York.

Hotelling, H. 1931. The economics of exhaustible resources. Journal of Political Economy 39, 137-175.

IPCC (Intergovernmental Panel on Climate Change), 1996. Climate Change 1995: Economic and Social Dimensions of Climate Change. Contribution of Working Group III to the second assessment report of the IPCC, Cambridge University Press, Cambridge.

IPCC, 2001. Climate Change 2001: Mitigation. Contribution of Working Group III to the third assessment report of the IPCC, Cambridge University Press, Cambridge.

May, A.D. 2003. Developing Sustainable Urban Land Use and Transport Strategies. A Decision Makers' Guidebook. PROSPECTS Deliverable 15.

Minken, H. 1999. A sustainability objective function for local transport policy evaluation. In: Meersman, H., van de Voorde, E. and Winkelmans, W. (Eds.), Selected Proceedings of the 8th World Conference on Transport Research. Volume 4: Transport Policy, Pergamon, Amsterdam.

Minken, H., Jonsson, D., Shepherd, S.P., Järvi, T., May, A.D., Page, M., Pearman, A., Pfaffenbichler, P., Timms, P. and Vold, A., 2003. Developing Sustainable Urban Land Use and Transport Strategies. A Methodological Guidebook. PROSPECTS Deliverable 14, TØI Report 619/2003, TØI, Oslo.

Newman, P and Kenworthy, J (1999) Sustainability and Cities. Overcoming Automobile Dependence. Island Press, Washington D.C.

OPTIMA Consortium, 1998. OPTIMA. Optimisation of policies for transport integration in metropolitan areas. Office for official publications of the European Communities, Luxembourg.

Parry, I.W.H. and Bento, A.M.R., 1999. Revenue Recycling and the Welfare Effects of Road Pricing. World Bank Policy Research Working Paper 2253, Washington DC.

Portney, P.R. and Weyant, J.P. (Eds.), 1999. Discounting and Intergenerational Equity. Resources for the Future, Washington DC.

Standing Advisory Committee on Truck Road Assessment (SACTRA), 1999. Transport and the Economy. Department of the Environment, Transport and the Regions, UK.

Varian, H.R., 1992. Microeconomic Analysis. Third Edition. W.W. Norton & Company, N.Y.

Wegener, M., 1994. Operational urban models: State of the art. Journal of the American Planning Association 60, 17-29.

Weitzman, M., 1998. Why the Far-Distant Future Should Be Discounted at Its Lowest Possible Rate. Journal of Environmental Economics and Management 36, 201-208.

2.6 A finite horizon model of an exhaustible resource

APPENDIX
A finite horizon model of an exhaustible resource [22]

27/05/2023

Harald Minken

Institute of Transport Economics

To apply the Chichilnisky approach in practice, we need to convert the problem to a problem of finite horizon. At the end of the appraisal period, more or less sustainable conditions are assumed to exist. Benefits from then on will have to be included in the appraisal. We model this as a control theory problem with a restriction on the end state and a scrap value. There is a single exhaustible resource (the extension to more resources is trivial). Consumption of this resource is denoted by c, and the stock is denoted by s. Stocks give utility. The current stock is s_0 and the targeted "sustainable" end-of-period level of the stock is s_1. We assume a separable utility function $u(c,s) = u_1(c) + u_2(s)$. The functions u_1 and u_2 are increasing and concave. Denoting the end of the appraisal period by t_1, the scrap value is

(A1) $$S\left(t_1, s(t_1)\right) = \left(1-\alpha\right)\left(u_1(0) + u_2(s(t_1)\right)e^{rt_1}$$

The level of consumption c(t) is our control function. The problem is:

(A2)
$$\begin{aligned} &Max \quad W_f = \int_0^{t_1} \alpha\left(u_1\left(c(t)\right) + u_2\left(s(t)\right)\right)e^{-rt}dt + S\left(t_1, s(t_1)\right)e^{-rt_1} \\ &\dot{s} = -c(t) \\ &s(0) = s_0 \\ &s(t_1) \geq s_1 \\ &c(t) \in \left[0, s_0 - s_1\right] \quad \forall t \in \left[0, t_1\right] \end{aligned}$$

[22] This is the appendix to a TOI report, dated 27/05/2003.

Except for the finite horizon – which does make a difference, as we will see – this is a Chichilnisky problem. The current value Hamiltonian is:

(A3) $$H^C\left(t, s(t), c(t), \lambda(t)\right) = \lambda_0 \alpha\left(u_1(c(t)) + u_2(s(t))\right) - \lambda(t)c(t)$$

The maximum principle applied to this problem is (see for instance Theorem 3.3 in Seierstad and Sydsæter (1987)):

Assume $\left(s^*(t), c^*(t)\right)$ is an admissible pair that solves the problem. Then there is a constant λ_0 and a continuous and piecewise continuously differentiable function $\lambda(t)$ such that

(A4) $$\lambda_0 = 0 \text{ or } 1 \text{ and } \forall t \in [0, t_1], \left(\lambda_0, \lambda(t)\right) \neq (0,0)$$

(A5) $$\forall t \in [0, t_1],\ c^*(t) \text{ maximises } H^C\left(t, s^*(t), c(t), \lambda(t)\right) \text{ on } [0, s_0 - s_1]$$

(A6) $$\text{Except for discontinuity points of } c^*(t),\ \dot{\lambda} - r\lambda = -\frac{\partial H^C(t, s^*, c^*, \lambda)}{\partial s}$$

(A7) $$\lambda(t_1) \geq \lambda_0 \frac{\partial S^*(t_1, s^*(t_1))}{\partial s} \text{ (with equality if } s^*(t_1) > s_1)$$

From (A2), we also immediately get

(A8) $$s(t) = s_0 - \int_0^t c(x)dx$$

If we had required that $s(t_1)$ should reach the target s_1 exactly at the end of the period, then instead of the inequality $s(t_1) \geq s_1$ that we included in (A2), the transversality condition (A7) would simply be deleted. In this case, the scrap value would have been determined already by the constraints in the problem. Whatever path of depletion is chosen, it would always end up in s_1, and thus the scrap value would have been a constant and could have been deleted from the problem. Thus the choice of the parameter α would be entirely without consequences for the solution. Before looking closer at the solution, it is worthwhile to digress and see if this fact can throw some light on our method of appraisal in PROSPECTS.

Some experiences with the PROSPECTS approach

The control theory problem (A2) is the same as the problem that we solve with our PROSPECTS method, except for two things. First, we include more constraints or penalty terms in PROSPECTS. This is not all that important – we might have included the same constraints in the control theory problem and got qualitatively similar results. But secondly, the utility functions u_1 and u_2 are much simpler than the corresponding parts of the objective function in PROSPECTS. They are determined solely by the consumption and stock of the single resource. In PROSPECTS, their levels are influenced by many other factors. Since we use many different policy instruments in conjunction, it must be thought that some of them could be used primarily to attain the end-of-period target s_1, while others might be used to maximise annual benefits subject to this constraint. If this was the case, these other instruments would also have been able to influence the scrap value. Thus the scrap value would not have been the same for all paths, and the parameter α would make a difference for the solution.

Experience from the UK in particular (Timms, May and Shepherd 2002) shows that the optimal solution is surprisingly insensitive to α. Thus there is virtually no trade-off between the objectives of reaching the resource target and maximising annual benefits at the end of the period. There might be two explanations for this. Either the target is extremely difficult to reach within the bounds on the use of the instruments, so that we have to use extreme policies and disregard annual benefits. In this case, the method collapses to a simple backcasting procedure. Or the two objectives are in fact only different expressions of the same. This might be due to the particular assumptions embodied in the transport and land use model, or might be a more general and objective feature of the system we are studying.

Since the approach allows us to change policies in the middle of the period, it should be possible to say which of the two possibilities that actually is true by comparing optimal policies at the start and at the end of the period.

Many studies confirm that any seemingly realistic CO_2 target for transport is very difficult to attain within acceptable bounds on the normal transport policy instruments, and that if it is reached, a lot of other environmental targets are automatically overreached. Whether or not this is also true for transport benefits is debated. Some studies like OECD's EST study sees no conflict between a CO_2 targets and an efficient transport system, while most do. Anyhow, the finding that the parameter α does not make a difference indicates a need for more instruments, that could make the trade-off between objectives more feasible. If we included technological development as an instrument in our studies, there would obviously have been more scope for trade-offs between objectives, and the issue of taking the interests of future generations into account would have been less tightly tied to a end-of-period CO_2 target.

Properties of the solution to problem (A2)

We now return to the problem at hand. Assume $\lambda_0 = 0$. Then according to (A5), $c^*(t) = 0$ for all t. Looking at (A6), the right hand side must be 0 and thus $\lambda(t) = Ce^{rt}$ for some constant C. But this contradicts (A7), because $\lambda(t_1)$ cannot attain 0 even if $s^*(t_1) > s_0$. Thus $\lambda_0 = 1$.

Since u_1 is concave, H^C as a function of c is concave. At any particular point in time, there are three solution candidates for (A5), namely c(t) = 0, c(t) = $s_0 - s_1$ and the inner solution.

If $c^*(t) = 0$, $\frac{\partial H^C}{\partial c} = \alpha \frac{\partial u_1}{\partial c}\Big|_{c=0} - \lambda \le 0$,

if $c^*(t) = s_0 - s_1$, $\frac{\partial H^C}{\partial c} = \alpha \frac{\partial u_1}{\partial c}\Big|_{c=s_0-s_1} - \lambda \ge 0$,

and if there is an inner solution, $\frac{\partial H^C}{\partial c} = \alpha \frac{\partial u_1}{\partial c} - \lambda = 0$.

In analogy with the infinite horizon problem (Heal 2000) it is natural to *guess* that if the resource is consumed at all, there will be an initial phase where consumption occurs, followed by a phase where nothing more is consumed. At any rate, implicit differentiation of (A8) with respect to t_1 yields $c(t_1) = 0$. Let t* be the point in time where this shift occurs (t* might be 0).

The marginal utilities u_1' and u_2' will be constants on [t*,t_1], while the inner solution applies on [0,t*] and thus u_1' is proportional to $\lambda(t)$.

Treating $\lambda(t_1)$ as a constant, the solution to the differential equation (A6) is:

(A9) $$\lambda(t) = \lambda(t_1)e^{-r(t_1-t)} + \alpha\int_t^{t_1} \frac{\partial u_2}{\partial s} e^{-r(\tau-t)} d\tau$$

Using (A7),

(A10) $$\lambda(t) \geq (1-\alpha)\frac{\partial u_2}{\partial s}\bigg|_{s=s^*(t_1)} \cdot e^{rt} + \alpha\int_t^{t_1} \frac{\partial u_2}{\partial s} e^{-r(\tau-t)} d\tau$$

If consumption stops before the target value s_1 is reached, there is equality in (A10).

Now if our guess was right, there might be some interval [t*, t_1] on which the partial derivatives involved in (A10) are constant and equal. For t in this interval, some simple calculation gives

(A11) $$\lambda(t) \geq \alpha e^{rt} \frac{\partial u_2}{\partial s}\left[\frac{1-\alpha}{\alpha} + \frac{1}{r}\left(e^{-rt} - e^{-rt_1}\right)\right]$$

The basic problem of transforming the Chichilnisky problem to a finite horizon problem can be seen from (A11). Apart from two minor differences – the unit of account which is α times the unit of account used in Heal (2000) and the current value formulation of the adjoint function used here – the real difference of (A11) from the stationary state shadow price given in Heal is that the capital value of the resource is reduced, and even vanishes when t = t_1. (On the other hand, (A11) is an inequality).

Somehow the missing capital value should be put back in in the form of a modification of the scrap value. Alternatively, we might delete the scrap value and use the shadow price from Heal with an ordinary cost benefit analysis objective function. However, what we loose then is the far-away benefits that are derived from other elements than the use and non-use of the exhaustible resource.

To actually solve the problem (A2) would require us to specify the utility function. However, some elements of the solution procedure can perhaps be outlined. Using (A8), the variable s(t) can be eliminated. Setting the two expressions for $\lambda(t)$, (A9) and $\lambda = \alpha\frac{\partial u_1}{\partial c}$ equal to each other would give an expression for consumption c*(t) on [0,t*] as a function of t*. Some other equation must be used to determine t*. Finally, some sufficiency theorem must probably be used to prove that the solution found is actually the right one, confirming our guess about the optimal path.

References

Heal, G. (2000). Valuing the Future. Economic Theory and Sustainability. Economics for sustainable earth series, Columbia University Press, New York.

Timms, P.M., May, A.D. and Shepherd, S.P. /(2002). The senitivity of optimal transport strategies to specification of objectives. Transportation Research A, 36(5), pp. 383-401.

2.7 Project selection with sets of mutually exclusive alternatives

Project selection with sets of mutually exclusive alternatives[23]

Harald Minken
Institute of Transport Economics
E-mail: hm@toi.no

Contents

[23] This is an Accepted Manuscript of the article published in: Economics of Transportation 6, 2016, 11-17. 2212-0122. The article has been published in final form by Elsevier at https://dx.doi.org/10.1016/j.ecotra.2016.06.001.

1 Introduction

We study the problem to maximise the net economic benefit of an investment plan by selecting from a portfolio of candidate projects within a given budget constraint. One example would be the national transport plans in countries like Norway and Sweden. Assuming independent projects, i.e. (1) all projects may be selected regardless of which other projects are selected, and (2) their benefits and costs stay the same regardless of which other projects are selected, the economic efficiency of the entire investment plan is maximised if projects are selected according to their benefit-cost ratio until the budget is exhausted. To be exact, this result requires projects to be infinitely divisible, but the divisibility matters only for the last project to be included in the plan, and so is of little consequence if projects are small compared to the budget.

Normally, however, the planning of a project involves a stage where a set of alternative concepts or designs are considered. A best alternative is chosen, and the plan is composed from the pool of all such best alternative solutions. This two-step procedure violates the assumptions underlying the benefit-cost ratio criterion, and in fact, neither the benefit-cost ratio nor the net present value of a project is a valid choice criterion in this case.

In this paper, we set out the correct criterion to use in this case. It is not the first time this criterion had been proposed. Actually, it was proposed as early as 1955 by Lorie and Savage, but even if it was commented upon by authors such as Weingartner (1963, 1966) and others in the sixties, it obviously got lost in the subsequent more and more complex development of the capital budgeting literature. We show that the criterion is the solution to a one-period continuous knapsack problem with mutually exclusive project alternatives, and that an approximate solution can be found by a simple iterative procedure, just like Lorie and Savage said.

Section 2 prepares for the derivation of the Lorie and Savage criterion in section 3. This it does by reminding the reader of how the benefit cost criterion is derived: It is the solution to a linear programming problem called the continuous knapsack problem with independent projects. The assumptions underlying this problem are necessary and sufficient conditions for the benefit cost criterion to be valid. Changing the assumption of independent projects to projects with mutually exclusive alternatives must produce a different criterion, namely the Lorie and Savage criterion, as shown in section 3. In section 4, we illustrate the way this criterion functions in a real life example from Norwegian transport planning. In section 5, we briefly discuss the situations when the new criterion might be of use and its implication for the possibility of local decisions. Section 6 concludes.

2 The benefit cost ratio

Judging from the HEATCO survey of how cost benefit analysis is practised in 25 European countries (HEATCO 2005a and b), some confusion still exists about the definition of costs to be used in the benefit cost ratio, about its relationship to the net present value and other commonly used indicators, and about the conditions for its validity as a decision-making tool. Even the HEATCO recommendations themselves (HEATCO 2006) are plainly wrong when they define costs (to be entered in the denominator of the ratio) as the resource consumption of transport providers and government, and benefits (to be entered in the nominator) as the resource gains of

travellers and third parties. This is shown in this section. We also show the necessary and sufficient conditions for the benefit cost ratio to be a valid criterion for project selection.

Nearly all of the countries surveyed in HEATCO report that they combine the benefit cost ratio and the net present value. Many of them provide a clear description of when to use the one or the other, but there seem to be some that use some undefined mix of them. Furthermore, fairly many countries use the internal rate of return to compare projects (a criterion that is not suitable for comparing mutually exclusive options, and that may produce wrong results unless all costs occur before all benefits), or even the payback period (a practise that does not take all relevant costs and benefits into consideration).

Assume that our objective is to maximise the net present value of a plan within a given budget constraint. The candidate projects are assumed to be infinitely divisible and mutually independent. That is, any fraction of the costs of a given project will produce a similar fraction of the benefits, and the costs and benefits of a candidate project is not at all dependent on which of the other projects that are included in the plan. There are no other objectives than maximisation of net present value, and no constraints or conditions other than the given budget constraint. We want to show that the necessary and sufficient condition to achieve our objective under these circumstances is that we select projects in descending order of their benefit cost ratio (with costs defined as net outlays over the relevant public budget) until the budget is exhausted. To keep within the budget, only a fraction of the last selected project can normally be implemented.

2.1 The solution to a linear programming problem

Let $\boldsymbol{b} = (b_1,\ldots,b_n)$ be the net present benefit of n candidate projects, some of which are to be chosen to form the plan of a government agency. Let $\boldsymbol{c} = (c_1,\ldots,c_n)$ be the vector of discounted net payments that the agency must incur if these candidate projects are to be included in the plan. We assume there is a constraint a on the net present value of the agency's budgets in the period we consider.

The assumption of such a constraint seems to contradict one of the implicit assumptions of discounting, namely free lending and loaning at the same interest rate. The contradiction is resolved if we assume that the constraint is imposed by a political decision at a higher level of government, as it usually is. Such a decision may make sense even if the margin between the lending and loan rate for the government is very small, because the agency's spending involves not just money, but real resources in short supply.

The n prosjects are infinitely divisible. That is, if we carry out only a part of a project, as measured by budget outlays, we will always achieve the same part of the project's net benefits. This is certainly not always reasonable, but it matters less and less the smaller the projects are as parts of the budget. Let $\boldsymbol{x} = (x_1,\ldots,x_n)$ be the parts of each of the projects that are implemented. Thus $x_j \in [0,1]$ for all x_j. Finally, we assume that all projects are independent of each other, i.e., no element of $\boldsymbol{b}$ and $\boldsymbol{c}$ are functions of $\boldsymbol{x}$. If this seems to be a problematic assumption in any given case, it can often be solved by forming all possible combinations of the interdependent projects and enter these combinations instead of the inter-dependent projects themselves. But what we have then are mutually exclusive alternatives, and the rule of section 3 must be applied.

Projects that do not require any part of the budget can be decided upon separately, and projects with negative net benefits should always be discarded. Thus we may assume without

problems that all elements of ***b*** and ***c*** are strictly positive and that all elements of ***b*** are larger than or equal to their corresponding element of ***c***.

The linear programming problem (LP1) based on these assumptions can now be formulated. Implicitly, it is also assumed that there are no binding restrictions other than the budget on the selection of projects. For example, there is no quantified target for the reduction of climate gas emissions.

$$\text{(LP1)} \qquad \max_{\mathbf{x}} \ \sum_{j=1}^{n} b_j x_j \quad \text{s.t.} \ \sum_{j=1}^{n} c_j x_j \le a \ \text{ og } \ x_j \in [0,1] \ \ \forall j$$

The solution to the problem (LP1) is to arrange the candidate projects after their cost benefit ratio b_j/c_j and select them from the top until the budget is used up. Say that the candidate projects are numbered so that $b_1/c_1 \ge b_2/c_2 \ge \ldots \ge b_n/c_n$. If we select them in the order 1, 2, 3, … and so on, we will ultimately come to a project number r such that the sum of the $r-1$ first costs c is less than the budget a, while the sum of the r first is greater than a. Formally, the solution can be written as Equation (1) on the next page.

The formal proof that (1) is indeed the solution requires use of the Simplex method, see any textbook in linear programming. An intuitive argument is this: Assume, contrary to (1), that the solution is to exclude some project with a higher benefit cost ratio b_j/c_j than at least one of the r projects selected by (1). If we take out a small slice of project r and replace it by a similar slice of this excluded project, the objective function must increase. Thus in the the optimal solution, all selected projects must have higher benefit cost ratios than any project not selected.

(1)
$$x_j = \begin{cases} 1 & \text{for } j = 1,\ldots,r-1 \\ \dfrac{a - \sum_{j=1}^{r-1} c_j}{c_r} & \text{for } j = r \\ 0 & \text{for } j = r+1,\ldots,n \end{cases}$$

Diagram 1 illustrates our finding. There, all projects are ordered by the benefit cost ratio (BCR) and entered in the diagram as columns of different height and width. The width of a column is its cost, c_j, and the height is the benefit cost ratio b_j/c_j. Since $c_j*(b_j/c_j) = b_j$, the area of column j represents the net present value of project j. The vertical line a represents the budget. The area of all columns to the left of a is the net benefit of all projects financed within the budget. It is seen that the only project that have to be divided is project 6. On the right side of the line a are the projects that are excluded from the plan. We have just concluded that the gros benefit of the plan is maximised if projects are selected according to their BCR. Since the cost of the plan always equals the constant a, this strategy also maximises the net present value of the plan.

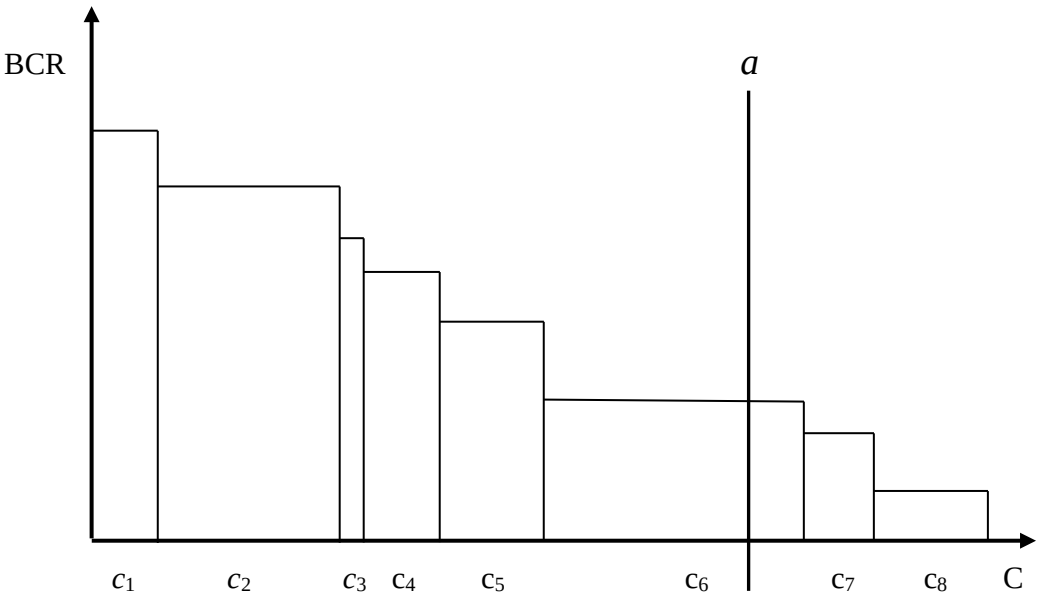


Diagram 1. 8 projects with costs along the C axis and benefit cost ratio along the BCR axis. Project 6 is only partially financed within the budget *a*.

2.2 The denominator

It is obviously of some importance to be clear about what budget constraint is considered as the binding one. It may be defined at the level of a plan or a programme or a whole sector of government. The higher the level, the more types of payment goes into the denominator ***c***. For instance, revenue from user charges must be included if they are earmarked to be used to finance the plan or to be used within the sector. If there is no earmarking, the revenues created by a project should not go into the denominator. They are however entered in the nominator, as part of the net benefits.

2.3 Maximising net present value of the plan within the budget constraint

We have just found that, given our assumptions, the net present value of a plan is maximised if projects are selected according to their BCR, defined as gros benefits divided by the discounted net payments that the agency must incur if this candidate project is included in the plan. It may be more intuitive to redefine ***b*** as the vector of net benefits, not gros benefits, and maximise net benefits instead.

To this end, define the net present value of project j as $b'_j = b_j - (1+\lambda)c_j$, where λ is the marginal cost of public funds, b_j is net benefits to travellers, transport operators and other affected parties, and c_j is the net present value of payments in and out over the relevant government budget. The net present value of the plan is

$$NPV = \sum_{j=1}^{n} b'_j x_j = \sum_{j=1}^{n} \left(b_j - (1+\lambda)c_j\right)x_j \tag{2}$$

Our problem becomes

$$\text{(LP2)} \qquad \max_{\mathbf{x}} \ \sum_{j=1}^{n} b_j' x_j = \sum_{j=1}^{n} \left(b_j - (1+\lambda) c_j \right) x_j \quad \text{s.t.} \ \sum_{j=1}^{n} c_j x_j \le a \ \text{ og } \ x_j \in [0,1] \ \ \forall j$$

Contrary to (LP1), some of the $b_j' x_j$ terms are likely to be negative, but these can be eliminated in advance without consequence for the optimal solution. The solution to (LP2) follows immediately from (LP1) by substituting b_j' for b_j. We have:

$$\text{(3)} \qquad \frac{b_j'}{c_j} = \frac{b_j - (1+\lambda) c_j}{c_j} = \frac{b_j}{c_j} - (1+\lambda)$$

The solution to (LP2) is therefore exactly the same as the solution to (LP1). All ratios are reduced by $1 + \lambda$, but that does not affect the ranking. Thus if the BCR is properly defined and related to a single binding budget, as it should, there is no reason to make a distinction between the BCR and the criterion they call RNPSS (ratio of NPV to public sector support), as HEATCO (2006) does. Actually, official Norwegian guidance uses the (LP2) formulation instead of (LP1).[24]

What we have shown in this chapter, is that given the assumptions, any procedure that produce the solution (1) may be used, but no procedure that does not produce solution (1) is valid.

3 The case of mutually exclusive alternatives

According to HEATCO (2005a and b), quite a few countries point to the benefit-cost ratio as the only correct criterion to use if the objective is to maximise the net present value of a plan that is constrained by a single budget. Some, as an old Norwegian manual (Finansdepartementet 1979), even care to mention that this criterion breaks down if there are interdependencies between the projects. But none of them propose any alternative criterion for the case of mutually exclusive alternatives. This is our task in this section. We start by explaining the general idea, before formulating and solving the problem in a more formal way.

Let us assume that we have $n - 1$ independent candidate projects plus a candidate project number n with two mutually exclusive alternative designs. We order the $n - 1$ candidates by the BCR. We use the formulation of this criterion given in (3), so that zero is the demarcation point between profitable and unprofitable projects. Assume that the last projects to fit into the budget all have BCR's close to a certain number k. Now we want to find out which one, if any, of the two alternative designs that deserve to be included in the plan at the expense of one or more of these marginal projects. If both alternatives have CBR below k, none of them qualifies. If only one of them has BCR above k, this

[24] See the website of the Norwegian Ministry of Finance. But the Ministry's experts were wrong in thinking that by substituting (LP2) solutions for (LP1) solutions, they had made a substantial improvement.

alternative should be included at the expense of one or more of the projects whose BCR is *k*. What about the case where both alternatives have BCR's above *k*?

If we use a diagram similar to Diagram 1, with BCR on the vertical and *C* on the horizontal axis, the net present value of the whole plan is equal to the area of all colums to the left of the budget line *a*. It is this area that we want to make as large as possible. Let us say that one of the alternatives has a BCR considerably above *k*, while the other has has a somewhat lower BCR, but still above *k*. Obviously, we must choose the alternative with the largest area above the *k* level. The areas are not only dependent on the columns' heights (the BCR) but also on their widths (the C).

Diagram 2 shows a such case. The area DEFG in the diagram is a string og projects with a BCR of *k*. The cost of the projects with benefit D is c_0, and the costs of the projects with benefit E plus F is $c_2 - c_0$. The first competing alternative of project *n* has cost $c_1 - c_0$, a net present value of ABE and a BCR equal to the *height* of ABE. The second alternative has a cost of $c_2 - c_0$, a net present value of BCEF and a BCR equal to the height of CF (or BE).

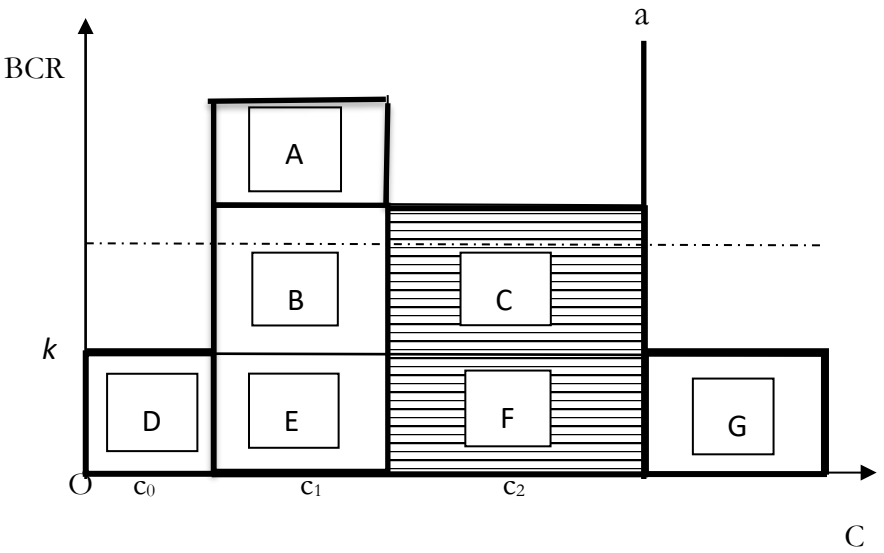


Diagram 2. A choice between mutually exclusive project alternatives ABE and BCFE.

As the area of A + B is smaller than the area of B + C, it is the second alternative that should be chosen, If however the *k* line had been at the dotted line instead, or even higher, the first alternative should be chosen. Thus the BCR of the marginal project included in the plan matters for the choice between the two alternatives of the n^{th} candidate project. The BCR of the two candidates is not a valid choice criterion in this case, or else project ABE would have won regardless of *k*. Likewise, the net present value is not a valid criterion, or else BCEF would have won in all cases. The reason is that the relative size of the boxes E and F has nothing to do with the choice here, since the net present value they represent is secured anyhow by choosing other projects than *n*. It is the area above the *k* level that matters.

Call the two alternatives *n1* and *n2*. The first alternative has a benefit above the *k* level (a surplus benefit) of W_{n1} = A + B, while the surplus benefit of the second alternative is W_{n2} = B + C. Let h_{in} denote the benefit cost ratio as defined by (3). We have:

$$W_{in} = \left(\frac{b_{in} - (1+\lambda) c_{in}}{c_{in}} - k \right) c_{in} = \left(\frac{b'_{in}}{c_{in}} - k \right) c_{ni} = (h_{in} - k) c_{ni}, \ i = 1, 2 \tag{4}$$

The indicator W_{in} can obviously be used to maximise economic efficiency when all projects except one are without alternatives. But we will now show that it is much more general than that. In fact, it is the choice criterion for all candidates to be included in a plan when at least one of them has mutually exclusive alternatives.

Assume that the plan in case is a national transport plan. Let P be the set of candidate projects, and index the elements of this set by $j \in P$. Call the set of mutually exclusive alternatives of project j by Aj, and index this set by $i \in Aj$. Observe that the number of elements in A_j may be different for each j, and that it may also consist of just one element. Let b_{ij} be the net present value of benefits to travellers, freight owners, transport sector operators and infrastructure providers from alternative i of candidate project j, and let c_{ij} be the net outlays over the constrained budget in case alternative j of project i is included in the plan.

We assume infinitely divisible projects and formulate the following linear programming problem to maximise the net present value of the plan:

$$\text{(LP3)} \qquad \max_{\mathbf{x}} \ \sum_{j \in P} \sum_{i \in A_j} \left[b_{ij} - (1+\lambda) c_{ij} \right] x_{ij} \quad \text{s.t.} \ \sum_{j \in P} \sum_{i \in A_j} c_{ij} x_{ij} \le a$$

$$\sum_{i \in A_j} x_{ij} \le 1, \quad j = 1, 2, ..., |P|$$

$$x_{ij} \in [0,1] \quad \forall j \in P \text{ og } i \in A_j$$

Here, a is the budget and $|P|$ is the number of candidate projects. Thus the first constraint is the budget constraint, while the second (or to be correct, the following $|P|$ constraints) says the the fractions of each alternative design of a project must sum to at most 1. We will see that in practice, this implies that at most one alternative will be chosen.

To select just one alternative i for each of the j projects can sometimes be done in millions of ways. It is difficult to test out every possibility. Thus to solve the problem, we formulate a similar problem that under certain circumstances will produce the same solution as (IP1), but that is much easier to solve, at least approximately. What we do is to delete the budget constraint from the problem and include it in the objective function instead, multiplied by an unknown parameter that we shall call k. This procedure is called Lagrangian relaxation. The new objective function becomes:

$$V(k) = \sum_{j \in P} \sum_{i \in A_j} \left[b_{ij} - (1+\lambda) c_{ij} \right] x_{ij} - k \left(\sum_{j \in P} \sum_{i \in A_j} c_{ij} x_{ij} - a \right) \tag{5}$$

We may perform some simple rearrangements of the right hand side of (5):

$$V(k) = ka + \sum_{j\in P}\sum_{i\in A_j}\left[\left(b_{ij} - (1+\lambda)c_{ij}\right) - kc_{ij}\right]x_{ij} = ka + \sum_{j\in P}\sum_{i\in A_j}\left(\frac{b_{ij} - (1+\lambda)c_{ij}}{c_{ij}} - k\right)c_{ij}x_{ij}$$
$$= ka + \sum_{j\in P}\sum_{i\in A_j}\left(h_{ij} - k\right)c_{ij}x_{ij}$$

To optimise the modified objective function $V(k)$ for a given k means to find the optimal fractions of all the x_{ij}. Since k is a parameter and not a variable, ka is a constant that does not affect the optimisation. The sum of sums of the last line is easily seen as the sum over all projects and all project alternatives of the indicator W_{ij} of (4). Maximisation of this expression obviously means to select from each candidate project the alternative with the highest indicator value, then select projects according to the BCR, starting with the highest value and proceeding until the budget is exhausted. This procedure only requires identification, for each project, of the project alternative with the highest value, then a simple ordering of the projects. Obviously, it can be done by in an EXCEL spreadsheet.[25]

But this optimisation is conditional on k. Thus we must also choose the k that produces the best result. And for each chosen k, we need to repeat the same procedure to select the optimal set of x_{ij}'s. This choice of optimal k is also quite simple. The lowest possible value is $k = 0$. If the whole budget is not used up at this level of k, this is the optimal solution k^* (and the problem becomes just to find the alternative of each candidate project with the highest net present value, and add all such values that are above zero). If not, our first task is to find a k large enough that the budget is not used up. We have then an interval between 0 and this k on which the optimal k, k^*, must lie. On this interval, a search algorithm may then be applied to find the lowest k that does not use up the whole budget. The choices that follow from using this value of k, produces the optimal plan.[26]

Observe that for each new choice of k, the computation of all w_{ij} must be repeated. Then the sum of all cost for the projects with positive w_{ij} must be computed and compared to the given budget. The k is then adjusted so as to utilise as much as possible of the budget, but not more. This approach utilises the fact that the optimal k is both the Lagrangian multiplier of the budget constraint and the BCR of the last project to fit into the budget. A possible algorithm may consist of the following steps:

Sub routine

For any given k, (a) compute $w_{ij}(k)$ for all projects with project alternatives, (b) for all projects j, select the i with the highest $w_{ij}(k)$, (c) eliminate all remaining j with $w_{ij}(k) < 0$, and (d) add together all remaining c_{ij}. Call this sum $C(j;k)$.

---

[25] The h_{ij}'s and the c_{ij}'s are output from the cost benefit analysis of every alternative of every candidate project, and is treated as given in the subsequent choice of projects and project designs for the plan.

[26] If projects are indivisible, our procedure does not garantee an optimal result, only a result close to the optimal. Since a small part of the budget is not used, it might be that some project that uses up more of the budget, but has a slightly lower indicator value than the ones we selected, might improve the economic value of the plan.

Main routine

1. Set $k_0 = 0$. If $\sum_j C(j;k_0) < a$, compute the net present value of the optimal plan consisting of all selected *ij* with $w_{ij}(k_0) > 0$, and stop.
2. If $\sum_j C(j;k_0) > a$, make a guess at a $k_1 > k_0$. Perform the subroutine. If $\sum_j C(j;k_1) < a$ increase k_1 until $\sum_j C(j;k_1) > a$. Call this value of k_1 $\overline{k}$.
3. Use some search algorithm on the interval $[0,\overline{k}]$ to find points k_2, k_3 etc. with smaller and smaller $\left|\sum_j C(j;k_n) - a\right|$ until for k^* and some prespesified ε, $\varepsilon > a - \sum_j C(j;k^*) > 0$. Perform the subroutine and stop.
4. Compute the net present value of the optimal plan, $V(k^*)$.

This procedure secures that all eliminated projects with only one alternative have BCR < k^* and all eliminated alternatives are dominated by another alternative, while all retained projects and alternatives have $BCR \geq k^*$. Thus there is no point in substituting a fraction of an eliminated alternative for a fraction of some retained project or alternative. Furthermore, there is very little space between $\boldsymbol{a}$ and $\sum_j C(j;k^*)$, so it is not much point in squeezing in a fraction of a deleted project there. Therefore, the routine produces a plan that is as close to optimum as one wishes.

4 An example

The E39 is a major road in the western part of Norway. At present, a ferry carries the traffic on E39 across one of the major fjords in the area, Bjørnafjorden. The main issue in this study is to find the best conceptual solution for the crossing of Bjørnafjorden in the future. A cost benefit analysis is part of the quality assessment of this choice of concept.

The competing concepts studied are:

K2	minor improvements on the current road
K3	bridges from island to island in the outer part of the fjord
K4A and K4C	long/short variants of a large bridge further into the fjord
K4D	like K4C, except with ferry connection instead of bridge
K5A and K5B	bridge solutions even further into the fjord

Table 1, based on Table 7-2 in Dovre and TØI (2012), shows the main results of the cost benefit analysis of different conceptual solutions in the E39 Aksdal-Bergen project.

From these numbers, we can compute h_i and c_i for i = 2, 3, 4A, 4C, 4D, 5A and 5B. We can then compute $w_i(k)$ for different values of k between 0 and 2. This is done in Table 2. To be precise: Investments and running costs from Table 1 are added to form c in Table 2. BCR in Table 1 is entered as h in Table 2. Finally, the indicator $w = (h - k)c$ is computed for the different values of k.

Table 1: E39 Aksdal-Bergen. Key numbers from the cost benefit analysis of the quality appraisal. (NOK Billion in 2012 prices).

	K2	K3	K4A	K4C	K4D	K5A	K5B
Investment	4,3	28,9	12,6	27,1	12,2	25,8	21,2
Running cost*	1,0	3,6	1,2	2,0	1,4	2,2	2,2
Gros benefit	5,5	62,9	26,7	56,5	34,8	55,7	52,3
Net present value	0,2	30,3	12,8	27,5	21,2	27,7	29,0
BCR**	0	1,0	1,0	1,0	1,7	1,1	1,4

* Maintenance and upkeep

** Benefit cost ratio (Net present value per NOK of National Road Authority budget outlays)

Table 2: E39 Aksdal-Bergen. Cost c (NOK billion in 2012 prices), Benefit cost ratio h and the indicator w(k) for the main alternatives at different values of k, the benefit cost ratio of marginal projects in the plan.

	K2	K3	K4A	K4C	K4D	K5A	K5B
c	5,3	32,5	13,8	29,1	13,6	28,0	23,4
h (=NNB)	0	1,0	1,0	1,0	1,7	1,1	1,7
w(0)	0	33	14	29	23	31	33
w(0,25)	-1	24	10	22	20	24	27
w(0,5)	-3	16	7	15	16	17	21
w(0,75)	-4	8	3	7	13	10	15
w(1)	-5	0	0	0	10	3	9
w(1,5)	-8	-16	-7	-15	3	-11	-2
w(2)	-11	-33	-14	-29	-4	-25	-14

We note in Table 2 that if the budget does not require projects to be more than just socially efficient (k = 0), there is a tie between K3 and K5B. If we strengthen our reqirement one notch, our rule will pick K5B alone as the best alternative solution. As k approaches 1, K4D – the alternative with the highest benefit cost ratio – will take over the lead. This is all as expected.

Actually, the alternative that was chosen was K4C. As we can see from Table 1, it has neither the highest net present value nor the highest benefit cost ratio. Alternative K3 was however eliminated because of unacceptable non-monetarised effects. This having been done, K4C emerged as one of several alternatives with about the same net present value. It was also this alternative that answered best to the purpose of the project, which was to build a fast connection between Stavanger and Bergen, the two major cities on the west coast. If this is the purpose of the project, there is of course nothing to prevent it from becoming the the decisive factor in the end. It will always be necessary to use judgement to supplement formal methods. But the reasons for the final choice should of course always be stated clearly.

K4D is identical to K4C except that it retains the ferry crossing. Thus it can function as a first stage in the construction of K4C, postponing the bridge until traffic levels have

grown sufficiently. We see that as the budget gets tighter and only extremely profitable projects can be realised, it is this first stage that becomes the best alternative. If we only compare K4C and K4D, K4C should be chosen for $0 \le k < 0,5$, while K4D takes over when $k \ge 0,5$. Thus, the tightness of the budget and the amount of profitable projects elsewhere have a bearing on the question of whether we should opt for a simple and cheap or a more expensive but better solution. This perspective has not been used explicitly on the choice of alternative in any project up until now, as far as I know.

At *k* near zero, our criterion becomes similar to the net present value criterion, while as *k* increases, it becomes more like the benefit cost ratio, Thus the global setting into which the project competes for funding, matters for the criterion to be used locally.

5 Remarks

5.1 Cases with mutually exclusive alternatives

We assumed that the task at hand was to select projects to the national transport plan. But there are many similar situations. In the initial exploratory planning stage, for instance, there are always competing designs of the project, and very often, it is clear that there exists at least an expectation that the amount of funds to be spent on transport investments is kept within certain limits. An urban transport plan is a case in point. In urban areas, there are also often interdependencies between projects either on the demand side or in construction. If very many projects depend on each other, one should formulate and solve a network design problem. If however the interdependencies consist of a number of small groups of interdependent projects, another possibility is to construct all possible combinations of the projects within a group. These form mutually excluding alternatives that should compete with the single projects and the combinations of other groups as outlined in (LP3).

5.2 Several constraints

In addition to the budget constraint, there may be targets in the form of constraints on for instance CO_2 emissions, the number of accidents etc. In that case, neither the BCR nor the indicator of section 3 is of any use. If there are just two constraints, we could probably add them both to the objective function, each with its own Lagrangian parameter. But the search procedure would be more difficult. Anyhow, a linear programming problem can always be formulated and solved with any commercial software for LP problems.

5.3 Who is responsible for the selection?

It is worth noting that unless we can propose a correct value of k^* in advance, the procedure of section 3 implies that, except for the projects with only one alternative, the final selection of the right alternative design can no longer be taken locally, but must be transferred to a central authority.

6 Conclusion

If the projects are infinitely divisibe and independent and the given budget is the only binding restriction, and if none of the projects exist in mutually exclusive alternatives, maximum total

net present value of the plan is achieved by ranking them according to their benefit cost ratio and selecting them from the top until the budget is exhausted. If there are mutually exclusive alternatives, neither the net present value nor the cost benefit ratio is a valid selection criterion. Instead, a selection criterion (a special indicator) that depends on the benefit cost ratio of the last project that is included in the plan should be applied.

Initially, assume this marginal BCR to be given. Then for each project, the alternative with the highest score on the selection criterion should be chosen. When this has been done, compute the sum of costs of all projects with an indicator value above zero, or alternatively with a BCR above the assumed marginal BCR. Adjust the assumed marginal BCR up or down and repeat the computation until until the cost of all projects with a positive indicator value is just a little below the given budget. This procedure utilises the fact that, by construction, the Lagrangian multiplier of the budget constraint and the BCR of the last project to fit into the budget are the same.

References

Dovre Group and TØI (2012) E39 Aksdal-Bergen. Quality Assessment of the Choice of Concept (QA1). (In Norwegian)

HEATCO (2005a) Deliverable 1. Ciurrent practice in project appraisal in Europe. http://heatco.ier.uni-stuttgart.de/

HEATCO (2005b) Annex to Deliverable 1. Country reports. http://heatco.ier.uni-stuttgart.de/

HEATCO (2006) Deliverable 5. Proposal for harmonised guidelines. http://heatco.ier.uni-stuttgart.de/

Lorie, JH and LJ Savage (1955) Three problems in rationing capital. *Journal of Business* **28**(4), 229-129.

Norwegian Ministry of Finance (1979) Program Analysis. (In Norwegian.) Tanum-Norli, Oslo.

Weingartner, HM (1963) The Excess Present Value Index – A theoretical basis and critique. *Journal of Accounting Research* **1**(2), 213-224.

Weingartner, HM (1966) Capital Budgeting of Interrelated Projects: Survey and Synthesis. *Management Science* **12** (7), 485-516.

2.8 Industrial reorganisation benefits revisited

Industrial reorganisation benefits revisited[27]

Harald Minken
Institute of Transport Economics
E-mail: hm@toi.no

Contents

[27] The final publication is available in: Journal of Transport Economics and Policy (January 2014), Vol 48 (1), pp. 53-63, and may also be downloaded at https://www.ingentaconnect.com/content/lse/jtep/2014/00000048/00000001/art00004#, ISSN 0022-5258.

ABSTRACT

Mohring and Williamson (1969) showed how transport improvements can give rise to industrial reorganisation benefits. When a monopoly owns all firms and bears the transport costs, these benefits are all captured in a transport CBA. Unpublished work by Jansson and Wall shows that in the free market case, industrial reorganisation benefits are proportionally larger and can no longer be captured by a transport CBA. These results are reproduced here in a simpler mathematical framework.

We show that the free market has too many firms and too little transport and is economically inefficient. This is why a transport improvement in this case produces wider economic benefits. Preferably, we should eliminate the inefficiency rather than counteract it with transport infrastructure building. A uniform price of the commodity at all locations might do the trick.

2 November 2012

1.0 Introduction

Consider a geographical area with consumers spread uniformly all over the area, and inputs to production available at the same fixed price everywhere. If there were no returns to scale in production, the economically efficient industry structure would be to have very small plants serving their immediate neighbourhoods, thus saving on transport costs. Mohring and Williamson (1969) studied the case where consumers are identical, each demanding one unit of the produced good if price is below some threshold, and zero units if price is above that level. Production is characterised by increasing returns to scale and constant per kilometre transport costs. They showed that if a monopolist owns all plants and bears all transport costs, profit maximisation results in a pattern of equally spaced plants at some distance from each other. A transport improvement in this model will result in *industrial reorganisation*, i.e. fewer plants further apart from each other, and more transport. The key result in the Mohring and Williamson (MW) paper is that the benefits from this reorganisation are perfectly captured in a transport cost benefit analysis.

Unpublished work by Jan Owen Jansson and Rickard Wall (JW) extends the MW model by treating the case where each plant maximises its profit separately. Markets are completely free, without barriers to entry and with no cost associated with relocation. Consumers bear the transport costs. In this case, industrial reorganisation benefits are larger than in the monopoly case, and a considerable part of the total benefits to a transport improvement are not captured by a transport CBA. These interesting unpublished results are the inspiration for the present work, in which I reproduce the JW results in a simpler mathematical framework and provide some explanations for them.[28]

The reason for the MW results seems to be that the monopoly, bearing as it does all social costs and catering to a constant demand, actually realises the socially optimal outcome, even if it prices above marginal cost. On the other hand, the outcome of free competition is too many plants and too little transport. Bearing in mind that Jara-Diaz (1986) has shown that with elastic demand functions and marginal cost pricing, the benefits of a transport improvement will all be captured by a transport cost benefit analysis, while if prices deviate from marginal costs this is no longer true, we might venture the hypothesis that it is not the marginal cost pricing per se, but the fact that the economy is at the social optimum that makes it unnecessary to analyse other markets than the one immediately affected by the improvement. A policy that could move the economy to the social optimum would obviously reap the industrial reorganisation benefits and eliminate the need to implement transport improvements with a negative net benefit before industrial reorganisation benefits are added.

In the sequel, we illustrate the MW and JW results in a setting that makes for much simpler mathematics than theirs, and show the inefficiency of the free competition outcome (sections 2.0 and 3.0). We then ask in section 4.0 what policy measures could be used to move the free competition outcome towards the social outcome, and what policies should accompany a transport improvement. Section 5.0 relates our findings to the current discussion about wider economic benefits, while 6.0 concludes.

---

[28] I am grateful to Jansson and Wall for allowing me to do this.

2.0 The model

Basically, we apply a model originally due to Salop (1979) and outlined in section 7.1.2 of Tirole (1988). The model adopts the same assumptions about demand and the constant-per mile transport cost as the MW paper. The assumption about production is very simple: It is assumed that the marginal cost of producing a unit of the good (a widget) is zero, but for each plant there is a fixed cost element *f*.[29] We will treat two polar cases, case A and case B. *Case A* corresponds to the assumptions made by Jansson and Wall: Transport costs are borne by consumers, and the market structure is free competition, no entry costs, costless relocation. *Case B* corresponds to the assumptions of the Mohring and Williamson paper: transport costs are borne by a monopolist who owns all plants. Case A will be our base case, while case B will be used for comparison.

The only substantial difference between our model and the MW and JW framework is the structure of space itself: In our model, all activity takes place at the circumference of a circle! The circumference has a length of 1 mile and the demand has a density of 1 everywhere.

On the one mile ring, any numbers of identical firms are free to establish themselves anywhere and start production of widgets, as long as they each incur the fixed cost *f*. We assume that *f* is small enough for the whole circumference to be served. By symmetry, firms establish themselves at equal distance from each other (and by some miracle, they relocate costlessly to accommodate newcomers if *f* or the transport cost *t* is reduced). So if there are *n* firms, each serve a market area of length 1/*n*, extending 1/2*n* miles to the left and 1/2*n* miles to the right. And by the free entry assumption, all firms charge a price *p* that gives zero profit.

This model is different from monopolistic competition since in monopolistic competition, a price change by one of the firms has only a negligible effect on the sales of all others. On the ring, however, each firm has two neighbours whose price-setting will affect it very much and determine the length of its market area.

In case A, the utility of a customer located at distance *x* from the nearest plant is given by:

(1) $$U_x = \max\left(R, \overline{s} - p - tx + R\right)$$

where

x is the distance from the consumer to the plant

t is transport cost per widget mile (assumed to be constant)

p is the f.o.b. widget price

$\overline{s}$ is the consumer surplus associated with consuming a widget

R is income.

If trade occurs, Roy's identity gives

[29] In Tirole there is a marginal cost *c*, which I set to 0 because demand is constant, and so the variable cost of production is a constant as well.

$$-\frac{\frac{\partial U_x}{\partial p}}{\frac{\partial U_x}{\partial R}} = 1,$$

so the demand of the customer located at x is 1 unless $\bar{s} - p - tx < 0$, in which case it is 0.

Aggregate demand in this model is 1. The demand of each firm is 1/n. Aggregate transport in widget miles, D, is

(2) $$D = 2n\int_0^{\frac{1}{2n}} x\,dx = \frac{1}{4}\cdot\frac{1}{n}$$

Total social cost C, the sum of production costs and transport costs, is

(3) $$C = nf + tD = nf + \frac{t}{4n}$$

Now let us find expressions for aggregate user benefits UB and aggregate profits Π, assuming there is a common price p for all firms and that the number of firms is n. The user benefits are

(4) $$UB = 2n\int_0^{\frac{1}{2n}} \left(\bar{s} - p - tx + R\right) dx = \bar{s} - p + R - tD = \bar{s} - p + R - \frac{1}{4}\cdot\frac{t}{n}$$

Total profits are

$$\Pi = p - nf$$

Aggregate welfare W is therefore

(5) $$W = UB + \Pi = \bar{s} + R - \left(nf + \frac{1}{4}\cdot\frac{t}{n}\right) = \bar{s} + R - C$$

Since $\bar{s} + R$ is a constant, welfare maximisation is the same as minimisation of total social costs. We also note that the price p cancels out when UB and Π is added and does not occur in the expression of W. Therefore you cannot achieve welfare improvements by setting p to a particular level, and in particular, price equal to marginal cost means nothing for welfare here. This is an illustration of the fact that when demand is constant, welfare maximisation means minimisation of total social costs.

2.1 Social optimum

Cost minimisation with respect to the number of firms gives

(6) $C_{\min} = \sqrt{tf}$ which is achieved at $n = n_{C\min} = \frac{1}{2}\sqrt{\frac{t}{f}}$

Since demand is constant, (7) characterises the socially efficient outcome. This is now to be compared with the case A (free market) solution and the case B (monopoly) solution. In fact, there are four solutions, since the two market assumptions can each be combined with an

assumption about who bears the transport cost. But initially we assume that in case A, the customers bear the transport cost, but in case B, the firm does.[30]

2.2 Case A: Free competition

Assume that firm *i* sets a price p_i and faces a price p from its left and right competitors. The borders of the market area of firm *i* will be at the distance *y* where consumers are indifferent between the two neighbouring firms, or the *y* that solves

$$p_i + ty = p + t\left(\frac{1}{n} - y\right)$$

Firm *i* then faces a demand of

$$D_i(p_i, p) = 2y = \frac{p + t/n - p_i}{t}$$

and consequently wants to choose p_i to maximise profits

$$Max\ \Pi_i = p_i \frac{p + t/n - p_i}{t} - f$$

The solution to this problem is

$$p_i = \frac{p}{2} + \frac{t}{2n}$$

Now by symmetry, $p_i = p$ and therefore $p = t/n$ is the price set by all firms. If they make a profit at this price, new firms will enter until

$$\Pi_i = p\frac{1}{n} - f = \frac{t}{n^2} - f = 0 \text{ for all firms } i.$$

This fixes the number of firms at n_A:

(7) $$n_A = \sqrt{\frac{t}{f}} \text{ and the price at } p_A = \sqrt{tf}$$

Comparing (8) with (7), we see that the free market has exactly twice as many firms as the socially efficient solution, so the free market is socially inefficient. Inserting (8) in (3) gives a social cost of 25% above the efficient solution, or

(8) $$C_A = \frac{5}{4}C_{\min} = \frac{5}{4}\sqrt{tf}$$

The inefficiency of the free market solution is due to a congestion externality. Space, whether circular, plane or abstract "product space", is a congestible resource, which is why economies with a spatial dimension will not always follow laws of non-spatial economics. The free

[30] The case of monopoly in production combined with customers paying the transport costs is briefly treated in footnote 4, while the free market case where producers pay the transport costs is treated in section 4.0 and in the appendix.

market solution in our model is the equilibrium solution of a non-cooperative game, in much the same way as user equilibrium on a congested network is. It is not to be expected that the outcome will be socially optimal without some form of regulation.

2.3 Case B: Monopoly

For case B, we will have to reformulate the utility function of consumers and change the UB and Π of (4) and (5) by moving the transport costs from UB to Π. In this case,

(9) $$U_x = \max\left(R, \bar{s} - p + R\right)$$

and so the demand of the customer 1 regardless of where he is located, unless $\bar{s} - p < 0$, in which case it is 0.

It is not a foregone conclusion that the monopolist wants to serve the entire market, and so we allow him to choose the range y on both sides of a plant that he will serve. Taking into account the amount of transport required (2), his profit maximisation problem is

(10) $$\max_{p,n.y} \Pi = 2n\int_0^y \left(p - tx\right)dx - nf \quad \text{s.t. } p \le \bar{s} \text{ and } y \le \frac{1}{2n}$$

As shown in the appendix, provided $\bar{s} \ge \sqrt{tf}$, the solution to this problem is

$$y = \sqrt{\frac{f}{t}}, \quad n = \frac{1}{2}\sqrt{\frac{t}{f}}, \quad p = \bar{s}$$

(11) $$\Pi = \bar{s} - \sqrt{tf}$$

$$W = U_x + \Pi = R + \left(\bar{s} - \sqrt{tf}\right)$$

This is precisely the social optimum solution. It is seen that $y = 1/2n$, and so the monopolist serves the entire market. Price is obviously above marginal cost, and even above the free market price in the case where the producers do not have to charge for transport. Consumers get *R* regardless of whether they buy or not. They are robbed of the whole consumer surplus. The monopolist gets the consumer surplus and bears the total social costs, which he keeps at the minimum level, thereby realising the social optimum. Since in this case the monopolist has captured all benefits and bears all costs, no wonder he chooses the socially optimal solution.[31]

3.0 Transport benefits and industrial reorganisation benefits

We now confirm that in the case B of this circular market, corresponding as it does to the assumptions of the original Mohring and Williamson paper, the benefits of a reduction in

[31] Assuming the customers bear the transport cost in the monopoly case gives a result in-between the two other results. The monopolist maximises the Π of (5) subject to $\bar{s}$ – p – t/2n ≥ 0 to get $n = \sqrt{t/2f}$ and a *C* 6% above the socially efficient *C*.

transport cost *t* can indeed be measured in transport. Inserting the n_{Cmin} of (7) or of (12) in (2), we get transport demand in this case:

(12) $$D_B = \frac{1}{2}\cdot\sqrt{\frac{f}{t}}$$

Doing a CBA in transport of a change in *t*:

(13) $$\Delta TUB_B = -\int_t^{t+\Delta t}\frac{1}{2}\cdot\sqrt{\frac{f}{y}}\cdot dy = \sqrt{tf} - \sqrt{(t+\Delta t)\,f} = -\Delta C_{\min}$$

Thus it is confirmed that all benefits are captured by a CBA in transport in case B.

From (8), the corresponding transport demand and transport CBA in case A are:

(14) $$D_A = \frac{1}{4}\cdot\sqrt{\frac{f}{t}}$$

(15) $$\Delta TUB_A = -\int_t^{t+\Delta t}\frac{1}{4}\cdot\sqrt{\frac{f}{y}}\cdot dy = \frac{1}{2}\left(\sqrt{tf} - \sqrt{(t+\Delta t)\,f}\right) = -\frac{1}{2}\Delta C_{\min} = -\frac{2}{5}\Delta C_A$$

where we have used (9) at the last equality sign in (16). Obviously, 3/5 or more than half the benefit is *not* captured by a CBA in transport in case A, confirming one of the main results of Jansson and Wall. Thus the social inefficiency case is also the case where benefits cannot be captured on the road in our model.

Finally, we define Consumer Surplus to Existing Traffic, CSET and industrial reorganisation benefits, IROB:

(16) $$CSET_B = -\Delta t\cdot D_B \quad \text{and} \quad CSET_A = -\Delta t\cdot D_A$$

(17) $$IROB_B = -\Delta C_{\min} - CSET_B \quad \text{and} \quad IROB_A = -\Delta C_A - CSET_A$$

From (13) and (15), transport in case A is only half the level of transport in case B. From (9), cost reductions in case A will be 5/4 the cost reductions in case B. Using these relationships, it is easy to verify that

(18) $$\frac{IROB_A}{CSET_A} = \frac{2}{3}\cdot\frac{\Delta C_{\min}}{\Delta t\cdot D_B} + \frac{IROB_B}{CSET_B}$$

Thus the ratio of industrial reorganisation benefits to benefits to existing traffic will be considerably larger in the free market case than in the monopoly case in our model. This accords with the other main finding of Jansson and Wall.

4.0 Regulating the free market

In the case A situation, the question of what can be done to move the solution towards the social optimum arises. Consider these five instruments:

An upper limit on entry. Since new firms will enter as long as there are profits to be reaped, the upper limit will also be the number of firms that enter, provided the regulation is tight enough. By setting $n = \frac{1}{2}\sqrt{t/f}$, this instrument brings about the optimum at no cost. But in

practical cases the right number of firms to allow will be difficult to assess, and the restriction will be impossible to enforce.

Transport investment. By making transport cheaper this policy leads to fewer firms. This allows for further exploitation of economies of scale in production. In many ways, this is what has happened in world production and trade in the last 50 years. But in our model at least, this policy does not remove the difference between the free market number of firms and the optimal number of firms. The former will still be twice the latter. Also, the policy is costly to the government.

A property tax on firms (or any tax on firms that is independent of the level of production). This leads to fewer firms and brings in money to the government. The money could be spent on transport infrastructure. Thus a combination of a property tax and transport investment is likely to achieve something close to the social optimum.

A transport tax, charged to consumers in proportion to their transport cost. This could help finance investments but leads to more firms entering and will never help bridging the gap between the free market solution and the social optimum.

Regulation to the effect that firms will bear transport costs. Under the conditions specified in the appendix, it is shown there that this could lead to the socially optimal solution. The main requirements are that firms are forced to charge the same price everywhere in their market areas, and that there is fierce price competition at the edges of each firms' market area.

Since a social optimum is achieved, all benefits may be measured in the transport market, and there are no wider economic benefits. It is quite remarkable that wider economic benefits may be wiped out (or, more realistically, drastically reduced) by such simple forms of regulation.

5.0 Discussion

Wider economic benefits are benefits from a transport policy that is not captured by a cost benefit analysis in the transport sector alone. Thus what we show in this article is that economies of scale in production may give rise to wider economic benefits in the free market case, even if in the monopoly case studied by Mohring and Williamson, it does not. We also show that there are ways of achieving the social optimum even under free market conditions, thus eliminating the need to compute benefits outside transport.

As some of these results can also be derived from models with a more realistic representation of space, possibly all results in the article are independent of the particular representation of space here. But apart from space, the model is also simple in other respects: The produced good is homogeneous, the factories are identical and equally spaced, the customers are identical and equally spread out in space, and the costs of production and transport are very simplified. It may nevertheless throw some light upon basic aspects of the relationship between increasing returns to scale in production and the conditions where wider economic benefits exist.

Important as it is for the geographical organisation of production, economies of scale are of course not the only form of market imperfection that may give rise to wider economic benefits – see for instance Venables (2007). Nothing is said of these in this article, but whether the results here carry over to other forms of market imperfection is an interesting question for the future.

Our results should in no way be taken to imply that we can forget about wider economic benefits in appraisal. We live in a second best world, and even if in some cases we might be able to design first-best policies, it will often not be possible to put them into practice. In that case, we need to include the wider benefits. But the search for a first-best policy might nevertheless have a role to play in appraisal.

6.0 Conclusion

Mohring and Williamson (1969) showed how transport improvements can give rise to industrial reorganisation benefits. In the case where a monopoly owns all firms and bears the transport costs, they showed that these benefits are captured in a transport CBA. As originally shown in unpublished work by Jansson and Wall, industrial reorganisation benefits are proportionally larger in the free market case, and can no longer be captured by a transport CBA. These results are reproduced here in a simple framework where production and consumption takes place on a circle. We show that the free market has too many firms and too little transport and is economically inefficient. As marginal cost pricing is irrelevant in this model, we venture the hypothesis that it is not marginal cost pricing per se but social optimality that allows all benefits to be captured in transport.

Suitable policies can move the free market to a social optimum. Ingenious regulation of the number of firms might do the trick, but is difficult to achieve. Transport investment financed by property taxation is likely to work well. A uniform price of the commodity, regardless of the location of the customer, will bring about the social optimum if there is strong enough price competition at the edge of the market areas. Thus, the model indicates that before adding wider economic benefits to the transport CBA, one should consider the nature of the inefficiencies in the economy and whether there exist simple pricing and regulatory measures that may eliminate them.

References

Jansson, J.-O. And R.E. Wall (2002) A new model for identifying and measuring re-organization benefits of improvements in transport infrastructure. Prepared for presentation at the 6th Workshop of the Nordic Research Network on Modelling Transport, Land-Use and the Environment, Haugesund, Norway, September 27-29 2002.

Jara-Diaz, S. (1986) On the Relation between User Benefits and the Economic Effects of Transportation Activities. *Journal of Regional Science* **26**(2), 379-391.

Mohring, H. and H.F. Williamson Jr. (1969) Scale and "industrial reorganisation" economies of transport improvements. *Journal of transport Economics and Policy* **3**(3), 251- 271.

Salop, S. (1979) Monopolistic competition with outside goods. *Bell Journal of Economics* **10**, 141-156. (Cited from Tirole (1988)).

Tirole, J. (1988) *The Theory of Industrial Organization*. MIT Press, Cambridge, Mass.

Venables, A.J. (2007) Evaluating Urban Transport Improvements: Cost Benefit Analysis in the Presence of Agglomeration and Income Taxation. *Journal of Transport Economics and Policy* **41**(2), 173-188.

APPENDIX

The monopolist's problem

Solving the integral in (11) and rearranging, the monopolist's problem in case B can be stated as

$$Max_{p,n,y}\Pi = 2pny - tny^2 - fn \quad s.t. \quad p \le \overline{s} \quad (\mu_1)$$
$$2ny - 1 \le 0 \quad (\mu_2)$$

The Kuhn-Tucker conditions are:

$$\frac{\partial L}{\partial n} = 2py - ty^2 - f - 2\mu_2 y \le 0 \quad (= 0 \text{ for } n > 0)$$
$$\frac{\partial L}{\partial p} = 2ny - \mu_1 \le 0 \quad (= 0 \text{ for } p > 0)$$
$$\frac{\partial L}{\partial y} = 2n\left(p - ty\right) - 2\mu_2 n \le 0 \quad (= 0 \text{ for } y > 0)$$
$$\mu_1 \ge 0 \quad (= 0 \text{ for } p < \overline{s})$$
$$\mu_2 \ge 0 \quad (= 0 \text{ for } y < 1/2n)$$

Observe first that $n = 0$ would give $\Pi = 0$ with p and y being irrelevant or immaterial. If there is any better solution than this, it must follow from $n > 0$. This makes the first K-T condition an equality, and since $f > 0$, it follows that $y > 0$ also. Thus the third condition is also an equality. From these two equalities we can compute the solution for y and an expression for μ_2:

$$y = \sqrt{\frac{f}{t}}, \quad \mu_2 = p - \sqrt{tf}$$

Since $\mu_2 \ge 0$, this expression implies p > 0, and thus equality in the second K-T condition. Now it can be seen that if $\mu_1 = 0$, the second K-T condition produces a contradiction, and so $p = \overline{s}$. Inserting this solution into our expression for μ_2, we notice that if $\overline{s} < \sqrt{tf}$, the last K-T condition is contradicted. Thus the trivial solution $n = 0$ is the only solution unless the parameters of the problem are such that $\overline{s} \ge \sqrt{tf}$. Assume this to be the case.

Now if y < 1/2n, $\mu_2 = 0$ which is only the case if by chance $\overline{s} = \sqrt{tf}$. Except for this coincidence, which we ignore here, the solution is either not to produce at all or to serve the entire market. In the latter case, $y = 1/2n$, and this gives us the entire non-trivial solution candidate solution:

$$y = \sqrt{\frac{f}{t}}, \quad n = \frac{1}{2}\sqrt{\frac{t}{f}}, \quad p = \overline{s}$$
$$\Pi = \overline{s} - \sqrt{tf} = W - R$$

Provided $\overline{s} \ge \sqrt{tf}$, this is the best solution candidate. Consumers get R regardless of whether they buy or not. They are robbed of the whole consumer surplus. The monopolist gets the

consumer surplus and bears the total social costs, which he keeps at the minimum level, thereby realising the social optimum.

The free market case when a uniform price is charged

If the firms bear the transport costs and somehow charge a unified price regardless of the location of the customer, and if firm *i*'s market area extends *y* miles in each direction, the profit function is

$$\Pi_i = 2p_i y - 2t\int_0^{1/2n} x\,dx - f = 2p_i y - ty^2 - f$$

We might assume that the entire market is served. Since all firm are identical, it means that each of them have a market slice of $1/n$. We also assume that all customers will buy, so $p_i \leq \overline{s}$ for all *i*. The profit of firm *i* can then be written

$$\Pi_i = \frac{p_i}{n} - t\left(\frac{1}{2n}\right)^2 - f$$

At the edge of the market area, the firm cannot charge a price that gives a positive profit, for if it does, it will be undercut by the neighboring firm with exactly the same distance to its production site. Consequently,

$$\frac{\partial \Pi_i}{\partial y} = 2p_i - 2ty = 0$$

This determines the price p_i as $p_i = ty$ or, substituting for *y*, $p_i = \frac{1}{2}\frac{t}{n}$.

Inserting this price into the profit function we find that

$$\Pi_i = \frac{1}{4}\frac{t}{n^2} - f$$

Free entry means that firms will enter the marked as long as there is a positive profit to be reaped. In equilibrium, all firms earn zero profit. Thus

$$\Pi_i = \frac{1}{4}\frac{t}{n^2} - f = 0, \text{ or } n = \frac{1}{2}\sqrt{\frac{t}{f}}$$

Thus the number of firms equals the socially optimal number.

All firms will charge the same price, so we set $p_i = p$ for all *i*. Inserting the equilibrium number of firms into the price expression $p = \frac{1}{2}tn^{-1}$, we find

$$p = \sqrt{tf} \quad \text{(as long as } \overline{s} \geq \sqrt{tf}\text{)}$$

It may now easily be verified that profits are zero, total costs are $\sqrt{tf}$, and user benefits are $UB = W = \overline{s} + R - \sqrt{tf}$. Thus everything is just as in the socially optimal solution.

2.9 A theory of freight values of time and reliability

A Theory of Freight Values of Time and Reliability[32]

Harald Minken
Institute of Transport Economics

Contents

[32] This unpublished paper has some overlap with the published article «A logistics cost function with explicit transport costs» by Harald Minken and Bjørn Gjerde Johansen, in Economics of Transportation,Volume 19, September 2019, 100116. The model is basically the same, but they apply it to different problems. See however acknowledgments at the end of the paper.

ABSTRACT

A logistics cost model including the different elements of transport cost is constructed, and freight values of time and reliability are derived by the envelope theorem. The value of time is the change in the optimal logistics cost per trip resulting from a marginal change in expected transport time, with the number of trips held constant at the initial level. Likewise, the value of reliability is the cost change resulting from a marginal change in transport time variance. It reflects the change in costs of stock-outs during lead time and/or the cost of holding safety stock to guard against it. It is found that the value of time includes a reliability term, as expected transport time matters for stock-outs and the level of safety stock. Both values can be assessed with data on freight flows at the level of the firm.

Keywords: Freight, appraisal, value-of-time, reliability

1 Introduction

Improving the speed and reliability of freight transport is considered a major transport policy objective in most countries, but the economic appraisal of reliability improvements in particular is not very well advanced. There are problems not just with assessing the value to firms and society of a certain reliability improvement, but also with assessing the size of the improvement itself.

Current state-of-the-art practice is to use freight values of time (VOT) and reliability (VOR) derived by stated preference methods (de Jong 2000). This approach was criticised by Bruzelius (2001). As an alternative, this paper derives VOT and VOR from a logistics cost minimisation problem. To go with it, a model to assess the size of the reliability improvement itself must be devised. This is left to future work.

The main idea is that firms will already have made provisions to insure against the consequences of delays and unexpected incidents in transport, namely in the form of a safety stock designed to maintain a set level of service. A policy measure that improves expected transport time and reduces uncertainty in transport will make it possible for firms to reduce their safety stocks at the same level of service or to improve the level of service at the same level of safety stock. This is how such measures may improve economic efficiency. In addition, there will be transport cost savings as vehicles can carry out more jobs per day, and as reliability improvements may reduce the planned slack between jobs.

In this perspective, it is less relevant to consider a single shipment and ask the firm about its willingness to pay for a time saving or an increase in the probability of timely delivery. At least, one has to make it clear if the improvement affects all shipments or if it is an unexpected one-off chance. If all shipments are affected, the relevant question is how the firm will adapt its inventory and service level policy and what it expects to gain from this. Or we can try to compute the gain ourselves, which is what we will do.

Since the derivation will be somewhat lengthy, we start in section 2 by providing an overview of the approach. In section 3, we identify the elements of logistics cost and formulate the logistics cost minimisation problem. The solution is left to the appendix. Section 4 derives and discusses the values of time and reliability, while section 5 concludes and points to future work.

2 Overview

We consider a simple but not untypical case where goods are supplied regularly from a source A to an outlet B. Transport may be door-to-door by truck or consist of distribution stages by truck and a line haul by any mode. The simple text book logistics cost minimisation problem in this case minimises the sum of inventory costs and ordering costs. Transport costs are considered to be per tonne and thus irrelevant for the problem as long as demand in tonnes per year is constant. The result of cost minimisation is a certain shipment size (the "economic order quantity") and a certain average inventory.

However, a closer look at transport costs reveals that not all of them accrue per tonne. The per tonne transport costs are loading and unloading at terminals plus the cost of the line haul (assuming the line haul carrier offers a freight tariff by the tonne). These costs may be left out of the cost minimisation problem. The picture is different for the stages where shipment size may affect the choice of vehicle size (vehicle capacity), i.e. distribution and door-to-door. Suppose there are upper and lower bounds on vehicle size. Then the cost of using the smallest available vehicle will be a cost per shipment, or a part of the ordering cost, and highly relevant for the cost minimisation problem. The cost of choosing a larger vehicle size will also be relevant, with the implication that vehicle size will have to be included as a choice variable in the cost minimisation problem. This is shown in section 3. With the relevant transport costs included, the firm chooses shipment size and vehicle size to minimise logistics cost subject to the constraints that shipment size cannot exceed the maximal vehicle size, vehicle size must be within a feasible range, and annual transport capacity must at least cover annual demand.

Now attention must be paid to uncertainty. We assume that lead time (including transport time as a part) and demand per hour (day, week) are both stochastic variables. This implies that demand during lead time is also a stochastic variable. It may be shown (Hadley and Within 1963, exercise 3.12) that stochastic lead time demand has the following expectation and variance:

(1) $$\mu_L = \mu_D \mu_T$$

(2) $$\sigma_L^{\ 2} = \mu_T \sigma_D^2 + \mu_D^2 \sigma_T^2$$

where μ_D and σ_D are expectation and variance, respectively, of demand per hour, μ_T and σ_T are the expectation and variance of lead time, and μ_L and σ_L are expectation and variance of lead time demand. Other things equal, the longer the transport time, the longer the expected lead time. Thus equation (2) shows that the longer the transport time, the more does the variance of demand contribute to lead time demand variance. Also, the mean level of demand will determine the impact of transport time variance. To the extent, then, that uncertain lead time demand influences the firm's optimal inventory and optimal stock out costs, a policy measure that reduces transport time variance will affect firms with different demand characteristics and transport distances very differently. If we assume normally distributed lead time demand, (1) and (2) is all we need to know about the distribution.

With stochastic lead time demand, the firm will need another policy variable, which we take to be the reorder point *R*, i.e., the inventory position that triggers the placement of a new order. A small *R* makes for a high probability of stock-out during lead time (and higher the higher σ_L is), while a large *R* makes for high average inventory costs. The difference between *R* and the expected lead time demand is the safety stock.

To formalise all of this we add two new elements to the logistics cost to be minimised, namely the cost of holding the safety stock and the annual expected cost of stock-outs, both largely determined by *R*. In this as in everything else except the transport costs, we follow standard inventory theory, as in Hadley and Whitin (1963). The firm's problem is to minimise all relevant logistics costs with respect to shipment size, vehicle size and reorder point, subject to the same constraints as in the deterministic case.

Obviously, the optimal safety stock and stock-out costs increase with σ_L. Consequently, a policy measure that reduces transport time expectation and variance will reduce these elements of cost and thereby increase economic efficiency. The effect depends on the transport distance and demand characteristics of the particular firm.

The solution to the logistics cost minimisation problem with transport costs, choice of vehicle size and uncertain lead time demand is given in the appendix. The solution – the logistics cost function – may potentially be useful to explain localisation, the choice of suppliers, the possibility of adopting Just-in-time (JIT) policies etc. It may be seen that economies of scale depend on the upper and lower bounds on vehicle size. In section 4, however, the formulation of the problem is used to derive VOT and VOR for cost benefit analysis.

Suppose lead time consists of transport time t_i on n transport stages plus a non-transport part T_0. All of the elements of lead time are independently distributed stochastic variables. Thus $\mu_T = \sum_i Et_i + ET_0$ and $\sigma_T^2 = \sum_i \operatorname{var} t_i + \operatorname{var} T_0$. Consider any marginal policy measure on one of the transport stages. It may be a shortening of the distance, a speed increase or whatever influences $\operatorname{var} t_i$. How will it affect the expected average annual logistics cost of our firm? To find out, we differentiate the logistics cost function with respect to distance a_i, mean transport time Et_i and variance $\operatorname{var} t_i$, recognising that a_i and Et_i appears as parameters directly in the logistics cost function while Et_i also appears in μ_T and hence μ_L and σ_L (equations (1) and (2)). Transport time variance on the stage i, $\operatorname{var} t_i$, only appears in the logistic cost function because it appears in σ_T and hence σ_L, which is an important determining factor of safety stocks and stock-out costs.

Hence if K^* is the expected average annual logistics cost, we compute the differential

$$dK^* = \frac{\partial K^*}{\partial a_i} da_i + \frac{\partial K^*}{\partial Et_i} dEt_i + \frac{\partial K^*}{\partial \operatorname{var} t_i} d \operatorname{var} t_i \tag{3}$$

We divide the result by the initial mean number of trips per year to find the cost change per trip, and reach the following conclusion:

1. The value of a saved kilometre is the kilometre dependent transport cost for the optimal (distribution) vehicle, or, for the main haul, the kilometre dependent transport cost times the shipment's proportion of the total load.

2. The value of a saved hour (VOT) is the cost of hiring and manning the optimal (distribution) vehicle for one hour (or, for the main haul, the time dependent transport cost times the shipment's proportion of the total load), plus the hourly depreciation of the shipment and the cost of capital tied up in the shipment, plus the cost of increased uncertainty associated with one more hour in transport. This third term of VOT is $VOR \cdot \left(\sigma_D / \mu_D\right)^2$, with VOR defined below.

3. Denote the density function of the standardised normal distribution by $f(.)$ and the cumulative probability function of the same by $F(.)$, the price of the good by p, the

holding cost per hour by *h*, the fixed per case cost of stock-out by π and the stock-out cost per hour by $\tilde{\pi}$. The value of a marginal decrease in the variance of transport time (VOR) is

$$VOR = \frac{1}{2}\left[\pi \frac{\mu_D}{\sigma_L} f\left(\frac{R^* - \mu_L}{\sigma_L}\right) + \left(ph + \tilde{\pi}\right)\left(1 - F\left(\frac{R^* - \mu_L}{\sigma_L}\right)\right)\right]\mu_D \quad (4)$$

VOR expresses the increase in stock-out costs when the firm has adapted optimally in the first place but experiences afterwards a marginal reduction in the variance of transport time. As it stands, (4) concerns the distribution stages of transport or door-to-door by truck, but if price equals marginal cost on the line haul and the line haul frequency is no cause of concern to the customers, the formula will also apply to line haul stages.

As a rule, a policy measure that reduces distance will also reduce mean transport time, and a measure that reduces mean transport time will also reduce transport time variance. Thus very often we will need to use all three types of unit values to assess the impact, and we need to complement our unit values by a theory of da_i, dEt_i and $dvart_i$ and how they relate to each other – a theory of delays.

It may seem that (4) is useless because the fixed per case cost of stock-out π and the stock-out cost per hour $\tilde{\pi}$ cannot be known with certainty. However, if either one of π and $\tilde{\pi}$ is assumed to be zero, the other one may be expressed in terms of commonly used service level indicators. Let S_1 be the expected number of delivery periods per year when stock-out does not occurs, and S_2 be the proportion of sales that can be delivered from shelf. Setting $\pi = 0$ we get

$$VOR = \frac{1}{2} ph\mu_D \frac{(1 - S_1)}{(1 - S_2)} \quad (5)$$

Here, VOR is proportional to $ph\mu_D$, the cost of holding expected hourly demand for one hour, and a factor involving the service level indicators. Section 4 rounds off by discussing the collection of data to assess this formula and derive average values of VOT for use in cost-benefit analysis. As with a stated preference approach to VOT and VOR, a survey of firms and their shipments is needed. The difference may be that while very much of the variation in SP analyses is considered unobservable, we do claim some knowledge of its causes.

3 The elements of logistics cost

Consider a supply chain were a commodity is supplied from a single source and sold at a single outlet. For each shipment, transport time will be uncertain, and there will also be uncertainty about demand in shorter periods of time. However, if demand is generated by a stationary stochastic process, annual demand might be so much more certain that we might regard it as given. The number of shipments per year will nevertheless vary from year to year, since as a rule, the shipments will not fall on the same date each year. Over a number of years there is an average expected annual cost of the operation, and we assume the objective of the firm in charge (the shipper) is to minimise it. The average expected annual cost consists of the transport cost, non-transport ordering cost, the inventory holding cost – including the cost of holding a safety stock to guard against uncertain demand during lead time – and the stock-out costs. We intend to measure costs from the point of view of society, but for the moment it

may be assumed that this coincides with the point of view of the shipper. Unreliable transport gives rise to excessive safety stocks, high stock-out costs, and problems in meeting just-in-time requirements. It might also induce the carrier to guard against contingencies by including various forms of slack in his offer to the shipper, thereby increasing the cost per shipment somewhat. Apart from that, by the law of large numbers, the transport cost per shipment can still be based on expected transport time.

3.1 Transport costs

Transport is either door-to-door or combined transport. In the case of combined transport, we assume that the line haul freight tariff is a price P per unit of the commodity. We also assume that the frequency of the scheduled line haul transport service is sufficiently high not to constrain the choice of shipment size. With respect to the door-to-door transport or the distribution stages of the combined transport, we assume a perfect market. Besides zero profit to the carrier, we take this to mean that the firm in charge of the operation (the shipper) can always choose the most appropriate vehicle size C from a continuous interval $\left[C_{\min}, C_{\max}\right]$, and that outside the time when the vehicles are engaged in our supply chain, they will be fully employed in other engagements also yielding zero profit.[33] It also means that the carrier can be counted upon to keep turn-around time and other unproductive time to a minimum. The vehicles employed in the door-to-door transport or the distribution stages of the combined transport are not supposed to pick up or carry other goods on the same tour. Thus empty backhaul is assumed.

We define the following parameters:

x annual expected demand in units of the commodity

$\bar{t}$ expected one-way transport time in hours

u turn-around time and other unproductive time per round trip in hours

a one-way transport distance in kilometres

t_l combined expected loading and unloading time per unit in hours

k distance dependent costs per kilometre

i vehicle capital costs per hour

η number of business hours per year.

These parameters are all strictly positive and pertain to the distribution stages or door-to-door case. Where appropriate, subscripts on the parameters will be used to denote link or stage of the transport.

The choice variables, also pertaining to the distribution stages or door-to-door-case, will be selected among the following:

Q shipment size in units of the commodity

C vehicle size (carrying capacity) in units of the commodity

---

[33] If there is a cost of moving the vehicles to the starting point of the operation, this can be included in the non-transport ordering costs.

Y maximum annual transport capacity in units of the commodity

N the expected number of vehicles (or vehicle hours employed per hour)

F frequency, i.e., the number of trips per hour.

Two linear relations will be assumed. Apart from the other assumptions and the values of exogenous variables and parameters, they constitute the empirical content of the model. First, we assume a linear relationship between vehicle size C and kilometre cost k:

(6) $$k = k_0 + k_1 C$$

where k_0, k_1 are positive constants. Thus we ignore the impact of speed on kilometre costs. Second, the relation between vehicle size and hourly vehicle capital cost is

(7) $$i = i_0 + i_1 C$$

where i_0, i_1 are constants. For a given type of vehicles, both relations are thought to fit data well.

Expected round trip time will be $2\bar{t} + u + t_l Q$. Two identities, $Y = \eta F C$ and $N = \left(2\bar{t} + u + t_l Q\right) F$ apply. The second applies even if the number of vehicles will have to be a whole number, because unproductive time u will have to adjust. Under the perfect market assumption, however, N can fairly safely be assumed to be continuous, because vehicles can be engaged just for the time they are needed. Since transport time is uncertain, then compared to the certainty case, the number of vehicles will have to be increased to guard against adverse impacts of delays on the subsequent transport engagements. We assume this has been done by incorporating a slack in u. We may also assume that no unnecessary trips are made, so $\eta F = xQ^{-1}$. Making use of this to eliminate F, the identities are:

(8) $$Y = C\frac{x}{Q}$$

(9) $$\eta N = \left(2\bar{t} + u\right)\frac{x}{Q} + t_l x$$

Equation (8) shows that if Q is going to be a choice variable, the other choice variable could be Y or C, but there is no need for both. Likewise, (9) provides an opportunity to eliminate the number of vehicles as a choice variable. We settle for (Q,Y) as our choice variables. Both are strictly positive.

Let w be the cost per hour of crew, appropriately adjusted by the ratio of paid hours to hours "on the road". The hourly cost of engaging a vehicle will then be $w + i_0 + i_1 C$. Let the hourly cost of hiring loading and unloading capacity be w_l. The expected average annual cost of transport K_T consists of distance dependent costs, time dependent costs and the cost of hiring loading and unloading capacity for as long as this takes. In the door-to-door case this amounts to

$$K_T = \left(k_0 + k_1 C\right) 2a\frac{x}{Q} + \left(w + i_0 + i_1 C\right)\left(2\bar{t} + u + t_l Q\right)\frac{x}{Q} + w_l t_l x =$$

$$\left[2k_0 a + \left(w + i_0\right)\left(2\bar{t} + u\right)\right]\frac{x}{Q} + i_1 t_l Q Y + \left(w + i_0 + w_l\right) t_l x + \left[2k_1 a + i_1\left(2\bar{t} + u\right)\right] Y$$

where (9) has been used in the first line to eliminate *N* and (8) in the second line to substitute *Y* for *C*. It will never be economical to operate larger vehicles than necessary, thus we may assume $Y = x$ (and consequently $C = Q$) if $Q \in [C_{\min}, C_{\max}]$, and $Y = C_{\min}\, x/Q$ if $Q \le C_{\min}$.

For the case of combined transport, let stage 1 be the first distribution stage, stage 2 be the line haul and stage 3 be the second distribution stage. The costs of the second stage (including the cost of an extra loading and unloading) will be *Px*, which is to be added to the third term in the above expression. The shipment size and expected average annual number of shipments will be the same for both distribution stages, thus *C* and *Y* will be the same for both stages. The number of vehicles employed might however be different for the stages, as it depends on $\bar{t}$ and *u*. Index the relevant variables of each stage by their stage number, let stage 1 be the only stage in the door-to-door case, and let δ be 1 in the combined transport case, 0 otherwise. The expected average annual transport costs are:

$$
\begin{aligned}
K_T = {} & \left\{2k_0\left(a_1+\delta a_3\right)+\left(w+i_0\right)\left[2\left(\bar{t}_1+\delta\bar{t}_3\right)+\left(u_1+\delta u_3\right)\right]\right\}\frac{x}{Q} \\
& +\left\{i_1 t_l\right\}QY+\left\{\left(w+i_0+w_l\right)t_l+P\right\}x \\
& +\left\{2k_1\left(a_1+\delta a_3\right)+i_1\left[2\left(\bar{t}_1+\delta\bar{t}_3\right)+\left(u_1+\delta u_3\right)\right]\right\}Y
\end{aligned}
\qquad (10)
$$

Only the costs of the smallest possible vehicle accrue per shipment. The additional costs depend on annual demand and the vehicle size. Thus to assume a fixed transport cost per shipment is only right if the vehicle size cannot be chosen, and to assume a fixed cost per tonne is only right if the costs of the line haul outweigh by far the cost of the distribution stages.

3.2 Ordering costs and inventory holding costs, no uncertainty

We need the following prices, unit costs and rates:

b non-transport ordering costs (per shipment)

p free-on-board unit price of the commodity

H inventory holding cost per year and dollar of stationary inventory

J inventory holding cost per year and dollar of mobile inventory

ε equals $1-\frac{x}{z}$, where *z* is annual production rate at the source

The non-transport ordering costs *b* includes any fixed cost associated with producing the commodities to be shipped (set-up costs) or with providing the transport service, and any cost of paperwork associated with the shipment. From the point of view of society, these costs at the source are relevant, and surely, they will also matter to the private decision maker.

The expected annual cost of the stationary inventory will be $\frac{1}{2}pH\left(1+\varepsilon\right)Q$. This accounts for the inventory at the source as well as at the outlet. If production is instantaneous (import) and well coordinated with shipments from the source, there need be no inventory at the source, but if the production rate equals the demand rate, average inventories at both points are equally large. Thus ε takes values between 0 and 1. *H* includes the cost of capital, variable warehousing costs and time-dependent depreciation.

The mobile inventory will be proportional to annual demand and to average time in transport. The inventory holding cost per dollar tied up in transport is not the same as *H*, since on one

hand, there are no warehousing costs, and on the other, there might be a higher cost of depreciation and damage. It can be written as $pJ\eta^{-1}\left(\overline{t}+t_l Q\right)x$.

3.3 Safety stock and stock-out costs

We assume that items are demanded one at a time. Demand is generated by a stationary stochastic process, demand per business hour being normally distributed with mean μ_D ($\mu_D = x\eta^{-1}$) and standard deviation σ_D.

Lead time, with transport time as a part, is the time from an order is placed until it is available at the outlet. Lead time is stochastic, with delays in transport a part of this uncertainty. Mean lead time is μ_T and its standard deviation is σ_T. When demand and lead time are stochastic, demand during lead time is also stochastic. Since it is composed of demand during each of the separate elements of lead time, and arguably, these are independent stochastic variables, then by the central limit theorem we may assume demand during lead time to be normally distributed with mean μ_L and standard deviation σ_L. Under these assumptions, equation (1) and (2) apply.

Let the inventory be continually monitored. Stock-outs (not being able to deliver from shelf) are allowed to occur, but at a cost. We assume that stock-outs are backordered (delivered later), but that backorders have a cost per instance plus a cost depending on the time until delivery can take place. From society's point of view, this is the cost to the customer of having to wait, but for the firm, it mainly involves lost goodwill. The firm trades off stock-out costs and inventory holding costs by fixing a reorder point *R* such that whenever the inventory position (on hand inventory plus shipments under way minus backorders) reaches *R*, a new shipment is ordered. Adding *R* as a policy instrument, we say that our firm is following a (*Q,Y,R*) policy. *R* might take on any value, including negative, which means that we collect backorders until they are so many as to make a new shipment worthwhile.

To be able to treat the case of stochastic lead time properly, we must assume that orders arrive in the same sequence in which they were placed.

Let *I* be the average physical (on hand) inventory excluding the mobile inventory and the inventory at the source, *E* be the average number of backorders per year and *B* be the expected number of backorders at any point in time. These variables are functions of *Q* and *R*. *E* is of interest if there is a fixed cost per stock-out, while *B* is of interest if the duration of stock-outs matter. In chapter 4 of Hadley and Whitin (1963) it is proved that, with a minor and mostly inconsequential simplification,

$$
\begin{aligned}
E &= E\left(Q,R\right) = x\cdot Q^{-1}\alpha\left(R\right) \\
B &= B\left(Q,R\right) = Q^{-1}\beta\left(R\right) \\
I &= \frac{1}{2}Q + R - \mu_L + B\left(Q,R\right)
\end{aligned}
\qquad (11)
$$

where

$$\alpha(R)=\sigma_L\left[f\left(\frac{R-\mu_L}{\sigma_L}\right)-\frac{R-\mu_L}{\sigma_L}\left(1-F\left(\frac{R-\mu_L}{\sigma_L}\right)\right)\right]$$

(12)

$$\beta(R)=\frac{1}{2}\sigma_L^2\left[\left(1+\left(\frac{R-\mu_L}{\sigma_L}\right)^2\right)\left(1-F\left(\frac{R-\mu_L}{\sigma_L}\right)\right)-\frac{R-\mu_L}{\sigma_L}f\left(\frac{R-\mu_L}{\sigma_L}\right)\right]$$

In (12), $f(.)$ is the density function of a normalised normally distributed variable, and $F(.)$ is its cumulative density function. We will need the derivatives of $\alpha(R)$ and $\beta(R)$, and since $F'(x)=f(x)$ and $f'(x)=-xf(x)$,

(13)

$$\alpha'(R)=(-1)\left(1-F\left(\frac{R-\mu_L}{\sigma_L}\right)\right)$$

$$\beta'(R)=-\alpha(R)$$

Both functions are monotonously decreasing, and as both approach 0 as R goes to infinity, they are everywhere positive.

If we let π and $\hat{\pi}$ be the fixed unit cost per backorder and the cost per year of a backorder respectively, the stock-out costs will be $\pi E(Q,R)+\hat{\pi}B(Q,R)$. By (11), the average excess inventory compared to the deterministic case will be $R-\mu_L+B(Q,R)$. These two terms will have to be added to the deterministic logistics cost to take account of uncertain demand and lead time. We are now in a position to add the transport costs, the ordering costs, the inventory holding cost and the stock-out costs to form the total annual logistics cost K in our case. It is:

$$K=K_T+\left(b+\pi\alpha(R)\right)\frac{x}{Q}+\left(pH+\hat{\pi}\right)\beta(R)\frac{1}{Q}+pJ\eta^{-1}(1+\delta)t_lQx$$

$$+\frac{1}{2}pH(1+\varepsilon)Q+pJ\eta^{-1}\left(\overline{t_1}+\delta\left(\overline{t_2}+\overline{t_3}\right)\right)x+pH\left(R-\mu_L\right)$$

or

$$K=\left[\gamma_1+\psi(R)\right]\frac{x}{Q}+\gamma_2Qx+\gamma_3QY+\gamma_4Q+\gamma_5x+\gamma_6Y+pH\left(R-\mu_L\right)$$

(14)

$$\gamma_1=2k_0\left(a_1+\delta a_3\right)+\left(w+i_0\right)\left[2\left(\overline{t_1}+\delta\overline{t_3}\right)+\left(u_1+\delta u_3\right)\right]+b$$

$$\gamma_2=pJ\eta^{-1}(1+\delta)t_l$$

$$\gamma_3=i_1t_l$$

$$\gamma_4=\tfrac{1}{2}pH(1+\varepsilon)$$

$$\gamma_5=\left(w+i_0+w_l\right)t_l+P+pJ\eta^{-1}\left(\overline{t_1}+\delta\left(\overline{t_2}+\overline{t_3}\right)\right)$$

$$\gamma_6=2k_1\left(a_1+\delta a_3\right)+i_1\left[2\left(\overline{t_1}+\delta\overline{t_3}\right)+\left(u_1+\delta u_3\right)\right]$$

$$\psi(R)=\pi\alpha(R)+\left(pH+\hat{\pi}\right)x^{-1}\beta(R)$$

Except for transport costs and inventory holding costs during transport, *K* is a standard expression of average annual logistics cost in the tradition of Hadley and Whitin.

The decision maker's problem is to minimise logistics cost subject to $C_{\min} \le C \le C_{\max}$ and $Y \ge x$. Using (8), this may be written

$$
\begin{aligned}
& Max_{Q,Y,R} \; -K \\
\text{s.t.} \; & -QY \le -C_{\min} x \quad (\lambda_1) \\
& \quad QY \le \; C_{\max} x \quad (\lambda_2) \\
& \quad -Y \le -x \qquad (\lambda_3)
\end{aligned}
\tag{15}
$$

Q and *Y* are supposed to be strictly positive, but no restriction is placed on *R*. The solution to this problem is in the appendix.

4 Values of time and reliability

Let $\mathbf{a} = (a_1, a_2, a_3)$ be non-stochastic, and let $\mathbf{t} = (t_1, t_2, t_3)$ be stochastic transport time for the three transport stages, with mean $\bar{\mathbf{t}} = (\bar{t}_1, \bar{t}_2, \bar{t}_3)$ and variance $\mathbf{vart} = (\operatorname{var} t_1, \operatorname{var} t_2, \operatorname{var} t_3)$. We assume stochastic independence of the t_i. For one of the stages, say *i*, what is the influence on *K** of a marginal change in distance a_i, expected transport time $\bar{t}_i$, and the variance of transport time, $\operatorname{var} t_i$?

To answer that, we need to clarify the relationship between transport time and lead time. In the logistics cost minimisation problem, we treated lead time expectation μ_T and standard deviation σ_T as parameters entering μ_L and σ_L (equations 6 and 7). Thus the impact of *Q* on μ_L and σ_L through loading and unloading time was ignored. To do otherwise would have made the problem very much more complex. Sticking to that simplification, we define lead time *T* by

$$
T = \sum\nolimits_i t_i + T_0 \tag{16}
$$

where T_0 is all non-transport elements of lead time, including loading and unloading. We assume that the t_i and T_0 are independently distributed. Thus

$$
\begin{aligned}
\mu_T &= \sum\nolimits_i \bar{t}_i + ET_0 \\
\sigma_T^2 &= \sum\nolimits_i \operatorname{var} t_i + \operatorname{var} T_0
\end{aligned}
\tag{17}
$$

Finally, expected hourly demand is $\mu_D = x\eta^{-1}$. With these clarifications we return to the problem at hand. By the envelope theorem, the derivative of the logistics cost function with respect to some parameter θ is the derivative of the Lagrangian evaluated at the point (*Q**,*Y**,*R**). The parameters of interest here are a_i, $\bar{t}_i$, μ_L and σ_L. None of them occur in the constraints. We first compute the derivate of the cost function with respect to μ_L. By (12), (13) and (14) we get:

$$\frac{\partial K^*}{\partial \mu_L} = \left[\pi x \left(1 - F\left(\frac{R^* - \mu_L}{\sigma_L} \right) \right) + \left(pH + \hat{\pi} \right) \alpha \left(R^* \right) \right] \frac{1}{Q^*} - pH$$

But by (A4) of the appendix,

(18) $$\frac{\partial K^*}{\partial \mu_L} = 0$$

Having established this, the expressions for the total derivatives of the logistics cost function with respect to a_i, $\bar{t}_i$ and vart_i simplify to

(19) $$\frac{dK^*}{da_i} = \frac{\partial K^*}{\partial a_i}$$

(20) $$\frac{dK^*}{d\bar{t}_i} = \frac{\partial K^*}{\partial \bar{t}_i} + \frac{\partial K^*}{\partial \sigma_L} \frac{\partial \sigma_L}{\partial \mu_T} \frac{\partial \mu_T}{\partial \bar{t}_i}$$

(21) $$\frac{dK^*}{d \operatorname{var} t_i} = \frac{\partial K^*}{\partial \sigma_L} \cdot \frac{\partial \sigma_L}{\partial \sigma_T} \cdot \frac{\partial \sigma_T}{\partial \operatorname{var} t_i}$$

For i = 1,3, equation (19), the envelope theorem and (8) yields:

(22) $$\frac{dK^*}{da_i} = \left(k_0 + k_1 C^* \right) \frac{2x}{Q^*}$$

Likewise by the envelope theorem we have

(23) $$\frac{\partial K^*}{\partial \sigma_L} = \left[\pi x f \left(\frac{R^* - \mu_L}{\sigma_L} \right) + \left(pH + \hat{\pi} \right) \sigma_L \left(1 - F\left(\frac{R^* - \mu_L}{\sigma_L} \right) \right) \right] \frac{1}{Q^*}$$

It will turn out to be convenient to express (23) in terms of VOR as defined in (4). To do so, we convert the annual unit costs $pH + \hat{\pi}$ and J to hourly costs, so define $h = H\eta^{-1}$, $j = J\eta^{-1}$ and $\tilde{\pi} = \hat{\pi}\eta^{-1}$. Then (23) can be rewritten

(24) $$\frac{\partial K^*}{\partial \sigma_L} = \frac{1}{2} VOR \cdot \frac{\sigma_L}{\mu_D^2} \cdot \frac{2x}{Q^*}$$

Now by (24), the envelope theorem, and equations (2), (8) and (17), equations (20) and (21) may be rewritten as follows in the case of i = 1,3:

(25) $$\frac{dK^*}{d\bar{t}_i} = \left[\left(w + i_0 + i_1 C^* \right) + \frac{1}{2} pjQ^* + \frac{1}{2} VOR \cdot \left(\frac{\sigma_D}{\mu_D} \right)^2 \right] \cdot \frac{2x}{Q^*}$$

(26) $$\frac{dK^*}{d \operatorname{var} t_i} = \frac{1}{2} VOR \cdot \frac{2x}{Q^*}$$

Equations (22), (25) and (26) are our main result. It is seen that the marginal change in expected average annual logistics cost when anyone of the variables a_i, $\bar{t}_i$, or $\operatorname{var} t_i$ changes, is proportional to the expected number of trips in the initial situation, *2x/Q**. The result (22) is almost tautological – the marginal distance cost per trip equals the average distance cost as

given by (6). According to (25), the marginal time cost per trip consists of time dependent transport cost plus two terms expressing the marginal non-transport time costs. The coefficient ½ in these terms is due to the fact that only half of the single trips in our model carry a load. For the trips that carry a load, however, there is a cost *pjQ** per hour of holding the inventory on wheels, and a second term which has not been reported in the literature before. A similar term appears in (26).

Obviously, a reduction in the mean transport time reduces the probability of stock-outs during lead time and/or the cost of holding safety stock to guard against it, as does a reduction in transport time variability. According to (4) and (25), the value of such improvements to a particular firm is crucially dependent on characteristics of the commodity flow at the level of the firm, such as the price of the good, the magnitude and standard deviation of demand, and the structure and size of stock-out costs.

So far the results apply only to the distribution stages, or to the case of door-to-door transport by truck. Now consider an improvement of the main haul stage, i.e. a change in a_2, $\overline{t_2}$, and $\operatorname{var} t_2$. Again with a reference to the envelope theorem, it is seen by (14) that these changes impact the logistics cost of our firm in three ways: By possibly changing the per tonne freight rate *P*, by the change in the cost of holding inventory in transport, brought about by a change in the $\overline{t_2}$ of γ_5, and finally by the impact of $\overline{t_2}$ and $\operatorname{var} t_2$ on σ_L, which enters $\psi(R)$. (By (18), the impact on μ_L might be ignored.) The impact of $\overline{t_2}$ and $\operatorname{var} t_2$ on σ_L is in no way different from the other stages, and likewise, there is no difference between the stages in the way the $\overline{t_i}$ of γ_5 changes the cost of holding inventory in transport. We may there conclude that the only difference is that the distance dependent transport cost of (22) and the time dependent transport cost of (25) must be substituted by the change in the freight rate. Under the conditions that the main haul freight rate equals marginal cost, possibly with a fixed mark-up, that frequency is unchanged or of no concern to the customer, and that the carrier's gain from increased reliability is negligible, the freight rate change may be decomposed in a distance dependent part, consisting of our shipment's proportional share of the distance dependent transport cost, and a time dependent part, consisting of our shipment's proportional share of the time dependent transport cost. Thus equations (22), (25) and (26) might be applied to the line haul stage also.

In our model, a marginal change in the transport cost parameters brings about both a change in the cost per trip and the annual number of trips. The latter is due to the marginal change in the shipment size. In other settings, where the demand for trips is supposed to be fixed, there might be reason to modify the marginal costs of trips as given by (22), (25) and (26) to implicitly account for the effect on the number of trips. Define the cost per trip G as $G = K^*Q^*(2x)^{-1}$, i.e., the annual cost divided by the number of trips. Thus,

$$\frac{\partial G}{\partial a_i} = \left(\frac{\partial K^*}{\partial a_i}Q^* + K^*\frac{\partial Q^*}{\partial a_i}\right)(2x)^{-1},\ \frac{\partial G}{\partial \overline{t_i}} = \left(\frac{dK^*}{d\overline{t_i}}Q^* + K^*\frac{dQ^*}{d\overline{t_i}}\right)(2x)^{-1}$$

$$\frac{\partial G}{\partial \operatorname{var} t_i} = \left(\frac{\partial K^*}{\partial \operatorname{var} t_i}Q^* + K^*\frac{\partial Q^*}{\partial \operatorname{var} t_i}\right)(2x)^{-1}$$

To evaluate these formulas, we use formulas (A5), (A7), (A9) and (A11) of the appendix, each corresponding to a particular candidate solution of the logistics cost minimisation

problem. Obviously, in case 2 and 4 the shipment size and thus the annual number of trips is unaffected by a marginal change, so there is no modification of the previous marginal costs per trip. Dividing the logistics cost function K^* of case 1 and 3 into two parts, K_1 and K_2, where K_1 is the square root term and K_2 is the rest of the function (see appendix), we get:

$$\frac{\partial G}{\partial a_i} = \begin{cases} \left(k_0 + k_1 C_{\min}\right)\left(2 + K_2/K_1\right) & \text{(case 1)} \\ k_0\left(2 + K_2/K_1\right) + k_1 Q^* & \text{(case 3)} \end{cases}$$

$$\frac{\partial G}{\partial \overline{t_i}} = \begin{cases} \left[\left(w + i_0 + i_1 C_{\min}\right) + \frac{1}{2}\left(\frac{\sigma_D}{\mu_D}\right)^2 VOR\right]\left(2 + K_2/K_1\right) + \frac{1}{2} pQ^*\left(j + \left(1 + K_2/K_1\right)h\right) & \text{(case 1)} \\ \left[\left(w + i_0\right) + \left(\frac{\sigma_D}{\mu_D}\right)^2 VOR\right]\left(2 + K_2/K_1\right) + i_1 Q^* + \frac{1}{2} pQ^*\left(j + \left(1 + K_2/K_1\right)h\right) & \text{(case 3)} \end{cases} \quad (27)$$

$$\frac{\partial G}{\partial \operatorname{var} t_i} = \frac{1}{2} VOR\left(2 + K_2/K_1\right)$$

Generally, an increase in the length, time and variability of transports makes trips more costly for two reasons: directly, and by reducing the optimal number of shipments, thereby calling for larger vehicles for each shipment, some increase in the cost of holding inventory, and some increased risk of stock-out. It is seen from (27) that the indirect effect is considerable.

To the extent that stated preference analysis produces higher values of time than time dependent transport costs plus the cost of holding Q^* during transport, this must be explained as respondents' awareness of the negative effects of long transport times on late deliveries (stock-out costs) or on the need to keep safety stocks to avoid stock-out. Perhaps it might also be because respondents realise that the optimal vehicle size and number of shipments will change. Alternatively, their valuation may not be based on long-term policy as we do here, but on a perceived opportunity to avoid the occasional negative consequences of their long-term policy.

4.1 Implementation issues

To evaluate the VOR, it seems we need the cost per occurrence and duration of backorders. In practice, it is probably best to assume that $\pi = 0$ and to introduce the service level criterion that a proportion S_2 of all sales should be from shelf. By the definition of E in equation (11),

$$S_2 = 1 - \frac{\alpha(R)}{Q} \quad (28)$$

From (A4) of the appendix, then, we derive

$$\hat{\pi} = \frac{pHS_2}{1 - S_2} \quad (29)$$

which is used in VOR to get

$$VOR\big|_{\pi=0} = \frac{1}{2}\frac{ph\mu_D}{\left(1 - S_2\right)}\left(1 - F\left(\frac{R^* - \mu_L}{\sigma_L}\right)\right) \quad (30)$$

Now $F\left(\left(R^* - \mu_L\right)\sigma_L^{-1}\right)$ is the probability that no stock-out occurs in a delivery period. This is sometimes used as a service level criterion, although a crude and unreliable one. If we call it S_1, we see that (30) is indeed identical to (5). Setting $\hat{\pi} = 0$ yields a slightly more complex expression for VOR:

$$VOR\Big|_{\hat{\pi}=0} = \frac{1}{4} ph\mu_D \left[\frac{S_2}{1-S_1} \frac{Q^* f}{\sigma_L} + \left(1 - S_1\right) \right] \tag{31}$$

The type of goods in question should determine which formula to use. If a stock-out must lead to delay of some activity, the cost per occurrence might be ignored and the cost per time unit is the important element, and the other way round if the activity can go on as planned provided some costly emergency action is taken.

The VOR of (30) or (5) of a particular firm specific goods flow can be assessed by collecting data from sales statistics and asking a few questions. The case of (31) is a bit more involved. If the firm uses S_1 as a service level criterion, there will be a one-to-one correspondence between S_1 and the corresponding safety factor $k = \left(R^* - \mu_L\right)\sigma_L^{-1}$. Then $f = f(k)$ may easily be assessed from a normal distribution table. There will still be the need to ask for Q and to assess σ_L. Anyhow, the data involved here is all about the policy of the firm, the demand for its product and the transport conditions. In principle, it should not be more costly to collect and process these data than to perform a stated preference analysis.

5 Conclusions and further work

Complementing an ordinary formulation of logistics cost with all the various elements of transport cost, an explicit logistics cost function for the case of a one-to-one supply chain has been derived. Values of time and reliability are derived by the envelope theorem. The main findings are the VOR formula (4), its simplified versions (30) and (31) and its role in the evaluation of expected transport time reductions and reductions of the variance of transport time (value of time formula (25) and value of reliability formula (26)).

These findings constitute a first step to establish a new framework for the appraisal of benefits to freight transport. To make the framework operational, data on the variables entering the formulas must be had. It is thought that this can be done by collecting and processing micro-data at the level of the goods flow of firms. Questions should concern the annual costs of the firm, not just single shipments. Steps to get the necessary data have been taken in an ongoing project to integrate logistics into the Norwegian and Swedish national freight transport models. The next step is to establish objective and reliable information about the impact of policies on transport time variance. Without it, the VOR is useless. This step is ongoing work now. It may be assumed that unexpected incidents of various kinds constitute a main cause of variability, and that "internal" causes (own driver and vehicle) and more foreseeable events such as road works and bad weather conditions make up the rest. The level of traffic will be an important factor in determining the consequences of such initial disturbances. Clearly, reliability enhancing policies will also to some extent affect expected transport time, and our μ_T and σ_T will not be independent variables.

When two firms using different suppliers but selling in the same area are merged, inventories can be reduced according to the "square root formula" (Tyagi and Das 1998). This is a clue to how the model can be modified to take account of more than one supplier.

Acknowledgement

The work reported here was carried out under contract no. AL90 B 2003: 26847 from the Swedish Road Administration's Research, Development and Demonstration Programme. Lars-Göran Mattsson pointed out errors and made useful comments on an earlier draft. Remaining errors and ambiguities are my own responsibility.

References

Bruzelius, N. (2001) The Valuation of Logistics Improvements in CBA of Transport Investments: A Survey. Report to SIKA, December 2001.

de Jong, G.C. (2000) Value of freight travel time savings. In: Hensher, D.A. and K.J. Button (eds) Handbook of Transport Modelling. Elsevier, Amsterdam.

Hadley, G. and Whitin, T.M. (1963) Analysis of Inventory Systems. Prentice-Hall Inc., Englewood Cliffs, N.J.

Tyagi, R. and Das, C. (1998) Extension of the Square-Root Formula for Safety Stocks to Demands with Unequal Variances. Journal of Business Logistics 19(2), 197-203.

Appendix

The logistics cost function

We want to solve the problem (16). Forming the Lagrangian L,

$$L = -K - \lambda_1\left(C_{\min}x - QY\right) - \lambda_2\left(QY - C_{\max}x\right) - \lambda_3\left(x - Y\right)$$

the Kuhn-Tucker conditions for a maximum are

A 1.
$$\frac{\partial L}{\partial Q} = \frac{\partial L}{\partial Y} = \frac{\partial L}{\partial R} = 0$$
$$\lambda_1 \geq 0 \left(= 0 \text{ if } QY > C_{\min}x\right)$$
$$\lambda_2 \geq 0 \left(= 0 \text{ if } QY < C_{\max}x\right)$$
$$\lambda_3 \geq 0 \left(= 0 \text{ if } Y > x\right)$$

From this we derive

A 2. $$\gamma_1 x Q^{-2} - \left(\gamma_2 x + \gamma_3 Y + \gamma_4\right) + \left(\lambda_1 - \lambda_2\right)Y = 0$$

A 3. $$-\gamma_3 Q - \gamma_6 + \left(\lambda_1 - \lambda_2\right)Q + \lambda_3 = 0$$

A 4. $$pHQ = \pi x\left(1 - F\left(\frac{R - \mu_L}{\sigma_L}\right)\right) + \left(pH + \hat{\pi}\right)\alpha\left(R\right)$$

Since the three constraints can be binding or non-binding, there are 8 cases to consider. However, if all three constraints or only the first and second constraint is binding, there is an

immediate contradiction. Also, if none or only the second is binding, there is contradiction in (A3). We are left with four cases:

Case 1: *Only the first constraint is binding*. Here, $\lambda_2 = \lambda_3 = 0$ and $QY = C_{\min} x$. From (A2) and (A3), then, we get

A 5. $$Q^* = \sqrt{\frac{x\left(\gamma_1 + \psi\left(R\right) + \gamma_6 C_{\min}\right)}{\gamma_2 x + \gamma_4}}$$

A 6. $$Y^* = C_{\min} x \sqrt{\frac{\gamma_2 x + \gamma_4}{x\left(\gamma_1 + \psi\left(R\right) + \gamma_6 C_{\min}\right)}}$$

Since $\psi\left(R\right)$ is a function of *R*, this is not an explicit solution. The optimal (*Q**,*R**) is found by starting by computing *R* from (A4) given that *Q* is set to Q_w, the value we get if we disregard uncertainty and set $\alpha\left(R\right)$ and $\beta\left(R\right)$ to zero, then use the resulting *R* to compute a new *Q*, etc. This iterative procedure converges quickly to the candidate (*Q**,*R**).

Since in this case, $Y > x$,

$$\sqrt{\frac{x\left(\gamma_1 + \psi\left(R\right) + \gamma_6 C_{\min}\right)}{\gamma_2 x + \gamma_4}} < C_{\min}$$

or $Q^* < C_{min}$.

Case 2: The first and third constraints are binding. $\lambda_2 = 0$. From the two binding constraints, we immediately get

A 7. $$Q^* = C_{\min}$$

A 8. $$Y^* = x$$

*R** follows from (A4). From the conditions that λ_1 and λ_3 are positive, we may derive

$$\sqrt{\frac{\left(\gamma_1 + \psi\left(R\right)\right)x}{\left(\gamma_2 + \gamma_3\right)x + \gamma_4}} \le C_{\min} \le \sqrt{\frac{x\left(\gamma_1 + \psi\left(R\right) + \gamma_6 C_{\min}\right)}{\gamma_2 x + \gamma_4}}$$

Case 3: Only the third constraint is binding. Here, $\lambda_1 = \lambda_2 = 0$ and

A 9. $$Q^* = \sqrt{\frac{\left(\gamma_1 + \psi\left(R\right)\right)x}{\left(\gamma_2 + \gamma_3\right)x + \gamma_4}}$$

A 10. $$Y^* = x$$

(A9) is not an explicit solution, and (*Q**, *R**) must be found from (A9) and (A4) by the same iterative procedure as in case 1. From the non-binding constraints, however, we have that $C_{\min} < Q^* < C_{\max}$.

Case 4: The second and third constraints are binding. Here, $\lambda_1 = 0$ and from the binding constraints,

A 11. $Q^* = C_{\max}$

A 12. $Y^* = x$

From the condition that $\gamma_2 \geq 0$ it follows that

$$\sqrt{\frac{\left(\gamma_1 + \psi\left(R\right)\right)x}{\left(\gamma_2 + \gamma_3\right)x + \gamma_4}} \geq C_{\max}$$

Also, $\lambda_3 \geq \lambda_2$.

The value function, i.e., the logistics cost function in our model, is determined by

1. If the R^* that solves $pH\sqrt{x\left(\gamma_1 + \psi\left(R\right) + \gamma_6 C_{\min}\right)\left(\gamma_2 x + \gamma_4\right)^{-1}} = -\frac{\partial\psi\left(R\right)}{\partial R}x$ satisfies
   $\gamma_1 + \psi\left(R^*\right) < \left(\gamma_2 + \gamma_4 x^{-1}\right)C_{\min}^2 - \gamma_6 C_{\min}$,
   $K^* = 2\sqrt{x\left(\gamma_1 + \psi\left(R^*\right) + \gamma_6 C_{\min}\right)\left(\gamma_2 x + \gamma_4\right)} + \left(\gamma_5 + \gamma_3 C_{\min}\right)x + pH\left(R^* - \mu_L\right)$

2. If the R^* that solves $pHC_{\min} = -\frac{\partial\psi\left(R\right)}{\partial R}x$ satisfies
   $\left(\gamma_2 + \gamma_3 + \gamma_4 x^{-1}\right)C_{\min}^2 \geq \gamma_1 + \psi\left(R^*\right) \geq \left(\gamma_2 + \gamma_4 x^{-1}\right)C_{\min}^2 - \gamma_6 C_{\min}$,
   $K^* = \left(\gamma_1 + \psi\left(R^*\right)\right)xC_{\min}^{-1} + \left(\left(\gamma_2 + \gamma_3\right)x + \gamma_4\right)C_{\min} + \left(\gamma_5 + \gamma_6\right)x + pH\left(R^* - \mu_L\right)$

3. If the R^* that solves $pH\sqrt{\left(\gamma_1 + \psi\left(R^*\right)\right)x\left(\left(\gamma_2 + \gamma_3\right)x + \gamma_4\right)^{-1}} = -\frac{\partial\psi\left(R\right)}{\partial R}x$ satisfies
   $\left(\gamma_2 + \gamma_3 + \gamma_4 x^{-1}\right)C_{\min}^2 < \left(\gamma_1 + \psi\left(R^*\right)\right) < \left(\gamma_2 + \gamma_3 + \gamma_4 x^{-1}\right)C_{\max}^2$,
   $K^* = 2\sqrt{\left(\gamma_1 + \psi\left(R^*\right)\right)x\left(\left(\gamma_2 + \gamma_3\right)x + \gamma_4\right)} + \left(\gamma_5 + \gamma_6\right)x + pH\left(R^* - \mu_L\right)$

4. If the R^* that solves $pHC_{\max} = -\frac{\partial\psi\left(R\right)}{\partial R}x$ satisfies
   $\left(\gamma_1 + \psi\left(R^*\right)\right) \geq \left(\gamma_2 + \gamma_3 + \gamma_4 x^{-1}\right)C_{\max}^2$,
   $K^* = \left(\gamma_1 + \psi\left(R^*\right)\right)xC_{\max}^{-1} + \left(\left(\gamma_2 + \gamma_3\right)x + \gamma_4\right)C_{\max} + \left(\gamma_5 + \gamma_6\right)x + pH\left(R^* - \mu_L\right)$

The Bolzano-Weierstrass theorem says that a continuous function on a non-empty compact set achieves a maximum and a minimum there. The existence of a solution to the problem is guaranteed by the Bolzano-Weierstrass theorem provided we can assume that there is a strictly positive minimal order quantity, because in that case, the feasible region is compact.

2.10 Production technology and cost functions in scheduled transport systems

Production technology and cost functions in scheduled transport systems

Harald Minken

Institute of Transport Economics

Contents

1 Introduction

In a wide range of analyses, there is a need to assess the cost of providing transport services. According to Small (1992, chapter 3) three general approaches to the estimation of cost functions may be discerned – the statistical approach, the accounting approach and the engineering approach. The choice of approach should be governed by the purpose of the analysis, the extent to which one is willing to make use of a priori theoretical or practical knowledge, and the availability of data.

The statistical approach to the estimation of transport cost functions is used to pass judgement on economies of scale, with the purpose of predicting the market structure following from deregulation or assess the need for regulation. Aggregate data from firms and flexible mathematical forms like the translog are used in these kinds of analyses. For the purposes of cost-benefit analysis or normative pricing theory, accounting cost functions are often used. Perhaps originating in CIPFA (1974), accounting cost functions typically treat costs as stemming from three separate cost drivers, the number of vehicles used in the peak period, transport work in tonne-kilometres or passenger kilometres, and transport volume in tonnes or passengers. Data in this case is taken from the accounts and other statistics from the firms actually operating in the study area, or typical unit values derived from statistical surveys and other sources are used. For the purpose of assessing operating costs in large freight transport and passenger transport models or in models to determine optimal vehicle routing, costs may be partitioned into time-dependent costs (wages, capital costs of the vehicles), handling costs at terminals and the costs of traversing the links, or time costs may be distributed to the links.

A reliable transport cost function may serve as a norm in analyses of cost inefficiency or as a basis for reimbursement to the firm in the procurement of public transport services. Modified to include terms depending on the firm's effort and "type", it may serve as a basis for procurement and regulation under asymmetric information.[34]

This article aims at deriving analytical cost functions that can be used alone or embedded in larger problems to serve all of these purposes. The approach is neither econometrical nor based on accounting data. It utilises instead the specific structure of scheduled transport service production to derive the production function. Minimising costs with the relationships describing the production technology as constraints gives us the cost function. This is an example of the engineering approach to transport cost functions. Its characteristic features are that it utilises a priori theoretical knowlegde about the production process and combines it with estimated or empirically derived relationships between some of the variables, such as the relationship between the price and the size of the vehicles, or fuel consumption as a function of vehicle size and speed. The result is fully specified production and cost functions with no further need of estimation. In principle, these can be tested by performing a second order Taylor expansion of the log of the cost function and comparing the parameters of this function with a corresponding estimated translog function. However, our cost function is based on the assumption that the route or line structure is exogenously given, and this should be remembered when interpreting the results.

Our approach is not without precedents. The basic relationships describing the production function can be found in for instance Williams and Abdullal (1993) and Gronau (2000). Combining these with an explicit expression for the time taken by a circular trip, as in Jara-

[34] Pedersen (1995) and Pedersen (2003) are examples.

Diaz (1982) and Jara-Diaz and Barro (2003), we are already there as far as the simplest transport operations are concerned.

The usefulness of cost functions of such simple operations to the analysis of more complex systems is crucially dependent on the definition of the system whose cost function we have derived – the *operation*. An operation must be large and complex enough to make it reasonable to assume that the transport company allocates its resources (vehicles, crew) in a fixed way between operations for as long as schedules stay in place. That way, the cost of the company is the sum of the costs of its operations. The problem of minimising this sum by redesigning the operations can be approached with explicit mathematical expressions for the individual operations. Thus, our cost functions are *conditional* cost functions in the sense that they take route structure as given, but their explicit analytical form may be of help in designing better routes.[35]

To achieve a high degree of separability between operations, output is defined as service capacity per time unit. Demand influences transport cost (a) because it determines loading and unloading time and cost, (b) because the actual load may mean a difference to fuel consumption or speed, and (c) because at equilibrium, the provided capacity should be sufficient to cater to all demands. These kinds of dependencies are typical of services, where production cannot be wholly separated from consumption. However, they take a relatively mild form in transport and can easily be accounted for. Whereas a hair stylist cannot produce anything meaningful without the customer being present in the chair, the production of transport capacity is a meaningful concept.

Of course, the actual relocation from origins to destinations of flows of goods or passengers must be considered the final products. Their production requires the combination of service capacity and user inputs. In normative pricing theory, user time costs and operating costs are added to form the total social cost to be mimimised (Mohring 1972, Turvey and Mohring 1975, Jansson 1980, 1984, and Gronau 2000). This procedure is structurally equivalent to the minimisation of total logistic costs in the provision of goods and produces the same square root formula (see for instance Daganzo 1996). Implicitly or explicitly, the minimisation of the sum of these two different kinds of cost assumes a decision-maker with the responsibility for both. Descriptive theory, on the other hand, must recognise that the transport operator and the user are often two different decision-makers with separate objectives. Thus, there is a use for a cost function covering only the operator's cost, and there is a need to model the commercial relationship between the operator and the user by deriving fares or freight rates from the model.[36]

The users' concern about frequency can be taken care of by defining frequency as an additional output. Based on this, one can proceed normatively to minimise the sum of user time costs and the operator's cost function with respect to frequency, taking demand as given, or, using demand functions that are sensitive to frequency to maximise profit (or social

[35] Jara-Diaz and Barro (2003) show that there may be considerable economies of scope involved in extending the geographical coverage and utilising the opportunities that this gives for redesigning routes. They admit, however, that there are no simple ways of studying them or even characterise them.

[36] McCann (2001) points to the anomalies created by using empirically derived rate structures as input to a minimisation problem whose solution has obvious implications for rates. However, it seems to me that his solution is only valid if the shipper and carrier have merged.

welfare, as the case may be), using constraints to secure that the provided capacity is sufficient to cater for the demand.

Section 2 defines concepts and lists assumptions. Section 3 and 4 derive the production and cost functions in the four cases of with and without frequency as an output and with and without the number of vehicles taking integer values only. In some of these cases, the square root formula appears, even without including user costs in the problem. In others, the accounting cost approach is vindicated. Implications for the structure of prices are derived. Section 5 discusses loading and unloading, and Section 6 discusses future work, especially with respect to taking account of uncertainty and unreliability, using speed as a choice variable and in-vehicle time as an additional input; and with respect to embedding the conditional cost functions in larger mathematical programming problems concerning route structure, fleet management, market outcomes etc. Section 7 concludes.

2 Concepts and assumptions

The model applies to passenger and freight and all modes. Loading and unloading is treated with freight in mind but may be simplified to suit passenger transport. Trains and boats and planes, buses and trailers are all called vehicles. Vehicles have three main characteristics: design speed, number of doors or outlets, and carrying capacity (most often called size). For trains and (to some degree) trailers, size has two dimensions, size of cars and number of cars. This two-dimensionality is only sketched and not analysed here but may easily be included in the analysis. It is assumed that for any feasible combination of design speed and number of outlets, the class of all vehicles with this combination of speed and outlets contains a continuum of vehicle sizes, from minimal to maximal size. The items to be transported are uniform in weight, volume and other characteristics of importance. Carrying capacity and loading and unloading capacity are measured in items per vehicle and items per unit of time, respectively.

The unit of time may be taken to be a day, a week or whatever is most suitable for the definition of an operation (see below). The distance unit is probably kilometres in most cases. All variables and parameters must be measured in the same time and distance units, even for instance acceleration and deceleration rates, or else conversion factors must be put into the formulas.

The decision-maker (mostly a transport company, a shipper) is called the operator. The operator is a price-taker in input markets. For any given operation and given outputs, she seeks to minimise costs. There is no uncertainty.

An operation is a transport service (a number of departures) on a circular route. The circular route consists of a number of nodes (two or more) connected by the same number of links. Some of the nodes (two or more) are designated as stops, at which loading and unloading might take place. The service must fulfil the following requirements:

1. From a wider class of feasible vehicle types, one and only one type is selected to perform the service.
2. From the whole set of stops, a subset (two or more) is chosen and used by each departure.
3. The service is repeated periodically, with a period length of one unit of time.

The definition of an operation does not require the departures to be spread out evenly over the period. Thus, there might be peak and off-peak periods in the basic period, but they will all be served by the same vehicles and using the same stops. If different vehicles sizes are used for peak and off-peak services, or if there are express services, there are two or more different operations taking place on the same circular route.

A timetable is implied by the requirement that the exact same service is repeated period after period. Periodical repetition is what makes the operation a scheduled transport system, or at least a building block of a scheduled transport system. It makes it sensible to assume that the operator commits the same vehicles to the operation for a long time, thereby justifying the assumption that the costs of the operation are separable from the costs of other operations. Also, it may be assumed that it is difficult or costly to change the timetable, as it involves acquiring new vehicles or selling vehicles, renegotiating contracts with customers or authorities, informing customers etc. This takes some time – probably at least months. Consequently, the cost function of the operation will be a "medium run" cost function (Jansson 1984). It is not a cost function at the stage of vehicle design or even route design, which would both have created many more substitution possibilities and a more "putty" production function.[37] But neither is the production function "clay", since the number of vehicles and their size, along with frequency and possibly loading and unloading intensity and operating speed can still be chosen.

In this article, speed will be taken as given, either by a regulatory authority or as design speed. This is realistic in most cases were no uncertainty is involved.

The output is service capacity per period or equivalently (since the route is given) capacity kilometres per period. In addition, quality variables like frequency, in-vehicle time and punctuality might be considered outputs. In this article, in-vehicle time is not considered an output, since neither speed nor the number of stops can be chosen. Punctuality is not considered, since there is no uncertainty.

3 The technology of a transport operation

Variables and parameters

Endogenous variables:

c	vehicle capacity (items/vehicle)
c_1	capacity of a car (train case)
f	frequency (departures/time unit)
g	fuel consumption per vehicle kilometre[38]

[37] The putty/clay distinction between different types of production functions was introduced in Johansen (1959).

[38] As an approximation, other distance-dependent inputs like oil, tyres and distance-dependent parts of maintenance and insurance are assumed to be proportional to fuel consumption, and the price of fuel is adjusted upwards to take that into account. Alternatively, g might have been considered a vector of inputs, but this would not really have provided much more scope for substitution.

g_s fuel consumption per stop (due to acceleration from stop)

k number of vehicles employed

l crew per vehicle

n number of cars per vehicle (train case)

t round-trip time

t_ℓ loading plus unloading time per item

u unproductive time per roundtrip (turn-around time and waiting to depart)

y the operation's service capacity (items/time unit)

Endogenous variables (input space):

v_0 number of vehicles allocated to the operation (size of fleet)

v_1 fleet capacity (items)

v_2 the operation's fuel consumption per time unit

v_3 loading and unloading intensity (items/time unit)

Exogenous variables:

d round-trip distance (sum of the distances of the links)

$\overline{f}$ highest allowable frequency under security regulations

s number of stops used

t_s time per stop (excluding loading and unloading time)

$\underline{u}$ lowest possible or allowable turn-around time

v marching speed (kilometres/time unit)

The exogenous variables stem from route design (d and s), regulations ($\overline{f}$, $\underline{u}$), and vehicle design and/or regulations (v, t_s).

Parameters:

ϕ load factor

λ number of simultaneous loading and unloading operations

r, w, w_0, w_1, w_2, w_l prices.

The load factor is defined in an unusual manner as transport volume divided by service capacity. Usually, the load factor is defined as item-kilometres divided by capacity kilometres. Denoting the average transport distance for each item by d' and the commonly used load factor by ϕ', $\phi' = (d'/d)\phi$. Our definition is the most convenient in case the actual load mainly influences loading and unloading time and costs, not in-vehicle time and costs.

Empirical content – linear relationships

The empirical content of the model consists of the description of the operation and the assumptions given above, the exogenous variables and parameters, and the following linear relationships:

(1) $$l = l_0 + l_1 c$$

(2) $$c = c_0 + n c_1$$

(3) $$g = g_0 + g_1 c$$

(4) $$r = r_0 + r_1 c$$

(5) $$g_s = g_{s0} + g_{s1} c$$

The first equation says that the crew size is a linear function of vehicle size. In most cases, $l_1 = 0$, but if not, this is admittedly just a coarse approximation. Equation (1) shows that even if there might be substitution possibilities between labour and capital at the level of vehicle design and choice of vehicle type, these are all gone when type has been chosen, and all that remains is to choose size.

The next equation relates vehicle capacity to the capacity of the cars. In this paper, we will not treat the case of many cars, so we use $n = 0$. The third equation relates fuel consumption per kilometre to vehicle size. In the fourth, r is the price of using a vehicle for a unit of time (for instance, hire price plus distance independent parts of maintenance and insurance). This price is a linear function of vehicle size. The empirical evidence for equation (3) and (4) is probably strong in most cases.[39]

Equation (5) says that the fuel consumption per stop (due to acceleration from the stop) is a linear function of vehicle size. This is mainly of interest in air transport.

Identities

Here, we assume one period only, i.e. departures are evenly distributed over the period, and there are no peaks and off-peaks.

The next three equations are identities defining round-trip time t, frequency f and service capacity y, respectively, provided the period is homogeneous.

(6) $$t = \frac{d}{v} + t_s s + t_l \phi c + u$$

The round-trip time is the sum of the time in motion, d/v, the time taken at stops, and unproductive time u. The time taken at stops consists of two terms. s is the number of stops and t_s is the extra time taken by decelerating from marching speed to full stop, making ready for loading and unloading, and accelerating to marching speed. It may also include average waiting time in case there is a chance that vehicles queue up at stops. In that case, t_s is a function of frequency. This can easily be incorporated into (6) with the help of queuing theory, but at the price of more complex mathematics when the model is solved. Since t_l is the

[39] With respect to buses, see Jansson (1984), chapter 5.

time taken to load and unload one item, and ϕ is the load factor, the term $t_l \phi c$ is the time taken by loading and unloading per round trip.

The next two identities are

(7) $k = tf$

(8) $y = cf$

To verify (7), check that the definition of frequency is reasonable for $k = 1$, then for $k = 2$ etc. Identity (8) should be self-evident.

These three relationships have of course been noted by earlier writers. Jara-Diaz (1982) has a fairly elaborate version of (6), while Williams and Abdullah (1993) uses a very simplified version. Relations (7) and (8) are used by many writers. These three relationships all but determine the production technology.

The production technology

By our assumptions, c is a non-negative real number. Assume that f too is a non-negative real number.[40] On the other hand, k must be an integer – an element of the natural numbers N. The identity (7) holds good in any case, because unproductive time u will have to adjust. Technical efficiency, however, requires that u is chosen as small as possible, or in other words that in every case, k is chosen to be as small as possible subject to $u \geq \underline{u}$. We define $\lceil k \rceil$, the least integer number of vehicles required, by

(9) $$\lceil k \rceil = \min \left\{ x \in N \mid x \geq \left(\frac{d}{v} + t_s s + t_l \phi c + \underline{u} \right) f \right\}$$

We then define the unproductive time $\tilde{u}$ that satisfies the inequality in (9) with equality by

(10) $$\lceil k \rceil = \left(\frac{d}{v} + t_s s + t_l \phi c + \tilde{u} \right) f$$

If, on the other hand, k can be any non-negative real number, efficiency requires

(11) $$k = \left(\frac{d}{v} + t_s s + t_l \phi c + \underline{u} \right) f$$

The difference between k in (11) and $\lceil k \rceil$ in (10) is

(12) $$k^{\triangleright} = \left(\tilde{u} - \underline{u} \right) f$$

$k^{\triangleright}$ is the "excess inventory of vehicles" caused by the indivisibility of vehicles. Intuitively, if increasing output by reducing this excess is possible, there will be increasing returns to scale.

Having thus eliminated the indeterminate variable u, we have two cases: the large k case where (8) and (11) applies, and the small k case, where (8) and (10) applies. We treat each in

[40] In many scheduled transport systems, a rigid schedule is applied. This means that departures occur "at the same time" every hour (or every day or week). The frequency per vehicle can then only take certain values. We do not analyse this case here.

turn. But first, let us note possible restrictions on the minimum and maximum vehicle size or on the maximum allowable frequency under safety regulations. They might apply in all cases:

$$c_{\min} \le c \le c_{\max}, \quad f \le \overline{f} \tag{13}$$

It might be noted that the existence of a minimum vehicle size or a maximum allowable frequency will generally invalidate the assumption of free disposability of inputs.

The case of large k:

The relations (8), (11) and (13) – taken together – express the technical possibilities of producing the product (y, f) in terms of the operational parameters (c, k, t_ℓ), or, alternatively, the technical possibilities of producing y as expressed by the operational parameters (f, c, k, t_ℓ). Solving for (y, f):

$$y = \frac{ck}{\frac{d}{v} + t_s s + t_l \phi c + \underline{u}} \tag{14}$$

$$f = \frac{k}{\frac{d}{v} + t_s s + t_l \phi c + \underline{u}} \tag{15}$$

Turning from the space of operational parameters to input space, we must somehow recognise the two-dimensional character of "capital" here, both as a number of vehicles and as a capacity per vehicle. The most convenient way to express this is to define the number of vehicles and the fleet capacity as two distinct inputs. The transformation to input space is performed by:

$$\begin{aligned} v_0 &= k \\ v_1 &= ck \\ v_2 &= \left(dg_0 + sg_{s0}\right) f + \left(dg_1 + sg_{s1}\right) y \\ v_3 &= t_l^{-1} \end{aligned} \tag{16}$$

Relations (3) and (5) and the identity $y = cf$ have been used in the third line. The intensity of loading and unloading operations – the number of items loaded and unloaded per unit of time – is naturally the inverse of the time taken to load and unload one item, as shown in the last line.

The equation system (16) together with (8) and (11) can be solved for the six variables y, k, c, f, t_l and v_2 as functions of v_0, v_1 and v_3. First k, c and t_l is found from line 1, 2 and 4 in (16), and substituted into (8) and (11) and line 3 of (16). Then these three equations are solved for y, f and v_2:

$$y = \frac{v_0 v_1}{v_0\left(\frac{d}{v} + t_s s + \underline{u}\right) + \phi\frac{v_1}{v_3}}$$

(17)
$$f = \frac{{v_0}^2}{v_0\left(\frac{d}{v} + t_s s + \underline{u}\right) + \phi\frac{v_1}{v_3}}$$

$$v_2 = \frac{v_0\left(dg_0 + sg_{s0}\right) + v_1\left(dg_1 + sg_{s1}\right)}{v_0\left(\frac{d}{v} + t_s s + \underline{u}\right) + \phi\frac{v_1}{v_3}} \cdot v_0$$

The functions for y and v_2 taken together describe efficient production in the case where frequency is not an output, while all three functions describe efficient production if frequency is an output. If one wishes, one may call (17) the production function for the (y, f) case, and the first and third lines the production function for the single output case. y and f are joint products – the decision about inputs to produce one of them determines the output of the other. The other peculiar feature of the production function (17) is of course that there is a relationship between the inputs that is not expressed by the first equation alone. In the terminology of Frisch (1953), v_2 is a "factor shadow". We believe such a technology is very seldom described in the literature, although it accords well with the impression that operators do not make decisions about fuel use when operations are designed. What they do, is to set the parameters k, c and t_l, which are simple transforms of v_0, v_1 and v_3, so as to minimise costs, including of course fuel costs.

Elasticity with respect to scale at the point **v** in the single output case is defined by $\varepsilon(\mathbf{v}) = \sum_i El_{v_i} y$, where $El_{v_i} y$ is the elasticity of output with respect to input i. These elasticities are:

(18)
$$El_{v_1} y = \frac{Ev_1}{Av_0 + Ev_1},\ El_{v_2} y = \frac{Av_0}{Av_0 + Ev_1},\ El_{v_3} y = \frac{Ev_1}{Av_0 + Ev_1}$$

$$\text{where } A = \frac{d}{v} + t_s s + \underline{u}, \quad E = \phi v_3^{-1}$$

There are increasing returns to scale if $\varepsilon(\mathbf{v}) > 1$, constant returns to scale if $\varepsilon(\mathbf{v}) = 1$ and decreasing returns to scale if $\varepsilon(\mathbf{v}) < 1$.[41] Since $El_{v_0} y + El_{v_1} y = 1$, there is constant returns to scale if loading and unloading intensity is fixed, but increasing returns to scale if it can be chosen.[42] These are global properties of the production function. Likewise, it can easily be shown that the elasticities of f with respect to v_0 and v_1 sum to 1 and that the elasticity of f with respect to v_3 equals $El_{v_3} y$, the elasticity of y with respect to v_3. Consequently, y and f

[41] We prefer to speak of economies of scale instead of economies of density of demand because output is defined as service capacity and frequency, not as transport volume.

[42] In passenger transport, the intensity of loading and unloading may be varied by choosing vehicle types with different number or size of doors, and by reducing or eliminating payment to the driver. The cost of these measures is reflected in the cost of vehicle types and fare collection systems.

increase in the same proportion along a ray in input space, and the conclusion about returns to scale is the same in the case of two outputs as the case of one output.[43]

At this point, it is perhaps wise to point out that queuing at the stops would have tended to decrease the returns to scale, and that the conclusions might be radically different if it is found that the linear relationships (1) – (5) must be replaced by more complex functions of *c*.

The case of small k:

It is believed that in most operations, the number of vehicles employed is small and the indivisibility of this factor cannot be ignored. Thus, we have three continuous and one integer input, and one or two continuous outputs. The production technology in this case can be found as follows: The four equations of (16) – with $\lceil k \rceil$ now taking the place of k – are solved for ($\lceil k \rceil$, c, f, t_l). The results are inserted in (8) and solved for y, and then again for f. The resulting (v_0, y, f) is then used in (9) and (10) to obtain conditions governing v_0, $\tilde{u}$ and v_2. We get:

$$v_0 = \min\left\{x \in N \mid Bx^2 + \left(Dv_1 - Av_2\right)x - Ev_1v_2 \geq 0\right\}$$

(19)

$$\text{where } A = \frac{d}{v} + t_s s + \underline{u}, \quad B = dg_0 + sg_{s0}$$

$$D = dg_1 + sg_{s1}, \quad E = \phi v_3^{-1}$$

Suppose v_3 has been set and v_1 and v_2 have somehow been preliminary determined. Then (19) determines v_0. With this v_0, the associated $\tilde{u}$ is determined by modifying A until the inequality in (19) is satisfied with equality.

The formula that determines v_2 is:

(20)

$$v_2 = \frac{v_0\left(dg_0 + sg_{s0}\right) + v_1\left(dg_1 + sg_{s1}\right)}{v_0\left(\frac{d}{v} + t_s s + \tilde{u}\right) + \phi\frac{v_1}{v_3}} \cdot v_0 = \frac{Bv_0 + Dv_1}{\tilde{A}v_0 + Ev_1} \cdot v_0$$

Equation (20) differs from the similar equation in (17) by having $\tilde{u}$ instead of $\underline{u}$. By (20), output (y, f) can be expressed in two ways, one involving $\tilde{u}$ and the other not:

(21)

$$y = \frac{v_0 v_1}{v_0\left(\frac{d}{v} + t_s s + \tilde{u}\right) + \phi\frac{v_1}{v_3}} = \frac{v_1 v_2}{v_0\left(dg_0 + sg_{s0}\right) + v_1\left(dg_1 + sg_{s1}\right)}$$

$$f = \frac{{v_0}^2}{v_0\left(\frac{d}{v} + t_s s + \tilde{u}\right) + \phi\frac{v_1}{v_3}} = \frac{v_0 v_2}{v_0\left(dg_0 + sg_{s0}\right) + v_1\left(dg_1 + sg_{s1}\right)}$$

[43] By our definition of the inputs, vehicle size is constant along any ray in input space, and y and f differ only by a multiplicative constant.

Even if (19)-(21) fully express the production technology in the integer case, the problem remains to actually determine the v_0 to use in each particular case. A distinction must be made between the case of two outputs *y* and *f*, and the case of only one input. In the first case, (21) can be solved for v_1 and v_2 as functions of *y*, *f* and v_0 and the least integer value of v_0 that satisfies (19) can be determined. In the case of only one input, v_1 as a function of v_0, v_2 and *y* is found from (21) and inserted in (19), and an iterative procedure must be used to find the values (v_0, v_2, $\tilde{u}$) that satisfies both (20) and the inequality in (19) with equality. If *f* is computed in the process of finding the correct (v_0, v_2, $\tilde{u}$), formula (12) can be used to assess the need for an increase in v_0. Unless there is a need to change v_0, $\tilde{u}$ is gradually increased until the solution is found. (The convergence of this process has not been studied.)

Thus, the production function in the case of small *k* is structurally similar to the large *k* case, but the process of finding it is more involved. The choice options are restricted by the integer requirement, since v_0 is determined endogenously. For small changes of v_1 and v_3, v_0 is fixed. In the range where this applies, v_0 is not a choice variable, although in a "wide range view", it is. An obvious way to take account of this is to derive conditional cost functions for a small range of integer values of v_0 and compare. Returns to scale in the range were the vehicle fleet is the same must be conceptually discerned from returns to scale over a range where it changes. Returns to scale in this case is best derived from the cost function.

The relevance of the cases

The relevance of the small *k* integer case has been argued. The relevance of the two-input, (*y*, *f*) case is obvious: It admits the formulation of profit maximisation or social efficiency maximisation problems where demand is elastic with respect to frequency, or of social cost minimisation problems where user costs are functions of frequency. What needs to be argued is perhaps the single output case. There are at least three instances where it is of interest. First, frequency is unimportant if the goods are very low value. Second, in well-developed markets with many suppliers, the frequency of one of them need not concern the customer unless she has somehow contracted to use this supplier only. And finally, there are loading and unloading operations. The "waiting time" between each return of a forklift to the store of goods to be loaded is totally irrelevant to costs, and only the capacity of the whole loading operation matters.

The relevance of considering loading and unloading capacity as an input must be assessed in each case. If there is a commercial relationship between the transport operator and the terminal operator, it all depends on whether or not quicker terminal services can be bought. If the transport operator controls the terminals, it depends on the time horizon of the planning and the options with respect to alternative use of the resources at the terminals.

4 The cost functions

We still assume one homogeneous period. The operational parameter t_l, or alternatively the input v_3, is assumed to be given. This makes the loading and unloading costs a constant that can be ignored in the minimisation problem. The extension to cases where t_l can be chosen will be considered in the next section. The cost of a minimal vehicle with crew will be denoted w_0, the incremental cost of increasing vehicle size will be called w_1, and the unit fuel cost is w_2. According to (1) and (3) - (5) and the section on loading and unloading, the cost

per unit time for the operation, expressed in terms of the operational parameters (*k*, *c*, *f*) or alternatively in terms of the inputs, is:

$$\begin{aligned} C &= w_0 k + w_1 ck + w_2 \left(dg_0 + sg_{s0} \right) f + w_2 \left(dg_1 + sg_{s1} \right) y \\ &= w_0 v_0 + w_1 v_1 + w_2 v_2 \\ \text{where } & w_0 = r_0 + wl_0, \quad w_1 = r_1 + wl_1 \end{aligned} \tag{22}$$

C is to be minimised, given the production technology. The resulting cost function will be the same, regardless of whether minimisation is performed with respect to operational parameters or inputs. There are four cases to consider. *Case 1* is the case where only *y* is an output and $k = v_0 \in R_+$, the non-negative real numbers. *Case 2* is the case where only *y* is an output and $k = v_0 \in N$, the natural numbers. *Case 3* has (*y*, *f*) as output and $k = v_0 \in R_+$, and Case 4 has (*y*, *f*) as output and $k = v_0 \in N$. From now on, the abbreviations A, B, D, and E defined in (19) will be used throughout. In practical applications, a term T_f could be added to B to account for taxes or other payments per departure, and a term T_y could be added to D to account for taxes on seats or passengers or the like. Taxes on labour, fuel and other inputs are assumed to be included already in the prices **w**.

Case 1: y is an output and $k = v_0 \in R_+$.

The problem in terms of operational parameters is

$$\begin{aligned} C\left(y, \mathbf{w}\right) = Min_{k,c,f} \; & w_0 k + w_1 ck + w_2 Bf + w_2 Dy \quad \text{s.t. } y = cf \\ & k = \left(A + Ec\right) f \end{aligned} \tag{23}$$

In input space, the constraints to use are the expressions for *y* and v_2 in (17). It is convenient to solve these two equations for v_0 and v_1 and insert these values into the objective function. Thus, the problem in terms of inputs becomes

$$C(y, \boldsymbol{w}) = Min_{v_2} \; w_0 \left(\frac{A(v_2 - Dy)}{B} + Ey \right) + w_1 \left(Ay + \frac{BE}{v_2 - Dy} y^2 \right) + w_2 v_2 \tag{24}$$

The results of both of these two minimisation problems are:

$$\begin{aligned} C\left(y, \mathbf{w}\right) &= \left[2\sqrt{w_1 E \left(w_0 A + w_2 B \right)} + \left(w_0 E + w_1 A + w_2 D \right) \right] \cdot y \\ k\left(y, \mathbf{w}\right) &= \left[A\sqrt{\frac{w_1 E}{w_0 A + w_2 B}} + E \right] \cdot y = v_0\left(y, \mathbf{w}\right) \\ c\left(y, \mathbf{w}\right) &= \sqrt{\frac{w_0 A + w_2 B}{w_1 E}} \qquad v_1\left(y, \mathbf{w}\right) = \left[E\sqrt{\frac{w_0 A + w_2 B}{w_1 E}} + A \right] \cdot y \\ f\left(y, \mathbf{w}\right) &= \sqrt{\frac{w_1 E}{w_0 A + w_2 B}} \cdot y \qquad v_2\left(y, \mathbf{w}\right) = \left[B\sqrt{\frac{w_1 E}{w_0 A + w_2 B}} + D \right] \cdot y \end{aligned} \tag{25}$$

Thus, it is confirmed that this case exhibits constant returns to scale, basically because there is an optimal vehicle size that is independent of *y*.

Case 2: y is an output and $k = v_0 \in N$.

We assume *k* is determined through the process indicated in the text concerning equations (19), (20) and (21). For any such *k*, the conditional cost minimisation problem in terms of operational parameters *f* and *c* becomes

$$C(y, \mathbf{w}; k) = Min_{c,f}\ w_0 k + w_1 ck + w_2 Bf + w_2 Dy \quad \text{s.t. } y = cf, \quad k \geq (A + Ec) f \tag{26}$$

Having formulated the Kuhn-Tucker conditions for an optimum, there will be two cases to consider; the case where the second constraint is binding and the case where it is not. It may be shown that costs in the non-binding case are always at least as large as costs in the binding case. Thus, if unproductive time can somehow be reduced, it always pays to do so. In the present case, the operational variables *f* and *c* provide options to reduce unproductive time.

The problem in terms of inputs features constraints derived from the expression for *y* and v_2 in (17), as in Case 1, but with inequality instead of equality. Put otherwise, *y* and v_2 in the integer case feature a $\tilde{u}$ which is at least as large as $\underline{u}$. Both problem formulations yield the same result, which is:

$$C(y, \mathbf{w}; k) = \left(w_0 + w_2 \frac{B}{A}\right) k + w_2 \left(D - \frac{BE}{A}\right) y + w_1 \frac{A}{k - Ey} ky$$

$$c(y; k) = \frac{Ay}{k - Ey} \qquad v_1(y; v_0) = \frac{Av_0}{v_0 - Ey} \cdot y \tag{27}$$

$$f(y; k) = \frac{k - Ey}{A} \qquad v_2(y; v_0) = \left(D - \frac{BE}{A}\right) y + \frac{B}{A} v_0$$

The optimal operational variables and inputs in this case are independent of input prices. There are increasing returns to scale in the range where *k* stays constant. In this range, it can easily be shown that average cost is a decreasing convex function. Average cost over a broader range will exhibit the well-known “saw-tooth” pattern with the maximum points getting smaller as *y* increases, while the minimum points stay the same. Thus, if we can always choose an output level where unproductive time is at its minimum, we experience constant returns to scale, while all points at the same distance from the nearest minimum point will form a pattern of decreasing average cost.

Case 3: (y, f) is output and $k = v_0 \in R_+$.

In the cases where frequency is acknowledged as an output, identity *y* = *cf* will mean that vehicle size is fixed, while equation (11) fixes *k*. Thus, there are no choice options. Nevertheless, this is an interesting case for two reasons. First, it might well be realistic, and second, the ensuing cost function has an interesting structure. Using (8) and (11) to eliminate *k* and *c*, *C* becomes:

$$C(y, f, \mathbf{w}) = (w_0 E + w_1 A + w_2 D) y + (w_0 A + w_2 B) f + w_1 E y^2 f^{-1} \tag{28}$$

The operator's cost depends on the number of departures (second term) and the headway (third term), exactly like the order costs and inventory costs of a shipper in the economic order quantity (EOQ) problem. The structure is also the same as the sum of operating costs and user costs in the "standard model, linear case" of Gronau (2000), provided users do not care about their waiting time. Thus, when frequency and vehicle size can be chosen, we get structurally the same results (our Case 1 and Gronau's equation 9b). Furthermore, if user costs are added to the operator's cost in (28) and the ensuing total social costs are minimised, we get Gronau's general result in his equation (13), but with different parameter values, since more elements of operating cost are structurally the same as user costs in our model.

Another possible interpretation of (28) is to see it as consisting of terms that depend on volume (the terms containing E) and terms depending on tonnekilometres (the other terms).

For completeness, we collect the "results" in this case:

$$
\begin{aligned}
&C(y, f, \mathbf{w}) = \left(w_0 E + w_1 A + w_2 D\right) y + \left(w_0 A + w_2 B\right) f + w_1 E y^2 f^{-1} \\
&k(y, f) = Af + Ey \qquad = v_0(y, f) \\
&c(y, f) = y/f \qquad\qquad v_1(y, f) = \left(Af + Ey\right) y/f \\
&\qquad\qquad\qquad\quad v_2(y, f) = Bf + Dy
\end{aligned}
\tag{29}
$$

The optimal operational variables and inputs are independent of input prices. There are constant returns to scale, as seen by multiplying *y* and *f* by the same positive constant. With respect to *y* alone, it is easily shown that average cost decreases in *y* if and only if *f* is larger than its cost minimal value (Case 1). Since the purpose of regarding *f* as an output must be to produce a higher *f* than its cost minimal value, we will in fact observe increasing returns to scale in *y* when some regard has been given to the quality variable *f*.[44] A higher frequency will however have to be bought at the price of higher fares or freight rates.

Case 4: (y, f) is output and $k = v_0 \in N$.

Once again, *y* and *f* determine *c*, and then k is determined as $\lceil k \rceil$. There are no choice options. Since unproductive time might be higher than its minimum and this cannot be adjusted, A and E will not appear in the formulas. The results are:

$$
\begin{aligned}
&C(y, f, \mathbf{w}; k) = w_0 k + w_1 k y f^{-1} + w_2 B f + w_2 D y \\
&c(y, f) = y/f \qquad v_1(y, f; v_0) = v_0 y f^{-1} \\
&\qquad\qquad\quad v_2(y, f) = Bf + Dy
\end{aligned}
\tag{30}
$$

Once again, the optimal operational variables and inputs are independent of input prices. In the range where integer *k* stays the same, there will be increasing returns to scale in (*y*, *f*) and considerable returns to scale in *y* alone. In this range, average cost in terms of *y* alone is a decreasing convex function. Average cost over a broader range will once again exhibit the "saw-tooth" pattern with the maximum points getting smaller as *y* increases, while the minimum points stay the same.

[44] That this would be the case, was also argued by Allen and Liu (1995).

Equation (30) may be seen as a specification of the accounting cost function (see Pels and Rietveld 2000), where the first term depends on the number of vehicles (in peak), the second on volume and the last two on tonne-kilometres. However, contrary to their rather harsh comments on this kind of function, Case 4 is altogether realistic in many cases. In particular, in none of the last three cases do optimal inputs depend on input prices.

Implications for the structure of freight rates and fares

The cost functions here are derived under restrictive assumptions. In particular, design speed and handling speed has not been assumed to be choice variables, the first for reasons of mathematical simplicity and the second because many periods will have to be considered first. Nevertheless, interesting implications with respect to prices can be drawn. It was seen that constant returns to scale (CRS) obtains in the cases where the number of vehicles can be approximated by a real number, and that increasing returns to scale were due to integer number of vehicles. Under CRS, marginal cost pricing and average cost pricing will coincide, while increasing returns requires the operator to apply average cost pricing. Thus, we will assume that prices are somehow set proportional to average cost.

Fares (or freight rates), understood as average cost per unit of capacity,

5 Loading and unloading

Daganzo (1996), chapter 3, describes loading and unloading as, in effect, a miniature transport operation. This is the view we will take here. We assume that the loading and unloading operation has only two stops. Furthermore, the cost of loading and unloading the loading vehicles (forklifts, jack-trolleys, cranes) is ignored or subsumed under other elements of cost. The output is items moved per time unit. Thus, frequency does not matter. On the other hand, we assume the size of the loading vehicles to be exogenously given, since they are often designed to move one pallet or one container. Most importantly, we will have to recognise that loading and unloading at a stop is not a continuously ongoing operation, but only takes place when a vehicle arrives. This makes it an extreme case of the peak and off-peak (multi-period) problem. We use the terms handling and loading and unloading interchangeably and use the subscript *h* to denote handling operations. We index stops by *s*.

For simplicity, we assume it is possible to run several handling operations at a stop simultaneously without interference between the operations. The number of outlets of the vehicle constrains the possibilities of doing this. The number of operations performed simultaneously, which we will denote λ, depends on the number of outlets and the proportion of the outlets that are actually used simultaneously. The degree to which loading and unloading are both performed on the same round-trip of the loading and unloading vehicle is captured by the load factor ϕ_{hs}. Often it can be set to ½ (empty backhauls), as for instance if the goods to be unloaded are kept in a different compartment than the goods to be loaded, or if the stops that are origins are separate from the stops that are destinations. If backhauls are not necessarily empty, the degree to which ϕ_{hs} is above ½ at each stop is determined by origin-destination matrix defined over the stops of the operation.

Output (items moved) per time unit in all the simultaneous handling operations at a stop is $\lambda\phi_{hs}y_h$, where y_h is the transport capacity of one handling operation. Loading and unloading time per item is the inverse of this, times two. Thus, we have:

$$t_l = \frac{2}{\lambda\phi_l y_l}, \quad v_3 = \frac{\lambda\phi_l y_l}{2}, \quad y_l = \frac{2v_3}{\lambda\phi_l} \tag{31}$$

Denote the cost per time unit of one loading/unloading operation by $c(y_l, w_l)$, where w_l is a vector of input prices. If all inputs have alternative use in the very short run, the cost of loading and unloading the amount ϕy per time unit is proportional to the time taken to do this. However, in most cases it is considerably more realistic to assume that, on the margin, acquiring more handling equipment to speed up the intensity of loading and unloading will mean more idle equipment for longer periods of time.

6 Possible applications

The cost functions have not yet been tested, nor have they been applied in larger models or problems. Empirical testing may take place in various ways, due to the analytical character of the functions. For instance, if empirical data on fares and the variables that determine fares in our models exist, a regression model may be formulated. Alternatively, empirically derived elasticities of fares or freight rates with respect to input prices, distance etc. may be compared to the theoretically derived elasticities.

Once empirically validated, the cost functions may be of use in a variety of problems. We sketch some of them briefly. First we introduce the demand side, then we extend the system under study to many operations, and finally we discuss the design of a system of many operations.

Consider a single, homogeneous operation. Index the links of the circular path on which the operation is run by *a*, ordered pairs of zones where trips or shipments originate and terminate by *w*, and ordered pair of stops by *r*. The share of demand on relation *w* that chooses stops *r* is denoted by x_{rw}. For our purpose we need not specify it except to note that it is a function of the vector of fares/freight rates $\mathbf{p} = \left(p_1, ..., p_r, ..., p_{s(s-1)}\right)$ and possibly frequency *f*. Assume Case 3 applies, so demand *is* responsive to frequency and the number of vehicles required is large. If the operator is able to choose fares and frequency to maximise profit, he will solve the following problem:

$$\begin{aligned} \Pi = & \underset{\mathbf{p}, f, y, \phi}{Max} \sum_r \sum_w p_r x_{rw} - C(y, f, \phi) \\ \text{st.} \quad & \sum_r \sum_w \delta_{ar} x_{rw} \le y \qquad a = 1, 2, ... \quad (\mu_a) \\ & \sum_r \sum_w x_{rw} = \phi y \qquad\qquad\qquad (\mu) \\ & \mathbf{p} \ge \mathbf{0},\ f \ge 0, y \ge 0, \phi \ge 0 \end{aligned} \tag{32}$$

where the Kronecker delta is 1 if link *a* is crossed on the way from the first to the second stop of the ordered pair *r*, and 0 otherwise. Lagrangian multipliers corresponding to the constraints have been noted in parentheses. The first set of constraints secure that nowhere along the

route are the vehicles "overcrowded", while the last constraint at long last defines the load factor. We have made the cost function an explicit function of this factor. Note that y and ϕ together play the role that actually transported volumes would have played if output had not been defined the way it was, and that the constraints provide a convenient bridge between capa-city, in terms of which the cost function is formulated, and actual demand.

Rearranging the Kuhn-Tucker conditions slightly, the conditions for a profit maximum are:

$$
\begin{aligned}
&(-1)\sum_w x_{rw} = \sum_r \left(p_r - \sum_a \mu_a \delta_{ar} - \frac{1}{y}\frac{\partial C}{\partial \phi} \right) \sum_w \frac{\partial x_{rw}}{\partial p_{r'}} \qquad r' = 1,2,...,s(s-1) \\
&\frac{\partial C}{\partial f} = \sum_r \left(p_r - \sum_a \mu_a \delta_{ar} - \frac{1}{y}\frac{\partial C}{\partial \phi} \right) \sum_w \frac{\partial x_{rw}}{\partial f} \\
&y\frac{\partial C}{\partial y} = y\sum_a \mu_a + \phi \frac{\partial C}{\partial \phi} \\
&\mu_a \left(\sum_r \sum_w \delta_{ar} x_{rw} - y \right) = 0 \qquad\qquad a = 1,2,...
\end{aligned}
\tag{33}
$$

In case the number of vehicles is small and integer, the procedure to find the right k to use must be performed simultaneously with the solution of (33). The Cases 1 and 2 where f is of no interest are handled by omitting the second line of (33). If fares are also exogenously given, the first line is dropped, and the profit maximisation consists simply in adapting output to actual demand. Thus, by an appropriate solution algorithm, we will be able to find the profit of the operation in all cases.

Another kind of problem that might be formulated and solved in much the same manner is the maximisation of social welfare. User benefits UB are if possible expressed as the potential function whose derivate with respect to generalised costs for each r is the negative of the demand x_r. Social welfare $W = UB + \Pi$. W might be maximised with respect to the same variables, or profit is maximised first with respect to the variables in the hands of the operator, and then W is maximised with respect to the remaining variables.

The extension to the case of one operator running many operations might simply be a case of including all operations in the objective function, be it profit or social welfare. However, some new features are likely to appear. First, the demand functions x_{rw} will probably be functions of the combined frequency of operations servicing the relation r.[45] In the limit, this makes demand indifferent to small changes in the frequency of a particular operation (if this particular operation is run by a different operator from the rest, the case of cost minimisation with only y as an input appears). Next, the operations might interact with each other on the supply side in at least three different ways: by creating congestion on the links, by creating queues at the terminals and by sharing the fixed element of handling costs at the terminals. Our cost functions are capable of incorporating these effects. Finally, a peculiar and potentially important form of economies of scope appears if the same kind (and size?) of vehicle is used in many operations. This is the reduction in the number of vehicles that will have to be kept in reserve at any point in time to allow for repair and maintenance. A square root law applies to the size of this safety inventory (Das 1975, Das and Tyagi 1997).

[45] See Williams and Abdullal (1993) for convenient formulations of this.

With the extension to many operations, the possibility of many operators also arises. Between them they constitute a transport supply market. If the form of competition between these suppliers has been determined, the market outcome can probably be determined along the lines of Williams and Abdullal (1993) and Williams and Martins (1993), even if our operations and cost functions are generally more complex than theirs.

Suppose the equilibrium conditions in any of these complex cases have been formulated. The question arises if total profit or social welfare can be increased by redefining the operations or by reallocating responsibility for the operations among the operators. This is a design problem with a bi-level structure. The upper level problem is a combinatorial problem. The lower-level problem is the problem to find the conditional equilibria obtaining in the supply markets and the trip/shipping markets given the design of the set of operations and the assignment of them to the operators. As demonstrated by Ivanova (2003), a mixed complementarity formulation of the equilibrium conditions in each of the markets allow the conditions to be added and solved simultaneously. This is potentially the best approach to design problems with equilibrium constraints and might constitute a possibility of systematically exploring the economies of scope pointed out by Jara-Diaz and Barro (2003).

Our cost functions were conditional on the given design of the basic elements of the operations – the circular paths and the stops. At that level, we found that constant return to scale obtained unless the number of vehicles had to be treated as an integer. Nevertheless, there might be considerable economies of scope if different operations could be rearranged under the authority of a single decision-maker. The hints in this section might provide a way to explore this question, but as we said, this work remains to be done.

7 Conclusions

Jara-Diaz and Barro (2003) say that “finding analytical cost functions for actual transport firms serving many OD-pairs over complex networks is simply infeasible …”. They are undoubtedly right as long as the actual movement of goods is taken as the outputs. Taking transport capacity as the output and making suitably simple assumptions, the picture looks different, as shown here. The key simplification is the concept of an operation. Returns to scale is defined with respect to service capacity and the key service quality variable, frequency, and increasing returns are shown to stem from the indivisibility of vehicles and the possibility of speeding up handling operations. The cost functions that result from these assumptions take as given the design of the circular path on which the operation is run, the stops, and the design of the vehicles except for their size. They are “medium term” cost functions in the sense that the number of vehicles, their size and the frequency (timetable) can be chosen.

The influence of actual demand on cost is modelled as simply as possible at this stage and is captured by the load factor and the handling time and cost. Demand is only fully considered at the level of profit maximisation or maximisation of social welfare. At this level, the complexity resulting from the OD-matrix and the existence of many operations and operators can be handled by formulating numerically solvable constrained optimisation problems. Finally, the design of the lines, stops and optimal allocation of responsibilities among the operators can be approached as bi-level combinatorial programmes with equilibrium constraints.

This article has only set the stage for such future work.

References

Allen, W.B and D. Liu (1995) Service quality and motor carrier costs: an empirical analysis. *Review of Economics and Statistics* **77**(3), 499-510.

Bobzin, H. (1998) *Indivisibilities. Microeconomic Theory with Respect to Indivisible Goods and Factors*. Physica-Verlag, Heidelberg.

Chartered Institute of Public Finance and Accountancy (CIPFA) (1974) Passenger Transport Operations. London. Cited from van der Veer (2002).

Daganzo, C.F. (1996) *Logistics Systems Analysis. 2nd Edition*. Springer-Verlag, Berlin.

Das, C. (1975) Supply and redistribution rules for two-location inventory systems: one-period analysis. *Management Science*, **21**(25), pp. 765-776.

Das, C. and Tyagi, R. (1997) Role of Inventory and Transportation Costs in Determining the Optimal Degree of Centralization. *Transportation Research E*, 33(3), pp. 171-179.

Gronau, R. (2000) Optimum Diversity in the Public Transport Market. *Journal of Transport Economics and Policy* **34**(1), 21-41.

Ivanova, O. (2003) The Role of Transport Infrastructure in Regional Economic Development. TØI Report 671/2003.

Jansson, J.-O. (1980) A Simple Bus Line Model for Optimisation of Service Frequency and Bus Size. *Journal of Transport Economics and Policy* **14**, 53-80.

Jansson, J.-O. (1984) *Transport System Optimization and Pricing*. John Wiley & Sons, Chichester.

Jara-Diaz, S.R. (1982) Transportation product, transportation function and cost functions. *Transportation Science* **16**, 522-539.

Jara-Diaz, S.R. and L.J. Barro (2003) Transport cost functions, network expansion and economies of scope. *Transportation Research E* **39**, 271-288.

Johansen, L. (1959). Substitution versus Fixed Production Coefficients in the Theory of Economic Growth: A Synthesis. *Econometrica*, 27(2), pp. 157-176.

McCann, P. (2001) A proof of the relationship between optimal vehicle size, haulage length and and the structure of distance-transport costs. *Transportation Research A*, **35**, 671-693.

Mohring, H. (1972) Optimization and Scale Economies in Urban Bus Transportation. *American Economic Review* **62**, 591-604.

Pedersen, P.A (1995). Public Regulation of a Transport Company with Private Information about Demand. *Journal of Transport Economics and Policy*, September, pp 247-251.

Pedersen, P.A (2003) On the optimal fare policies in urban transportation. *Transportation Research Part B*, 37(5), pp. 423-435.

Pels, E. and P. Rietveld (2000) Cost functions in transport. Chapter 19 in Hensher, D.A. and K.J. Button (eds.) *Handbook of Transport Modelling*. Elsevier.

Small, K.A (1992) *Urban Transportation Economics*. Fundamentals of Pure and Applied Economics 51. Harwood Academic Publishers, Chur, Switzerland.

Turvey, R. and H. Mohring (1975) Optimal bus fares. *Journal of Transport Economics and Policy*

Van der Veer, J.P. (2002) Entry deterrence and quality provision in the local bus market. *Transport Reviews* **22**(3), 247-265.

Williams, H.C.W.L and J. Abdullal (1993) Public Transport Services under Market Arrangements. Part I: A Model of Competition between Independent Operators. *Transport Research B* **27B**(5), 369-387.

Williams, H.C.W.L and D. Martin (1993) Public Transport Services under Market Arrangements. Part II: A Model of Competition between Groups of Services. *Transport Research B* **27B**(5), 389-399.

2.11 Congestion pricing should affect the cost of capital

Congestion pricing should affect the cost of capital[46]

Harald Minken

Institute of Transport Economics

Contents

[46] This working paper was originally written in June 2008. Although I have not been able to confirm it, I have reason to believe that in its present form, it was included as an appendix to Deliverable 4 of the EU-funded project ENACT – Designing Appropriate Contractual Relationships. ENACT studied the simultaneous use of public private partnerships (PPP) and marginal cost pricing (MCP).

Recently, the toll rings in several of the urban regions of Norway have been developed further into more complex systems of rings, and the charges are differentiated by location of the ring or arm, time of day, mode (freight, public transport and private transport) and the vehicle's CO_2 emission. In the best of cases, this may very well be called marginal cost pricing. Together with funds from the national transport plan, the main part of the toll revenue (or nearly all of it, in Oslo's case) is allocated to public transport services, maintenance and investment. Thus, the conflict between marginal cost pricing and stable and reliable financing of projects is currently being felt in Norway, as marginal cost pricing is sacrificed to secure the infrastructure plans. The possibility of GPS-based road pricing in many countries in the not-too-distant future makes it even more important to discuss reasonable ways to handle such problems.

This is the reason why I have picked up the old note again. I realise that the paper in its present form lacks a suitable introductory section and a clearer conclusion. Even the derivation itself could be made clearer. It is my plan to produce a better version soon.

Oslo, January 1st, 2021. Harald Minken

ABSTRACT

Plans to fund a project by user charges are subject to systematic and unsystematic risk. Unsystematic risks include the risk of wrong projections of expected future demand, while systematic risk stems from the covariance of demand and the general economic climate. Theoretically, congestion pricing implies that the charges be changed if the level of congestion changes. It is shown that the cost of capital raised by such charges is much higher than if a fixed charge was applied. This might induce project owners to keep the charge fixed for long periods of time, or to apply congestion charging but transfer the risk to government.

Key words: Congestion pricing, cost of capital, risk
JEL classification: H230, H440

Congestion pricing should affect the cost of capital

Normally, transport project benefits are approximately proportional to demand. However, if the project involves congestion pricing, the major part of the benefits – the revenue – is proportional to some power of the demand – quite possible to demand to the power of 4 or 5. This makes revenue from marginal cost pricing appreciably more volatile than demand itself and will affect both unsystematic and systematic risk of the project. The high systematic risk raises the cost of capital in the project. The risk will however be the same regardless of whether the project is undertaken by government or the private sector.

1 Congestion pricing (deterministic case)

Suppose demand is $x = D(G)$, where G is generalised travel cost, and

(1) $$G = p + \omega t(x)$$

where p is the congestion charge and ω is the value of travel time. The travel time function $t(.)$ is increasing and convex. For simplicity, the marginal costs c of maintaining the infrastructure and operating the charging system are assumed to be constant, and environmental costs and accident costs are ignored. The welfare function W can be written

(2) $$W = \int_G^\infty D(y)\,dy + (p - c)D(G)$$

The first term of (2) is the user benefits, and the second term is the net revenue to the operator. Maximising W subject to (1) yields the optimal congestion charge p^*:

(3) $$p^* = c + \omega x t'(x)$$

Inserting p^* into the net revenue $\pi = (p - c)x$ yields the net revenue from optimal charging:

(4) $$\pi^* = \omega x^2 t'(x)$$

2 Unsystematic revenue risk

Suppose a private operator is considering to bid for a contract to build and operate some congestible infrastructure that is going to be financed by congestion charges. Little is known about future demand. Making wrong predictions in this case is a form of unsystematic risk, since prediction errors will hardly correlate with future variations of his other sources of income.

We consider two forms of uncertainty about the prediction of future demand: (1) uncertainty about the basic level of demand and (2) uncertainty about demand elasticity. In the first case, the operator assumes demand to be of the form $x = \varepsilon d\left(G\right)$, where the operator thinks he knows $d(G)$ but is uncertain about ε and want to assess the consequences of getting it wrong. In the second case, he assumes demand is $x = ae^{-\lambda G}$, and that there is uncertainty about the consequences of getting λ wrong.

Differentiating (4) with respect to ε and λ, respectively, we reach the following results for the two cases:

(5) $$\frac{d\pi^{\varepsilon}}{\pi^{\varepsilon}} = \left(2 + El_x t'\left(x\right)\right)\frac{d\varepsilon}{\varepsilon}$$

(6) $$\frac{d\pi^{\lambda}}{\pi^{\lambda}} = \left(2 + El_x t'\left(x\right)\right)\cdot El_G x \cdot \frac{d\lambda}{\lambda}$$

where the superscripts on π denote the uncertain factor in each case.

As can be seen, an error in the prediction of ε of 10 per cent, say, translates into a much larger error for the predicted revenue. In fact, if the most commonly used travel time function, $t = t_0\left(1 + \alpha x^{\gamma}\right)$, is used, the elasticity of the first derivative of the travel time function is $\gamma - 1$. If $\gamma = 4$, as is often assumed, the error in the estimation of demand translates into a fivefold bigger error (in percentage terms) in the estimation of revenue. A similar, but probably slightly less dramatic result is reached for λ.

Presumably, the operator of a project will not choose congestion pricing of his own will. Congestion pricing will have to be imposed as a form of regulation. Now, compare the two congestion pricing cases to the corresponding cases when there is no congestion pricing, but price cap regulation. With price cap regulation, the operator is not free to increase prices. Assuming it is not in his best interest to reduce prices either, we may identify this with p being exogenous. Thus net revenue is simply $\pi^{PC} = \left(p - c\right)x$. Differentiating, we get.

$$\frac{d\pi^{PC,\varepsilon}}{\pi^{PC,\varepsilon}} = \frac{d\varepsilon}{\varepsilon}$$

$$\frac{d\pi^{PC,\lambda}}{\pi^{PC,\lambda}} = El_G x \cdot \frac{d\lambda}{\lambda}$$

Thus in the price cap cases, an error in the estimate of the demand parameters translates into an erronous prediction of revenue of the same size in percentage terms. Obviously, investors

should be advised that marginal cost pricing amplifies the consequences of wrong demand estimates in a way that price caps do not do. The likely consequence of this is that they will demand a higher return on their capital than in the price cap case.

It must be pointed out that there is nothing in the above argument that does not apply if the investor and operator is government itself. The basic reason for the result is the explicit and implicit power of *x* in the net revenue formula (4). Thus, even if the government is the investor and operator of the infrastructure, there is every reason to avoid prediction errors and to be on guard against the possible volatility of congestion charging revenues.

What is likely to happen in practice is that the marginal cost pricing principle is sacrificed for the sake of a stable revenue. A charge once chosen will not be easily adjusted downwards if demand (and congestion) turns out to be less than expected, and will probably not be allowed to be increased if demand is higher than expected. The two objectives, congestion charging as a means of financing infrastructure and congestion charging as a means of improving the efficiency of the transport system, will have to fight it out. Quite possibly, the outcome will be given in advance in the form of guarantees to the operator and the investors that the congestion charging principle will be modified if revenues fall short of expectations.

3 Systematic risk

The relevant risk to any rational investor is not the variability of the returns of the investment itself, but the variability of all his assets and sources of income taken together. Thus when a new investment opportunity is considered, it is the covariance between the returns on the new project and his existing assets that matters.

Suppose he considers investing in an infrastructure project financed by congestion charging. Demand is a stochastic variable *X*, and the returns on the investment will be $\Pi = g(X) = \omega X^2 t'(X)$. His other assets are a stochastic variable *Y*. (If the investor is government, *Y* ought to be national income. If all national income stems from tradable assets, *Y* can be set equal to returns on the market portfolio, R_m.)

Being a rational investor, our man wants to compute $\operatorname{cov}(g(X),Y)$. Assuming both *X* and *Y* to be normally distributed, Stein's lemma applies. It says

$$\operatorname{cov}(g(X),Y) = E(g'(X)) \cdot \operatorname{cov}(X,Y)$$

Consequently,

(7) $$\operatorname{cov}(\Pi,Y) = \omega\left[2E(Xt'(X)) + E(X^2 t''(X))\right]\operatorname{cov}(X,Y)$$

Assuming once more that $t = t_0(1+\alpha x^\gamma)$, equation (7) simplifies to

(8) $$\operatorname{cov}(\Pi,Y) = (\gamma+1)\omega E(Xt'(X))\operatorname{cov}(X,Y)$$

The corresponding equation for the price cap case is simply

(9) $$\operatorname{cov}(\Pi^{PC},Y) = (p-c)\operatorname{cov}(X,Y)$$

Let us compare (8) and (9). The term $\omega E\left(Xt'\left(X\right)\right)$ in (8) is equal to $p - c$ with p set to the expected optimal congestion charge. Thus if the price cap p is in the same range, the relevant risk in the congestion charge case is once again some 5 times larger than the price cap risk (assuming $\gamma = 4$). The proportion between the beta's for the two projects will be the same, since the CAPM beta is just a normalisation of the covariances of (8) and (9), with $Y = R_m$.

Alexander et al (1996) found that companies under price cap regulation seem to be exposed to much higher levels of systematic risk (much higher beta's) than companies under rate-of-return regulation and other low-powered incentive schemes, and that the likely consequence of that is that they will face a higher cost of capital. Our finding here implies that companies under the regulatory regime implied by congestion pricing will be exposed to even higher levels of systematic risk than those under price cap regulation. The likely outcome of that is costly capital. Once more we will have to point out that this is the case regardless of whether the operator is a public or a private company, and regardless of whether the sources of capital are private or public funds. The systematic risk is in no way diminished by allocating seemingly cheap government money to the project or by using an artificially low discount rate in the calculation of the net present value of the toll revenue. It may be transferred to government by issuing guarantees to the operator or investors. But the only way to reduce it is to sacrifice the congestion charging principle itself.

The efficiency gains of congestion charging are obvious even if we account fully for the cost of capital invested in projects that depend on congestion charging revenue for their implementation. The difficult task will be to avoid sacrificing efficiency for the sake of cheaper capital.

Reference

Alexander, I., C. Mayer and H. Weeds (1996) Regulatory structure and risk and infrastructure firms. Policy Research Working Paper 1698, World Bank.

2.12 Assessing the benefit cost ratio of a marginal increase in the maintenance budget

Assessing the benefit cost ratio of a marginal increase in the maintenance budget[47]

Harald Minken

Institute of Transport Economics

Contents

[47] This is an unpublished paper from around 2016, based on TOI reports 1380/2014 and 1460/2015.

ABSTRACT

We propose a method to assess the benefit cost ratio of a marginal increase in the maintenance budget. It may be applied to any collection of facilities that are to be maintained within the same budget. The purpose is to assess whether the long-term budget level should be increased, and to compare the economic value of doing so with alternative uses of the money. In addition, the method makes it possible to identify inefficiency in the allocation of the budget to the facilities or the broad classes of facilities that it covers. To achieve this with a minimum of data and knowledge about the conditions and deterioration rates of the facilities, the model is kept as simple as possible. Existing knowledge about optimal rehabilitation policies is utilised. There is no binding budget constraint for each year, only a binding long term budget level. The model consists of the Kuhn-Tucker conditions of a non-linear programming problem. It may be programmed and easily solved using a spreadsheet. The output is a set of optimal rehabilitation frequencies for each class of facilities, and a non-negative Lagrangian multiplier that is identified as the benefit cost ratio of a marginal increase in the maintenance budget.

1 Introduction

In many parts of the public sector, decision makers face a difficult choice between allocating scarce funds to new infrastructure projects on the one side or to maintenance and rehabilitation of existing infrastructure on the other. Strategic transport plans at the national, regional or urban level is a case in point. Although the main political interest is often on the investment side, a considerable part of the available funds must also be set aside to secure the continued functionality of existing facilities. Thus the plan will consist of an investment budget and a maintenance budget.[48] How should we determine the share of each of them? The choice would be easier if it was possible to make a direct comparison of the economic value of a marginal increase in the maintenance budget with a marginal increase in the investment budget. To make the comparison, one would obviously have to make assumptions on how the extra funds will be used. We assume that in both cases, actions, plans and projects are chosen to maximise economic efficiency. This is certainly not the most realistic of assumptions, but it makes a useful distinction between the question of efficient use of the existing maintenance budget and the question of the economic value of a budget increase.

On the investment side, the economic efficiency of a plan is maximised if projects are chosen according to their benefit cost ratio. The marginal value of a budget increase or reduction is given by the benefit cost ratio of the marginal project. In this article, we propose an easy method to assess the marginal economic value of funds allocated to maintenance. The two measures may be directly compared. Although it is perhaps easiest to establish the necessary data for road applications, our model is generic and apply to any collection of objects the use of which generates costs for the users and the agency in charge, and that are to be maintained and periodically rehabilitated under a budget constraint. They may be single facilities, each with their own characteristics, or facilities may be lumped together and be represented as single objects if they share the same relevant characteristics.

[48] In our usage, the maintenance budget covers both ordinary annual maintenance and periodically recurring rehabilitation. The rehabilitation frequencies are our policy variables, while the funds needed for ordinary maintenance are supposed to follow from that.

Under normal circumstances, the deep layers of most transport infrastructure facilities are not subject to noticeable degradation by use, and so do not need to be replaced in full. The upper layers of the construction, however, are subject to degradation and need to be renewed repeatedly. Not doing so means extra costs for the users in the form of decreased speed, accelerated vehicle wear and tear and increasing risk of accidents. Ultimately, neglected maintenance and postponed rehabilitation will also affect the fundamental parts of the construction, leading to even larger costs and inconveniences for users and to agency cost increases in the form of unplanned corrective maintenance and repair.

Rehabilitation is however expensive. While too long intervals between rehabilitations impose unnecessary high costs on users as well as the agency in charge, so will too short intervals. Consequently, depending on the size of the costs in each case, there must be an economically optimal rehabilitation cycle. Our task is to find the optimal rehabilitation cycles for many facilities that are to be rehabilitated under a common budget constraint. Both deterioration rates and the effect of rehabilitation are assumed to be deterministic, but these relationships and other factors that affect user costs and agency costs may be specific for each facility. Thus, in principle, our model will be a deterministic system-level bottom-up maintenance and rehabilitation (M&R) model with a budget constraint. For simplicity, we assume that preventive maintenance is carried out according to a fixed plan in any circumstance, while corrective maintenance costs and user costs, both separately and as a whole, increase in a known way with time from the last rehabilitation. We also assume that the level of use is constant, and that there is no technological change or other factors that affect deterioration rates or costs. The budget constraint is a long-term average level that may be violated in the short run. The time horizon is infinity. The initial condition of all facilities is assumed to be the best possible, and this best condition is recreated at each rehabilitation.[49] Thus the only decision that matters is the interval between rehabilitations.

The current article extends existing know-how in the following ways: It provides an easy way of comparing the economic efficiency of increasing the maintenance budget with an increase in the investment budget. By way of an example, it points out a possibly general relationship between the size of user costs relative to the agency costs on one hand, and the possibility of adaption to the budget constraint by prolonging the interval between rehabilitations on the other. In contrast to Sathaye and Madanat (2012), for instance, it applies discounting within a rehabilitation interval, thereby treating facilities with different optimal rehabilitation intervals more fairly. It allows for a large number of facilities without the need for progressively more complex solution methods, and so it may possibly be extended to include explicit optimisation of preventive maintenance as well as rehabilitation in the future.

The model is not intended as an actual planning tool. Thus there is no mention of need for inspection of the state of the objects, the need for revision and updating of the plan, etc. While it provides clues to the optimal rehabilitation intervals for the facilities, it is first and foremost an aid to decisions about the size of the maintenance budget part of a transport plan, either in absolute terms or relative to the investment budget. A budgeting tool, not a planning tool. It turns out, however, that these decisions are as dependent on good data on the state and

[49] Or rather the best condition that can be achieved in an efficient way.

degradation of the objects, the size of the user costs and the efficiency of rehabilitation policies as any actual planning tool.

The next section provides a short review of the relevant literature. It is followed by the main section, where the non-linear programming problem is formulated and solved and the economic value of a marginal budget increase is found. A hypothetical example is provided, and a few comments are made on how the model may be extended. The final section concludes.

2 Literature review

Economic analysis of maintenance and rehabilitation of transport infrastructure involves a set of facilities and a description of the possible states that each facility may be in (the state spaces), in addition to descriptions of how facilities deteriorate with time and use, and of the user costs associated with the states of each facility, the actions (or policies or strategies) open to the agency in charge, the effect of each action on the state in each of the states, and the cost to the agency of each of the possible actions. The objective of the analysis is usually to find strategies (actions over a certain time period) that minimise the sum of user costs and agency costs over this period. Strategies may be constrained by the agency's annual budget or other requirements.

Since the advent of the private car on a mass scale, applications to road pavements has constituted the largest and arguably richest part of this field of research. Recent reviews of different strands of the road pavement literature can be found in Lee and Madanat (2015), Gao (2011) and Durango-Cohen and Madanat (2008). A classical work in the field is Small et al (1989). Their analysis trades off the costs of construction and rehabilitation (resurfacing) to derive cost efficient standards for the construction and operation of American highways. Paterson (1987, 1990) are classical references to road pavement deterioration and the effect of actions. They are still much used in academic work today, as for instance Ouyang and Madanat (2004), Lee and Madanat (2014, 2015) and Sathaye and Madanat (2012).

The reason why we refer to the pavement deterioration literature here is because even if our model does not explicitly mention it (our cost functions are monetary expressions of the dynamics of each facility in reduced form), deterioration and the effect of actions must be addressed whenever the model is applied.

Broadly, the modern pavement literature divides into single facility and system-level approaches. The latter can further be sub-divided into top-down and bottom-up approaches. Top-down means that the facilities are homogeneous with respect to deterioration and cost function parameters. Such models cannot be used for actual maintenance planning, but may provide general guidance for high-level decisions that can be a starting point for such planning. Or they may include two levels, one in which the funds are allocated to local districts, an one in which the local agencies are assumed to plan their activities subject to their budgets. A simple top-down model of this type is Tsunokawa and Hiep (2008). Increasingly, however, the norm now is bottom-up, where each facility is modelled individually, with its own deterioration function and other specific traits.

Our interest is primarily in system-level models with budget constraints, but some very useful results from the single-facility models of Li and Madanat (2002) and Ouyang and Madanat (2006) carry over to a broader context. In particular, they show that the optimal steady state policy is to rehabilitate whenever the facility reaches a certain threshold state. At each rehabilitation, the facility should be improved up to its best achievable state. The optimal thresholds were robust with respect to uncertainty in the deterioration model. These results were proved for continuous time and state space, deterministic deterioration with the markov property, and deterministic effects of the chosen policy. Rehabilitation was supposed to be the only policy instrument available.

The results were subsequently utilised to solve multi-facility models such as Sathaye and Madanat (2011, 2012) and Lee and Madanat (2014, 2015). Together with efficient heuristic solution algorithms, the adoption of these results to solve large models is part of the explanation why there has been a switch to large bottom-up models, as well as a shift from stochastic to deterministic models. These results will apply and be used in the present model as well.

Models may have more than one state variable, and more than one policy variable. For instance, one state variable may reflect ridership quality and determine the user costs, while another reflects the strength of the construction and determines the rate of deterioration. Then maintenance and rehabilitation are geared toward improving ridership quality and reconstruction is used to improve the strength. Alternatively, preventive maintenance may also be used to improve ridership quality and slow down deterioration. In Gu et al (2012), rehabilitation and maintenance are the two policy options. In Lee and Madanat (2015), the integrated modelling of construction, reconstruction, rehabilitation and maintenance seems to bring us full circle back to the problems that occupied Small et al (1989). Finally, discrete bottom-up multi-facility and multi-policy models with budget constraints include Karabakal et al (1994), Dahl and Minken (2008), and Gao (2011).

Railway maintenance and rehabilitation problems are in some respects more difficult that its road counterparts, because the track consists of many different but equally important components, some of which may affect user comfort, others affect travel time reliability, and still others do not affect travellers at all until they break down. The components are all interrelated in the sense that normal traffic requires all components to function as intended (Mattsson 2007), and more often than not, there are economies of scope in maintenance and rehabilitation (Furuya and Madanat 2012).

3 The model

Consider a set of J infrastructure objects, indexed by j. Every object has its own construction cost, rehabilitation cost, user cost and annual agency cost. Thus optimal frequency between rehabilitations will be object specific. We assume that the objects can live infinitely if it is not too long between rehabilitations.

The cost of constructing object j, as measured in fixed prices, is C_{0j}. Later on, it will have to be rehabilitated with a fixed frequency every n_j-th year. Time is measured from the year of construction, so the first rehabilitation will take place in year n_j. At each full rehabilitation, the

time is reset. The rehabilitation cost $R_j(n_j)$ is an increasing function of the length of time between rehabilitations. It is reasonable to assume $R_j(0) = 0$ and the derivative $R_j'(n_j) > 0$.

The users of object j derive a fixed annual utility E_j from the use. In addition, they have annual costs $C_{Bj}(t)$ that depend on time t since the last rehabilitation. Time t goes from 0 to n_j and $C_{Bj}'(t) > 0$. The infrastructure owner (the agency) have annual costs $C_{Ej}(t)$ of the same form.

Table 1 summarises the notation.

Table 1 Notation

Investment cost	C_{0j}
Time dependent annual user costs	$C_{Bj}(t)$
Time dependent annual agency costs	$C_{Ej}(t)$
Fixed user benefit	E_j
Number of years between rehabilitations	n_j
Rehabilitation cost	$R_j(n_j)$
Discount rate	ρ

The discount rate ρ is used to discount benefits and costs of a rehabilitation period to the start of the period. The net benefit of object j over a complete rehabilitation period is denoted $V_j(n_j)$. We have:

$$V_j\left(n_j\right) = \int_0^{n_j} \left\{E_j - \left(C_{Bj}\left(t\right) + C_{Ej}\left(t\right)\right)\right\} e^{-\rho t} dt - R_j\left(n_j\right) e^{-\rho n_j} \tag{1}$$

The net benefit of the object over an infinitely long lifetime, $W_j(n_j)$, is the discounted net benefit of an infinitely repeated chain of rehabilitation periods, minus the initial investment C_{0j}. To find an expression for $W_j(n_j)$, we first note that multiplying $V_j(n_j)$ with the discount factor $\exp\left(-\rho n_j\right)$ transforms $V_j(n_j)$ in any of the rehabilitation periods to its net present value in the period before. Except for C_{0j}, $W_j(n_j)$ is an infinite sum of $V_j(n_j)$ terms. The first is already transformed to the net present value at the time of construction, the second is transformed to its net present value at the time of construction by multiplying with the discount factor once, the third by multiplying with the discount factor twice, and so on to infinity. As our discount factor is strictly less than 1, a well known formula for convergent geometrical series can be applied. The result is that in the case at hand,

$W_j\left(n_j\right) = -C_{0j} + V_j\left(n_j\right) \cdot \left(1 - e^{-\rho n_j}\right)^{-1}$. We define $W\left(\mathbf{n}\right)$ as the sum of $W_j(n_j)$ over all j.

Written out in detail and rearranging, then summing over all objects:

$$W\left(\mathbf{n}\right) = \sum_{j=1}^{J} \left(\left(\frac{E_j}{\rho} - C_{0j} \right) - \left(1 - e^{-\rho nj}\right)^{-1} \left\{ \int_0^{n_j} \left(C_{Bj}\left(t\right) + C_{Ej}\left(t\right)\right) e^{-\rho t} dt + R_j\left(n_j\right) e^{-\rho n_j} \right\} \right) \tag{2}$$

The sum of agency costs and rehabilitation costs for all J objects have to keep within a given long-term budget constraint of the form B/ρ, where B is the long-term average annual budget:

$$\sum_{j=1}^{J} \frac{1}{1 - e^{-\rho n_j}} \left\{ \int_0^{n_j} C_{Ej}\left(t\right) e^{-\rho t} dt + R_j\left(n_j\right) e^{-\rho n_j} \right\} \leq \frac{B}{\rho} \tag{3}$$

Our objective is to maximise $W(\mathbf{n})$ subject to the budget constraint. This is a non-linear programming problem with non-negative variables. To solve it, we form the Lagrangian function $L(\mathbf{n}, \mu)$ with the Lagrangian multiplier μ, and set the J partial derivatives of it equal to 0.[50] This produces J equations, which together with the requirement that $\mu \geq 0$ and that if the budget is not binding, $\mu = 0$, constitutes the Kuhn-Tucker conditions for maximum. In this case, each of the J equations involve μ and one and only one of the n_j. Details of the computation are given in Appendix 1. The form of any of the J Kuhn-Tucker conditions are given in Equation (4).

(4)

$$G_j(n_j) - H_j(n_j) = \frac{\mu}{1+\mu}\left[\frac{\rho}{1-e^{-\rho n_j}}\int_0^{n_j} C_{Bj}(t)e^{-\rho t}dt - C_B(n_j)\right]$$

where

$$G_j(n_j) = \frac{\rho}{1-e^{-\rho n_j}}\left[\int_0^{n_j}\left(C_{Bj}(t)+C_{Ej}(t)\right)e^{-\rho t}dt + R_j(n_j)\right]$$

$$H_j(n_j) = C_{B_j}(n_j) + C_{E_j}(n_j) + R_j'(n_j)$$

To understand this equation, note that for each object in the model, there is only one decision to be made. That decision is when to initiate the first rehabilitation. Per definition, all later rehabilitations will follow from this. At any time in the first rehabilitation period, the question is: Should we rehabilitate now, or should we postpone it for one more year? $G_j(n_j)$ is the infinite annuity we get if we decide to rehabilitate, while $H_j(n_j)$ is the cost of carrying on for one more year, including the marginal increase in the rehabilitation cost. If the annuity is larger than the present annual cost, we postpone the decision by one more year, but it the reverse is true, we are already a little late.

Together with the budget constraint, the system of J equations of the form (4) makes it possible to compute the optimal rehabilitation frequencies n_j^* and the optimal Lagrangian multiplier μ^*. The algorithm for doing so is particularly simple, and may even be programmed in a spreadsheet:

Start by choosing $\mu = 0$, then compute each of the rehabilitation frequencies given this μ. (We assume the integrals are solvable.) Check if the budget constraint is satisfied in the solution by adding together the discounted agency costs of all objects. If it is, the solution has been found. If not, choose a somewhat larger μ, say 2, compute the rehabilitation frequencies anew and check the budget constraint. If the constraint is not satisfied, increase μ further until it is. Having found a μ that satisfies the constraint, the final task is to find out how much μ can be

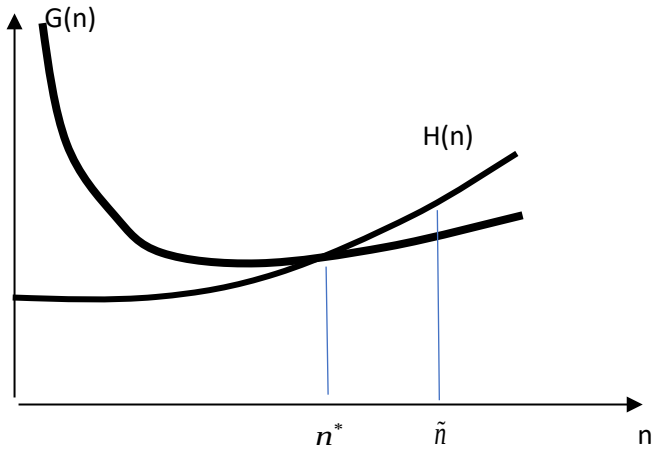
[50] In the general case, all we can say is that the J partial derivatives of $L(\mathbf{n}, \mu)$ are all non-positive (Sydsæter et al 2005). But in this case we may safely assume that no n_j is zero, which implies that strict equality applies.

reduced while still satisfying the constraint. This can be done by trial and error or with a search algorithm such as the golden section.

Figure 1 illustrates how the two curves $G_j(n_j)$ and $H_j(n_j)$ from Equation (4) determine the optimal rehabilitation frequency for object *j*. The form of the curves is derived in Appendix 2. As can be seen from Equation (4), the point n^* where the curves cross is the solution if $\mu = 0$ in optimum, i.e. if the budget constraint is not binding. As shown in Appendix 2, if $\mu > 0$, the right hand side of Equation (4) must be strictly negative. That means that the $H_j(n_j)$ curve lies above the $G_j(n_j)$ curve, so solutions with a binding budget constraint all lie to the right of the crossing point in the figure. One such solution is $\tilde{n}$ in the figure. The larger μ is, the farther to the right lies the solution. Points to the left do not represent anything – neither a solution nor points were there is no solution. This is because the Lagrangian multiplier is non-negative by the formulation of the programming problem.

The important but rather obvious consequence of this is that what the binding budget constraint does, is to lengthen the period between rehabilitations. This does not mean that the infrastructure facilities will have to deteriorate in the long run – it is perfectly possible for a stable long run solution to exist even with fairly long time between rehabilitations. But the longer between rehabilitations, the higher will the average costs be, in particular the user costs. On the other hand, it is also possible that no solution exists and that the facility deteriorates beyond repair. This is seen if we note that the factor $\mu(1+\mu)^{-1}$ must be positive but below 1. Thus there is plenty of room for the difference between $H_j(n_j)$ and $G_j(n_j)$ to

Figure 1. Determining the optimal rehabilitation frequency of an object in the unconstrained and a constrained case.



outgrow the left hand side of Equation (4) for some specifications of the cost functions and the budget. Points with no solution will lie to the far right.

In principle, then, there are three possible situations: An unconstrained solution, a constrained solution and no solution.

3.1 The interpretation of μ*

The Lagrangian multiplier at optimum, μ^*, can be interpreted as the increase in the value function (the objective function at optimum) by a marginal increase of the associated constraint. This is proved by use of the envelope theorem, see Sydsæter et.al. (1991). In our case, the multiplier represents the increase in the value function $W\left(\mathbf{n}^*\right)$ by a marginal increase in B/ρ. What we want to find is however not the effect of a marginal change in the net present value of keeping the budget level B forever, B/ρ, but the effect of a marginal change in the annual budget B itself. The calculation goes like this:

(5) $$\frac{\partial W^*}{\partial B} = \frac{\partial W^*}{\partial\left(B/\rho\right)} \cdot \frac{\partial\left(B/\rho\right)}{\partial B} = \frac{\partial W^*}{\partial\left(B/\rho\right)} \cdot \frac{1}{\rho} = \mu^* \frac{1}{\rho}$$

In the case of chosing (independent) investment projects to a plan under a budget constraint, the solution is to rank the projects with positive net present value by their benefit cost ratio and include them in the plan in descending order until the budget is exhausted. In that case, the marginal increase in the net present value by increasing the budget to include the best of the excluded projects is measured by the benefit cost ratio of that project. Thus there is a strong correspondence between the cost benefit ratio in the investment case and the optimal Lagrangian multiplier in the case of a marginal increase in the maintenance budget. A cost benefit analysis of the latter case would go like this:

Gross benefit: $\mu^* \cdot \rho^{-1}$ (see equation (5))

Cost of the policy: one dollar forever, or $1 \cdot \rho^{-1} = \cdot \rho^{-1}$

Benefit cost ratio: gross benefits divided by cost, μ^*

Net present value: $\mu^* \cdot \rho^{-1} - \rho^{-1} = \left(\mu^* - 1\right)\rho^{-1}$

Thus $\mu^* > 1$ (or $\mu\left(1+\mu\right)^{-1} > \frac{1}{2}$) would be required for a maintenance budget increase to be worthwhile. But if marginal investment projects have benefit cost ratios above 1, and if maintenance and investment compete for funds, a μ^* above the BCR of the marginal investment projects would be required. (Taking the rough character of our method into consideration, one might need a rather large difference between μ^* and BCR before a clear decision can be made.)

4 Example

The model was applied to rehabilitation of the track of 20 Norwegian railway lines. Ten of them are retained in this test example.

The agency costs considered consisted of preventive and corrective maintenance costs plus rehabilitation costs. In work by an agency within the National Railway Administration it was shown that the annual increase in each of the two types of maintenance costs were best explained by linear regression models. The rehabilitation costs were found not to depend on

time. These estimates were distributed by kilometres of track to the ten railway lines and used for the illustrative example.

The estimates did not however take into account that the activity level depends on the politically determined budget, and not only on annual wear and tear. Thus it is probable that efforts to reduce the backlog, both in the form of rehabilitation work and more preventive maintenance, have led to misspecification of these cost models (Ben-Akiva and Ramaswamy 1993, Chu and Durango-Cohen 2008). Remember that in our model, backlogs are assumed to have been eliminated at time zero.

User costs were computed from statistics on late arrivals. A certain part of the late arrivals where reported to be caused by problems with the track. The user cost was computed by multiplying the extra train-hours with the average number of passengers and the official value of time. A statistical analysis showed that an exponential function explained the development of user costs with time. Again, time in this case was historical time, not time since the last major rehabilitation. Also, the user costs were so narrowly defined that they where smaller than the agency costs by a factor of 5.

These flaws turned out to illustrate important features of the model. As Table 2 shows, unless B was very close to the annuity of agency costs (including the rehabilitation costs), it turned out to be impossible to equate the two by adjusting μ. There were simply not enough flexibility in the cost structure to use rehabilitation frequencies as instruments to achieve sustainable low-cost maintenance strategies. Why did the model fail to be of use in this case?

Remember that the right hand side of equation (4) contains only user costs. This is a sign that whenever the budget is binding, efforts to reduce the agency costs can only be achieved by increasing the user costs. Therefore, the size of the interval of possible budgets on which a solution is possible depends on the size of the user costs compared to the agency costs, and on how quickly each of them develop with time. In Table 3, where we have increased the user costs by a factor of 10 and left everything else as before, the interval of possible budgets is much larger, as the agency costs are much more sensitive to a change in μ.

Table 2 The railway example. Original user costs (MNOK).

Average annual cost	B	μ	Optimal rehabilitation frequency			- W^*
			Mean	Min	Max	
97,0	97,0	0	62,6	34,8	95,0	2810
96,5	96,5	1,4	65,0	36,7	98.4	2816
96,4	96,4	2,8	65,8	37,3	99,4	2820
96,35	96,35	5,6	66,5	37,8	100,2	2825
96,32	96,32	1000	67,7	38,4	101,3	2825

* The constant part of W is ignored

B = 97,0 is the borderline between unconstrained cases and cases where the budget is binding, while B = 96,32 is the smallest possible long-term equilibrium budget. Budgets lower that that will inevitably lead to periodically recurring maintenance backlogs if the functional relationships of the model are to be trusted. The optimal rehabilitation frequencies are perhaps a little too high to be fully confident of that.

The reduction of the budget from the point B = 97 to B = 96,32 is just 0,7 per cent, which shows that there is very little scope for adapting to the budget by prolonging the rehabilitation frequency. At the original user costs, the model is in fact of limited interest. At even lower user costs, as in the case of bridge rehabilitation, the only practical options are to hit the correct rehabilitation frequency or to incur regularly recurring backlogs.

Table 3 The railway example. User costs multiplied by 10 (MNOK).

Average annual cost	B	μ	Optimal rehabilitation frequency			- W
			Mean	Min	Max	
107,8	107,8	0	48,8	22,7	73,7	5871
100,7	100,7	1,4	55,4	28,4	83,2	5961
98,8	98,8	2,8	58,2	30,9	87,9	6055
97,5	97,5	5,6	60,9	33,4	92,5	6181
96,32	96,32	1000	67,58	38,4	101,2	6673

In Table 3, the interval of long-term sustainable budget levels is larger (10.6 per cent). In Tables 4 and 5, we have taken away the annual growth of preventive maintenance. As expected, this produces longer rehabilitation frequencies. (In this case, they obviously become too large.) The interval between the budgets of μ = 0 and μ = 1000 is 2,3 per cent in Table 4 and 30 per cent in Table 5. Thus the relative impact of a tenfold increase in user costs is about the same regardless of the annual size of the growth of preventive maintenance costs, but taking away maintenance cost growth also shifts the absolute impact by a factor of about 3.

Table 4 The railway example. No growth in preventive maintenance, original user costs (MNOK).

Average annual cost	B	μ	Optimal rehabilitation frequency			-W
			Mean	Min	Max	
45,22	45,22	0	133,5	64,5	208,4	1577
44,1	44,1	1,4	156,6	81,3	234,6	1591
43,83	43,83	2,8	168,3	89,9	247,8	1603
43,7	43,7	5,6	181,8	99,6	263	1619
43,5	43,5	1000	263,8	132,5	391,7	1678

Table 5 The railway example. No growth in preventive maintenance. User cost ten times original (MNOK).

Average annual cost	B	μ	Optimal rehabilitation frequency			-W
			Mean	Min	Max	
62,2	62,2	0	74,8	27,8	136	4884
51,7	51,7	1,41	96,2	39,7	163,9	5022
48,75	48,75	2,8	107,6	47,1	178,3	5171
46,3	46,3	5,6	122,5	56,7	195,5	5407
43,5	43,5	1000	235,8	128,1	325	6868

The final output of our model is μ^*, the shadow price of an increase in B/ρ. The example indicates that for a firm conclusion to be drawn on the profitability of increasing the long-term maintenance budget of a given agency, it is of the utmost importance that data are as accurate as possible. The model itself is simple, but the real work is to establish good data and functional relationships. In particular, the user costs need to be at least of similar size as the agency costs.

When μ is above 1000 and the model breaks down, there is no choice in the long run but to pay what it takes to keep the infrastructure in the best of shapes. Thus it is not only impossible to assess the profitability of increasing the maintenance budget, it is also unnecessary to do so.

The conclusion of the test example is that the reliability of the model depends very much on the quality of the data, and that the range of budget levels for which there is a long-term equilibrium maintenance solution depends crucially on the size of user costs relative to agency costs.

5 Extensions

5.1 More cost categories

Users also incur costs of rehabilitation, for instance because a stretch of road need to be closed down during the work, or because passing is only possible at reduced speed. This can be included if rehabilitation cost are split in one part that goes into the budget and one that does not. Likewise, the agency incurs costs in the form of corrective maintenance if the condition becomes bad enough. Expected costs of this type can be added to the agency's cost function.

5.2 Realistic initial conditions

The simplicity of the model is based on the assumption that results on the optimal rehabilitation policy from Li and Madanat (2002) and Ouyang and Madanat (2006) can be trusted to apply in our context. If so, it will not take too long before all facilities have reached the steady state that is assumed to exist in the model. The cost of bringing them there can be separated out from the costs in the model and estimated together with the backlog of rehabilitation activity that ideally should have taken place before the present time, but have been postponed.

Engineers tend to estimate this backlog on the assumption that facilities can be brought to the best condition all at the same time. While this leads to overstating the backlog, it fits with the assumption in our model that all facilities are in the best of conditions at time zero. Therefore, the backlog as estimated by engineers can be added to the optimal costs of the model as a first approximation to the total optimal cost of the whole set of facilities. Next, however, the starting time of the steady state of the facilities could be spread out evenly over a period of time of approximately the same length as the average interval between rehabilitations, before discounting the stream of backlog costs and infinite steady state costs back to the original starting point. This could be done using information about the last major rehabilitation of all of the facilities, or in a more approximate way.

Thus the assumption of a starting point where all facilities are in their best condition can be relaxed if necessary. We should note, however, that information about the total discounted cost of each facility or of the whole set of facilities are not necessary for the estimation of the intervals between rehabilitations or the benefit cost ratio of increasing the budget. These model outputs are not affected by the assumptions made about the condition of facilities at time zero.

5.3 Incorporating routine maintenance

Another questionable assumption in our model is that there is only one available policy instrument, rehabilitation. It is shown in for instance Labi and Sinha (2003, 2005), Lamptey et al (2008), Rashid and Tsunokawa (2012), Gu et al (2012) that total costs can be much reduced by applying relatively low-cost policies that slow down deterioration and thus lengthen the period between rehabilitations. The costs and effects of applying such instruments in an optimal or near-optimal way should therefore be taken account of when defining the annual agency costs and the deterioration function. If this is done, the need to define such low-cost policies as a separate policy instrument is at least reduced if not eliminated. Alternatively, a bi-level problem might be formulated, with optimisation of the preventive or routine maintenance inside each rehabilitation cycle as the lower level problem.

5.4 Other forms of inefficiency

Applying our model to find the benefit cost ratio of increasing the maintenance budget in steady state is only the final stage in an analysis of the possibility of increasing economic efficiency in rehabilitation. The first stage is to assess the cost of eliminating backlogs. It is rather meaningless to speak of the benefit of doing so: it is a necessity, not an option. Only then can one begin to speak of the optimal long-term policy and value of increasing the budget. But until it has been done, the backlog is a form of inefficiency.

The second stage is however inefficiency in the allocation of funds to the individual facilities. As long as the interval between rehabilitations is sub-optimal, there will be an efficiency gain to be had even without increasing the budget. Thus one should always check the optimal rehabilitation frequencies in the model output by comparing them to standard engineering advice. If there is a discrepancy, it might be that the costs need to be revised, or it might be that the model output is judged to be better from an economical point of view than the engineering advice. We have not gained enough experience with the model yet to say more about this here, but it is obviously something to keep an eye on to avoid completely wrong conclusions.

The third form of inefficiency is inefficiency in budget allocation, which has been the subject of this article. We may also talk about a fourth form, which however lies outside the scope of this article: It is inefficiency in the setting of standards in maintenance and rehabilitation. One aspect of this is covered in more complex models than this one, and that is to find an economically efficient upper threshold for the rehabilitation effort (the lower threshold follows from finding the optimal interval between rehabilitations). But there are many more aspects to be studied by cost benefit analysis, such as the design parameters of the infrastructure, and the supplementary constructions and services that is provided together with the basic services that the facility can offer. This will have to be the subject of future research.

6 Conclusion

A method has been devised to assess the economic value of transferring funds from the investment budget (or any other source) to the maintenance budget. The method utilises previous knowledge about the optimal steady state rehabilitation strategy, namely that a rehabilitation should be initiated whenever the facility reaches a threshold state, and that the maximum effective intensity should be applied in each case. The result is a model that is easy to solve even for a large number of facilities, and that produce a benefit cost ratio of increasing the maintenance budget and optimal steady-state rehabilitation cycles for each facility. Some of the simplifying assumptions behind the model are thought to be less important. For instance, the assumptions that a full rehabilitation to the best possible state is the only policy instrument available, may be justified if there are clear and near optimal rules for preventive maintenance and minor repairs, and if the effect of these policies on the state of the facilities is built into the user cost and agency cost. Other assumptions, such as the constant technology and constant levels of traffic, are less realistic. However, the model is not intended as a detailed long-term planning tool. Our example shows the importance of good data and solid estimates of costs for reliable results to be achieved, and that the scope for finding constrained solutions depends on the size of user cost compared to agency costs.

It is important to make it clear that the model does not treat the case where the level of maintenance and rehabilitation falls behind a long-term sustainable level. Thus it is assumed that such backlogs are eliminated before a constrained or unconstrained optimal long-term rehabilitation policy is established. This having been done, there may still be two forms of inefficiency in rehabilitation. The first is inefficiency in the allocation of funds to the individual facilities, which means that under the prevailing conditions, some facilities get rehabilitated too seldom and some too often. The second is inefficiency in the distribution of funds between investment and maintenance and rehabilitation. Both of them can be assessed with this model.

Acknowledgements

Even if this paper was never published, the author want to thank two anonymous referees for useful comments. Thanks also to Vegard Østli and Marius Fossen for help with the example.

Appendix 1

Derivation of Equation (4)

The Lagrangian is

$$L(\mathbf{n},\mu)=\sum_{j=1}^{J}\left(\left(\frac{N_j}{\rho}-C_{0j}\right)-\frac{1}{1-e^{-\rho nj}}\left\{\int_0^{n_j}C_{Bj}(t)e^{-\rho t}dt+(1+\mu)\left(\int_0^{n_j}C_{Ej}(t)e^{-\rho t}dt+R_j(n_j)e^{-\rho n_j}\right)\right\}\right)$$

where μ is the Lagrangian multiplier.

Differentiating $L(\mathbf{n},\mu)$ with respect to any n_j and setting the resulting expression equal to 0 produces one of the J first order conditions (they are all of the same form):

$$\frac{\partial L}{\partial n_j}=-\frac{1}{1-e^{-\rho n_j}}\left\{C_{B_j}(n_j)e^{-\rho n_j}+(1+\mu)\left(C_{E_j}(n_j)e^{-\rho n_j}+R_j'(n_j)e^{-\rho n_j}-\rho R_j(n_j)e^{-\rho n_j}\right)\right\}$$

$$+\frac{\rho e^{-\rho n_j}}{\left(1-e^{-\rho n_j}\right)^2}\left\{\int_0^{n_j}C_{Bj}(t)e^{-\rho t}dt+(1+\mu)\left(\int_0^{n_j}C_{Ej}(t)e^{-\rho t}dt+R_j(n_j)e^{-\rho n_j}\right)\right\}=0$$

Factoring out $e^{-\rho n_j}\left(1-e^{-\rho n_j}\right)^{-1}$ and rearranging:

$$\frac{\rho}{1-e^{-\rho n_j}}\left\{(1+\mu)\left\{\int_0^{n_j}\left(\left(C_{Bj}(t)+C_{Ej}(t)\right)e^{-\rho t}dt+R_j(n_j)e^{-\rho n_j}\right)\right\}-\mu\int_0^{n_j}C_{Bj}(t)e^{-\rho t}dt\right\}$$

$$-(1+\mu)\left(C_{B_j}(n_j)+C_{E_j}(n_j)+R_j'(n_j)\right)+\mu C_{B_j}(n_j)+(1+\mu)\rho R_j(n_j)=0$$

Now collect the two R_j terms together. The result is an almost unnoticeable change in the first line, where $R_j(n_j)e^{-\rho n_j}$ becomes just $R_j(n_j)$, while the $R_j(n_j)$ term in the second line vanishes:

$$(1+\mu)\frac{\rho}{1-e^{-\rho n_j}}\left[\int_0^{n_j}\left(C_{Bj}(t)+C_{Ej}(t)\right)e^{-\rho t}dt+R_j(n_j)\right]-\mu\int_0^{n_j}\frac{\rho}{1-e^{-\rho n_j}}C_{Bj}(t)e^{-\rho t}dt$$

$$-(1+\mu)\left(C_{B_j}(n_j)+C_{E_j}(n_j)+R_j'(n_j)\right)+\mu C_B(n_j)=0$$

Keep the (1 + μ) terms on the left hand side and move the μ terms to the right:

$$(1+\mu)\frac{\rho}{1-e^{-\rho n_j}}\left[\int_0^{n_j}\left(C_{Bj}(t)+C_{Ej}(t)\right)e^{-\rho t}dt+R_j(n_j)\right]$$

$$-(1+\mu)\left[C_{B_j}(n_j)+C_{E_j}(n_j)+R_j'(n_j)\right]=\mu\left[\frac{\rho}{1-e^{-\rho n_j}}\int_0^{n_j}C_{Bj}(t)e^{-\rho t}dt-C_B(n_j)\right]$$

Dividing by $(1+\mu)$ on both sides brings us to equation (4) in the text.

Appendix 2

The figure

The parenthesis on the right hand side of Eq. 4 is negative as long as $C_{Bj}(t)$ is monotonously increasing in t. This is because with $C_{Bj}(t)$ increasing in t, $C_{Bj}(n_j)$ is larger than $C_{Bj}(t)$ for all t (except for $t = n_j$, at which point it is equal to $C_{Bj}(n_j)$, of course). Thus we have the following inequality:

$$\int_0^{n_j}C_{Bj}(t)e^{-\rho t}dt<C_{Bj}(n_j)\int_0^{n_j}e^{-\rho t}dt$$

Solving the integral:

$$C_{Bj}(n_j)\int_0^{n_j}e^{-\rho t}dt=\frac{1-e^{-\rho n_j}}{\rho}C_{Bj}(n_j)$$

which means that the parenthesis on the right hand side of Eq. (4) is negative. The whole right hand side consists of this parenthesis multiplied with the non-negative Lagrangian term $\mu(1+\mu)^{-1}$. Thus it is negative if the budget constraint is binding and zero if not.

The *left hand side* of Eq. (4) is $G_j(n_j) - H_j(n_j)$. In terms of a figure of the two curves $G_j(n_j)$ and $H_j(n_j)$ it means that the two curves cross at a single point, namely at the n_j that is the solution to the unconstrained problem, i.e. the point that solves the problem if $\mu = 0$. To the right of this point, the $H_j(n_j)$ curve lies above the $G_j(n_j)$ curve. Here we find the n_j's that solve the problem if the constraint is binding.

With continuous cost functions, $\lim_{n_j\to 0}H_j(n_j)=C_{B_j}(0)+C_{E_j}(0)+R_j'(0)$. Since costs obviously are not zero, not even immediately after a rehabilitation, the H_j curve starts a bit up on the Y axis. It is reasonable to assume that without rehabilitation, the cost to the users as well as the cost to the agency increase with time, and that rehabilitation costs increase at an accelerating rate at the start. Thus H_j is increasing. It might also be convex, at least in the first years. This explains the form of the H_j curve.

The $G_j(n_j)$ curve is a little bit more difficult to analyse. It consists of two factors, the annuity factor and the bracketed annual cost expression. As n_j approaches zero, the first goes to infinity and the second goes to $R_j(0)$.[51] Thus $G_j(n_j)$ starts at infinity. The derivative of $G_j(n_j)$ is:

$$-\frac{\rho e^{-\rho n_j}}{1-e^{-\rho n_j}}\left[\left(C_{Bj}\left(n_j\right)+C_{Ej}\left(n_j\right)\right)-\frac{\rho}{1-e^{-\rho n_j}}\int_0^{n_j}\left(C_{Bj}\left(t\right)+C_{Ej}\left(t\right)\right)e^{-\rho t}dt\right]$$
$$+\frac{\rho}{1-e^{-\rho n_j}}\left(R_j'\left(n_j\right)-\frac{\rho e^{-\rho n_j}}{1-e^{-\rho n_j}}R_j\left(n_j\right)\right)$$

By the same argument as we used to see that the right hand side of Eq. (4) was negative, the bracketed expression in the first line here is positive. Also, it is clearly not infinitely large. The expression outside the bracket, on the other hand, approaches minus infinity as n_j goes to 0. We may perhaps assume that $R_j'\left(n_j\right)$ is small in size compared with the other cost items. If so, $G_j(n_j)$ falls steeply ar first, then become flatter and flatter. At some point it will begin to rise again, because in the end, the first derivative of $G_j(n_j)$ approaches $R_j'\left(n_j\right)$.

It is now possible to draw an approximate figure of the two curves, see Figure 1 in the text. In principle, it should be possible to find out if the minimum point of G occurs to the left or to the right of the point where $H = G$, but we leave this out here.

References

Ben-Akiva, M. and Ramaswamy, R. (1993) 'An approach for predicting latent infrastructure facility deterioration', *Transportation Science*, Vol. 27, no. 2, pp. 134-153.

Chu, C.-Y. and Durango-Cohen, P.L. (2008) 'Estimation of dynamic performance models for transportation structure using panel data', *Transportation Research B*, Vol. 42, pp. 57-81.

Dahl, G. and Minken, H. (2008) 'Methods based on discrete optimization for finding road network rehabilitation strategies', *Computers and Operations Research*, Vol. 35, no. 7, pp. 2193-2208.

Durango-Cohen, P.L. and Madanat. S. (2008) 'Optimization of inspection and maintenance decisions for infrastructure facilities under performance model uncertainty: A quasi-Bayes approach', *Transportation Research A*, Vol. 35, pp. 1074-1085.

Furuya, A. and Madanat, S. (2012) 'Accounting for Network Effects in Railway Asset Management', *Journal of Transport Engineering*, Vol.139, no.1, pp. 92-100.

Gao, L. (2011) 'Optimal Infrastructure Maintenance Scheduling Problem under Budget Uncertainty'. PhD dissertation, University of Texas at Austin.

[51] The integral probably goes to zero as n_j approaches 0, so for that part of the costs, l'Hopitâl's rule must be applied.

Gu, W., Ouyang,Y. and Madanat, S. (2012) 'Joint optimization of pavement maintenance and resurfacing planning', *Transportation Research B,* Vol. 46 (2012), pp. 511-519.

Karabakal, N., Bean, J.C. and Lohmann, J.R. (1994) 'Scheduling pavement maintenance with deterministic deterioration and budget constraints'. Technical Report 94-18, College of Engineering, University of Michigan.

Labi, S. and Sinha K.C. (2005). 'Life-cycle evaluation of flexible pavement preventive maintenance', *Journal of Transportation Engineering,* Vol. 131, no. 10, pp. 744-751.

Labi, S. and Sinha, K.C. (2003) 'The effectiveness of maintenance and its impact on capital expenditures'. Paper 208, Joint Transportation Research Program, Purdue University.

Lamptey, G., Labi, S. and Li, Z. (2008) 'Decision support for optimal scheduling of highway pavement preventive maintenance within resurfacing cycle', *Decision Support Systems,* Vol. 46, pp. 376-387.

Lee, J. and Madanat, S. (2014) 'Joint optimization of pavement design, resurfacing and maintenance strategies with history-dependent deterioration models', *Transportation Research B,* Vol. 68, pp. 141-153.

Lee, J. and Madanat, S. (2015) 'A joint bottom-up solution methodology for system-level pavement rehabilitation and reconstruction', *Transportation Research B,* Vol. 78, pp. 106-122.

Li, Y. and Madanat, S. (2002) 'A steady-state solution for the optimal pavement resurfacing problem', *Transportation Research A,* Vol. 36 (2002), pp. 525-535.

Mattsson L.-G. (2007) 'Railway capacity and train delay relationships'. Page 129-150 in: A.T. Murray and T.H. Grubesic (eds.) Critical Infrastructure: Reliability and Vulnerability, Springer Verlag, Berlin.

Paterson, W.D.O. (1987) 'Road deterioration and maintenance effects: Models for planning and management'. John Hopkins University Press, Baltimore.

Paterson, W.D.O. (1990) 'Quantifying the effectiveness of pavement maintenance and rehabilitation.' Proceedings at the 6th REAAA Conference, Kuala Lumpur, Malaysia.

Ouyang,Y. and Madanat, S. (2004) 'Optimal scheduling of rehabilitation activities for multiple pavement facilities: exact and approximate solutions', *Transportation Research A,* Vol. 38 (2004), pp. 347-365.

Ouyang,Y. and Madanat, S. (2006) 'An analytical solution for the finite-horizon pavement resurfacing planning problem', *Transportation Research B,* Vol. 40 (2006), pp. 767-778.

Rashid, M.M. and Tsunokawa, K. (2012) 'Trend curve optimal control model for optimizing pavement maintenance strategies consisting of various treatments', *Computer-Aided Civil and Infrastructure Engineering,* Vol. 27, no. 3, pp. 155-169.

Sathaye, N. and Madanat, S. (2011) 'A bottom-up solution for the multi-facility optimal resurfacing problem', *Transportation Research B,* Vol. 45, no. 7, pp. 1004-1017.

Sathaye, N. and Madanat, S. (2012) 'A bottom-up optimal pavement resurfacing solution approach for large scale networks', *Transportation Research B,* Vol. 46, pp. 520-528.

Small, K. A., Winston, C., and Evans, C. A. (1989). ‘Road Work; A new highway pricing and investment policy’. The Brookings Institution.

Tsunokawa, K. and Hiep, D.V. (2008) ‘A unified optimization procedure for road asset management’. 6th ICPT Conference.

Sydsæter, K., Strøm, A. and P. Berck (2005) ‘Economists’ Mathematical Manual’. Fourth edition, Springer Verlag, Berlin.

2.13 Environmental input output analysis and ecological footprints

Environmental input output analysis and ecological footprints[52]

Harald Minken

Institute of Transport Economics

Contents

[52] This is TOI working paper TØ/1332/2001, produced for the Prospects project. It has never been published before.

1 Input-output analysis

Consider a national economy consisting of n production sectors. A part of the output from the i'th sector is sold to other sectors who use it as input to their production, and another part of it is sold to end users. We will ignore intra-industry sales. Let X_i be sales in sector i in a particular year measured in real prices (that is, correcting for inflation). Let X_{ij} be sales from sector i to sector j and S_i be sales to end users, all measured in real prices. Obviously, we will have a system of equations like this:

$$(1)\quad \begin{array}{l} X_{11} + X_{12} + ... + X_{1,n-1} + X_{1n} + S_1 = X_1 \\ \vdots \\ X_{n1} + X_{n2} + ... + X_{n,n-1} + X_{nn} + S_n = X_n \end{array}$$

where we have assume the diagonal elements X_{ii} to be zero.

Define the coefficients a_{ij} by

$$(2)\quad X_{ij} = a_{ij} X_j$$

for all i and j between 1 and n. We assume these a_{ij} to be constants in the interval [0,1) and their sum over i to be less than 1 for all j. That is, the inputs to sector j from others sectors are assumed to cost less than the sales from j. Define **A** to be the matrix with elements a_{ij}. **X** is the column vector $(X_1, ...,X_n)'$ and **S** is the column vector $(S_1, ...,S_n)'$. Now the system (1) can be written **AX + S = X** or

$$(3)\quad (\mathbf{I} - \mathbf{A})\mathbf{X} = \mathbf{S}$$

where **I** is the identity matrix (a matrix where the diagonal elements are 1 and other elements are 0). If we consider **S** to be exogenously given, we can solve the system (3) for **X.** Under our assumptions, (**I -A**) will have an inverse, and thus there exists a unique solution for any **S**. The solution is

$$(4)\quad \mathbf{X} = \left(\mathbf{I} - \mathbf{A}\right)^{-1}\mathbf{S}$$

There is little reason to believe that the a_{ij} will in fact the constants, considering that each sector really produces a range of different products, has a choice of a wide variety of techniques, etc. But they may be sufficiently constant to make input-output analysis a good approach to a range of planning and forecasting problems.

The end use **S** can be partitioned into consumption, gross investment and export – possibly subdivided by end user sectors like households and the public sector. Suppose **S = C** + **J** + **Y**,

where **C** = (C_1, ...,C_n)´ is a column vector of consumption expenditure on goods and services from the different sectors, **J** = (J_1, ...,J_n)´ is a column vector of expenditure on goods and services used for investment purposes (gross investment) and **Y** = (Y_1, ...,Y_n)´ is a column vector of exported goods and services.

The inputs to the production processes that we have ignored up to this point are inputs not produced within any of our sectors, say imports and labour and capital services. We assume that when imports **Z** = (Z_1, ...,Z_n) are added to a sector's expenditure on goods and services from other sectors, the remaining difference between sales and costs is the value added in the sector. This makes the input-output analysis an accounting system for the whole national economy. This is illustrated in table 1. Value added in the sectors is the row vector **R** = (R_1,...,R_n). Total value added (gross domestic product GDP) is denoted by *R*, total export is *Y*, total import *Z*, total consumption *C* and total investment *J*. Imports used directly for consumption is Z_C and imports used directly for investment is Z_J.

Table 1. Input-output table for an open economy

Demand Supply	Sector 1	...	Sector n	Con sumption	Invest- ment	Export	Sum
Sector 1	$a_{11}X_1$	...	$a_{in}X_n$	C_1	J_1	Y_1	X_1
⋮	⋮	⋱	⋮	⋮	⋮	⋮	⋮
Sector n	$a_{n1}X_1$	...	$a_{nn}X_n$	C_n	J_n	Y_n	X_n
Import	Z_1	...	Z_n	Z_C	Z_J	-	Z
Value added	R_1	...	R_2	-	-	-	R
Sum	X_1	...	X_n	C	J	Y	

Since the column and row sum should be the same, $R + Z = C + J + Y$. Thus the available resources GDP plus imports are used for consumption, investment and export.

Consider now the matrix of non-produced inputs to the production sectors. In the case illustrated in table 1, it is

$$\begin{pmatrix} Z_1 & \cdots & Z_n \\ R_1 & \cdots & R_n \end{pmatrix}$$

If we may assume that these inputs too are required in fixed proportions to the total sector output, we can divide each column j of this matrix by X_j to get the fixed coefficients b_{ij}:

$$(5) \quad Z_j = b_{1j} X_j \text{ and } R_j = b_{2j} X_j$$

Let the matrix of b_{ij} elements be called **B**. **B** is a (2, n) matrix in our case, but might have any number of rows, for example if we split imports into imports of energy and other imports. But

to continue with our example, let the sum over j from 1 to n of the Z_j be Z_0. (We disregard imports directly for consumption at the moment). We call the column vector of non-produced inputs **Q**, **Q** = $(Z_0, R)'$. Now we have

(6) $\mathbf{Q} = \mathbf{BX}$

Obviously, the required amounts of non-produced inputs to cater for any end use **S** can be found by first using (4) to find **X** and then inserting this **X** in (6), or directly by solving $\mathbf{Q} = \mathbf{B}(\mathbf{I} - \mathbf{A})^{-1}\mathbf{S}$. Note however that **S** as we have defined it does not include the direct end use of imports. To get the total import requirement *Z*, $Z_C + Z_J$ must be added to the Z_0 found by (6). Total consumption and gross investment must also be found by adding Z_C and Z_J respectively.

The matrix of non-produced input requirements and the **B** matrix of requirements per dollar of output both measure the input requirements in real prices, just like the matrix of requirements for produced inputs and the **A** matrix. There is however no problem in forming matrices in analogy with the **B** matrix, but in physical units. This can be useful in environmental analysis, to which we turn in the next section.

2 Applications to environmental analysis

An early contribution to the literature on environmental analysis by way of input-output analysis is Leontief and Ford (1971). They consider the following extension of the **A** matrix defined above:

- Add rows showing the output of pollutant g per unit of output from sector i. These will consist of elements a_{gi}, where i runs from 1 to n as before and g runs from n + 1 to m if there are m – n pollutants. The pollutants will be measured in physical units, while sector output will still be measured in money. Thus a_{ig} is pollution of type g per dollar of output from sector i.
- Add columns representing pollution abatement sectors. Each abatement sector is engaged in eliminating a particular pollutant, so there will be as many abatement sectors as there are pollutants. The elements a_{ig} will represent the expenditure on good i per unit of eliminated pollutant g, and the elements a_{gk} (where g and k both run from n + 1 to m) will represent the output of pollutant g per unit of eliminated pollutant k.

In their study, data for the abatement activities were lacking, and so only the first bullet point, the addition of new rows showing levels of pollutants per dollar of output, were implemented. Most later studies, as far as we can see, have also disregarded abatement activities. We will do the same.

The matrix consisting of the elements a_{gi} of pollution per euro of sales from sector i will be denoted **M**. **M** is an (m – n, n) matrix. Emissions in physical units as a function of end use of the sector outputs will be denoted by **E**. **E** is a column vector $(E_{n+1},...,E_m)'$. As **M** is constructed in exactly the same way as the **B** matrix above, adding rows to table 1 showing the amount of some substance required or created by the sector output levels, obviously the vector **E** of pollution created by the end use **S** is

(7) $\mathbf{E} = \mathbf{MX} = \mathbf{M}(\mathbf{I} - \mathbf{A})^{-1}\mathbf{S}$

In exact analogy to the case of non-produced inputs, we have to add to **E** the pollution created by end use activities like household consumption and investment. These amounts too may be proportional to the level of activity. If relevant, we will also have to add the pollution created by the imports.

The fixed coefficient assumptions are perhaps a dubious feature if this method is used for long term forecasting. However, technological change may of course be modelled by changing the coefficients. The strong sides of this method are that it is much more consistent than even the most elaborate life-cycle analysis (LCA) in tracing all indirect effects, and that it is simple to apply if there already exists some input-output model for the region and if the sectors are aggregate enough to make the job of setting pollution coefficients manageable.

3 Ecological footprints – a special case

The ecological footprint of a population in a region is the productive area that is necessary to sustain the current consumption patterns of this population in the long run. It consists of "energy producing land", the built area, cropland, gardening land and pasture land, and productive forests. The area must be thought of as consisting of areas spread out over the whole world, and not necessarily situated in the region where the population lives. The energy producing land must be thought of as the area needed to produce the energy in a renewable form.

It is convenient to express the ecological footprint in hectares per capita. This indicator of sustainability was developed about ten years ago by the Canadian W.E. Rees and fellow workers and has attracted considerable interest since.

Bicknell et al (1998) compute ecological footprints from input-output data. Basically, their method consists in defining land as a non-produced input. Their paper is not perfectly clear on every point of methodology, but a **B** matrix for their method would probably consist of four rows: "domestic" land requirements per dollar of sector output, energy imports per dollar of sector input, other imports and value added. Total land requirements to sustain a certain pattern of end uses would then have to be found from an equation like $\mathbf{Q} = \mathbf{B}(\mathbf{I} - \mathbf{A})^{-1}\mathbf{S}$, plus additional calculations to find the land required to produce the imported energy (one of the elements of **Q**), the land required for the production of other imports (another element of **Q**) and the land required for end uses (the built area).

4 Relevance to PROSPECTS

We assume that pollution from transport can be computed directly from the transport models. Likewise, pollution from direct energy uses in the households (heating, lighting and appliances) could probably be computed from data on the number of households, the average residential area per household and statistics on different kinds of energy use per household

and type of housing. So when will input-output analysis and a formula like (7) be useful to us?

Probably it could be used to compute total city emission of CO_2, NO_x, SO_2 etc. from all consumption activities except transport and direct energy use in the household, or even including transport if no reliable transport model is at hand. The "indirect" pollution through the use of goods and services whose production is polluting will however not vary between *strategies*. It may be different in different *scenarios* according to income, demographic and technological assumptions.

The method is in line with the definition of the urban system that was outlined in Deliverable 1. Remember that we were primarily occupied with the sustainability of urban consumption. The sustainability of the systems of production to which the urban export industries belong, we said, must be judged on a wider scale than the city and is a problem not addressed in PROSPECTS. To judge about the sustainability of urban consumption we must however address environmental consequences and natural resource requirements of the production of the goods and services that enter into urban consumption, regardless of where it takes place. Environmental input-output analysis and the ecological footprint method do exactly this in a consistent manner.

The indicators of pollution, energy use and sustainability produced by the input-output method have the important property that they are model based and thus can be computed for assumed future states (scenarios).

Obviously, the method is all the more interesting if there already exists a regional model for the urban area with an input-output table at the core. As far as I know, some integrated land use and transport models include input-output relationships. In these cases one may build on the **A** matrix of the model to develop the environmental indicators.

The environmental and energy indicators that we develop from input-output analysis are for the city as a whole and will not be available at the zonal level. Thus if location within the city is an issue (local pollution, degradation of valuable sites within the city, traffic accidents), other methods must be used.[53] However, for energy use and the total environmental impact of the level of urban consumption, the input-output approach will be well suited.

References

Bicknell, K.B., R.J. Ball, R. Cullen and H.R. Bigsby (1998) New methodology for the ecological footprint with an application to the New Zealand economy. *Ecological Economics* **27**, 149-160.

Leontief, W. and D. Ford (1971) Air pollution and the economic structure: empirical results of input-output computations. In: Brody, A. and A.P. Carter (eds) *Input-Output Techniques. Proceedings of the Fifth International Conference on Input-Output Techniques*. North-Holland Publishing, Amsterdam.

[53] This might not be the case for some of the most elaborate land use/transport models.

TØI is an applied research institute that carries out research and study assignments for businesses and public agencies. TØI was established in 1964 and is organized as an independent foundation. TØI develops and disseminates knowledge about transport with scientific quality and practical application. The department has an interdisciplinary environment with 90+ highly specialized researchers.

The department conducts research dissemination through TØI reports, articles in scientific journals, books, seminars, as well as posts and interviews in the media. The TØI reports are available free of charge on the department's website www.toi.no.

The institute participates actively in international research collaboration, with particular emphasis on the EU framework programs.

TØI covers all means of transport and thematic areas within transport, including traffic safety, public transport, climate and environment, tourism, travel habits and travel demand, urban planning, ITS, public decision-making processes, business transport and general transport economics.

Postal Address:
Institute of Transport Economics
Gaustadalléen 21
N-0349 Oslo
Norway

Email: toi@toi.no

Business Address:
Forskningsparken
Gaustadalléen 21

Phone: +47 22 57 38 00

Web address: www.toi.no