

Economics 416: Cost-Benefit Analysis: Lecture Notes

Part One: The Measurement of Social Costs and Benefits

Please Note: The material that is covered in Part One of the course can be found in the first four chapters of the Boardman et. al. textbook. The most important of these chapters is Chapter Four.

(a) Cost-Benefit Analysis and Program Evaluation

Cost-benefit analysis is a public sector decision-enabling and planning tool, a measure of program effectiveness and project efficiency, and a branch of applied welfare economics in the public finance stream.

The cost-benefit analysis of a project or program uses the *net social benefits* (NSB) criterion for evaluation purposes, where NSB equals social benefits (B) less social costs (C). Efficient social choices involve the selection of those projects for which net social benefits are highest. The maximisation of net social benefits is the public sector equivalent to the maximisation of profits in the private sector.

The net social benefits criterion, $NSB = B - C$, is used for ranking projects or programs rather than the benefit-cost ratio, B/C , because benefit-cost ratios can be manipulated by redefining costs as negative benefits, and benefits as negative costs. Just as private firms maximise profits, and not the profit margin or the rate of profit, in cost-benefit analysis we again maximise a dollar value rather than a ratio between two different dollar values.

The nine basic steps of cost-benefit analysis are:

- Specify the set of alternative projects
- Decide whose benefits and costs count (standing)
- Catalogue the impacts and select measurement indicators (units)
- Predict the impacts quantitatively over the life of the project
- Monetize (attach dollar values to) all impacts
- Discount benefits and costs to obtain present values
- Compute the net present value (NPV) of each alternative
- Perform sensitivity analysis
- Make a recommendation based upon the NPV and sensitivity analysis

(b) Cost-Benefit Analysis and Economic Efficiency

The promotion of economic efficiency is an important normative goal. Economic efficiency includes both management (or production) efficiency, which involves cost minimisation in producing the chosen output levels, and allocation (or Pareto) efficiency, which involves choosing the right output levels to maximise net social benefits (NSB), or social benefits (B) minus social costs (C). The main technique used to assess economic efficiency is *Cost-Benefit Analysis*.

The net social benefits criterion suggests that one should adopt only those programs and/or combinations of projects that have positive net social benefits. The adoption of a program with positive net social benefits involves an improvement in economic efficiency, or a potential Pareto improvement.

If net social benefits are positive, it is possible to find a set of transfers, or side payments, that make at least one person better off without making anyone else worse off. Thus, programs that generate positive net social benefits obey the Kaldor-Hicks criterion, which states that a program should be adopted only if those who will gain from the program could fully compensate those who will lose and still be better off.

Among those programs and/or project combinations which are potential Pareto-improving and obey the Kaldor-Hicks criterion, one should select those programs or project combinations which maximise net social benefits. These project combinations are said to be Pareto-efficient, by which one means that no alternative selection of projects can make at least one person better off without making anyone else worse off. A Pareto-efficient allocation implies that no further potential Pareto-improving program changes are available to be used. All available program changes would fail to satisfy the Kaldor-Hicks criterion.

(c) The Maximization of Net Social Benefits Criterion

Cost-benefit analysis involves more than financial analysis. Two measurement rules are:

- (a) value social benefits as the direct budgetary revenues (R) associated with the project plus (less) any increase (decrease) in *social surplus* that occurs in the primary output markets impacted by the project; and
- (b) value social costs as direct budgetary expenditures (project outlays, E) less (plus) any increase (decrease) in *social surplus* that occurs in the primary input markets impacted by the project.

Within these measurement rules, social surplus refers to the sum of *consumer surplus* (CS) and *producer surplus* (PS). Consumer surplus measures the amount that consumers would be *willing to pay* for a desirable good less the amount that they actually do pay for the good (transaction outlays). Producer surplus measures the amount that producers actually receive from the sale of a good (transaction receipts) less the *opportunity costs* of producing the good. Social surplus is, therefore, the difference between willingness to pay and opportunity costs.

It follows that $NSB = dCS + dPS + dGR$, where dGR is the net change in budgetary revenues (R-E), and dCS and dPS are, respectively, the net changes in consumer and producer surplus that are observed in the primary output and input markets combined. The basic cost-benefit criterion is to invest in those projects which have the highest positive NSB values.

(d) Elasticities, Incidence Effects, and Deadweight Losses

Suppose that supply costs in an undistorted competitive market increase by c dollars. This increase will lower the market-clearing quantity from $Q(0)$ to $Q(1)$. When this occurs, the loss of *social surplus* is equal to

$$\text{Social surplus loss} = c [(Q(0) + Q(1)) / 2].$$

What portion of this social surplus reduction is a loss in consumer surplus, and what portion is a loss in producer surplus? What is the *incidence* of the social surplus loss?

It turns out to be the case that the consumer surplus and producer surplus loss proportions are given, respectively, by $b = e(s) / [e(s) + e(d)]$, and $(1-b) = e(d) / [e(s) + e(d)]$, where $e(s)$ is the elasticity of supply and $e(d)$ is the absolute value of the elasticity of demand. Thus, the loss in consumer surplus is equal to

$$\text{Consumer surplus loss} = b c [Q(0) + Q(1)] / 2, \text{ where } b = e(s) / [e(s) + e(d)],$$

and the loss in producer surplus is equal to

$$\text{Producer surplus loss} = (1-b) c [Q(0) + Q(1)] / 2, \text{ where } (1-b) = e(d) / [e(s) + e(d)].$$

Since the difference between the new demand side price, $p(d)$, and the unadjusted supply side price, $p(s)$, is also equal to c , this analysis can also be applied to the case of an excise tax or specific duty, t , which is equal to c . However, in this case, there is a revenue fall-in to government which is equal to

$$\text{Government revenue gain} = [p(d) - p(s)] Q(1) = t Q(1).$$

But the government revenue gain is smaller than the social surplus loss, the difference representing a *deadweight loss* from government taxation. The deadweight loss is equal to

$$\text{Deadweight loss} = t [Q(0) + Q(1)] / 2 \text{ less } t Q(1) = t [Q(0) - Q(1)] / 2.$$

It follows that there is an *excess burden of taxation*, which implies that the social cost of each taxation dollar raised is larger than one dollar because the market equilibrium is distorted by taxation. The *marginal excess burden of taxation* (METB) is given by the ratio of the deadweight loss to the tax revenues raised. Each dollar of tax revenue raised costs society $(1 + \text{METB})$ dollars. Notice also that $(1 + \text{METB})$ is the ratio of the social surplus loss to the tax revenues raised. Thus,

$$\text{METB} = [Q(0) - Q(1)] / 2Q(1), \text{ and } (1 + \text{METB}) = [Q(0) + Q(1)] / 2Q(1).$$

From a theoretical point of view, government revenue effects within a cost-benefit analysis framework should be augmented by $(1+\text{METB})$, but this is often ignored in practice. The size of the $(1+\text{METB})$ adjustment depends upon the degree to which taxation distorts market equilibrium, and this will vary for different kinds of taxes. In the case of an excise tax or specific duty, the distortion is larger the larger are the demand and supply elasticities.

(e) The Measurement of Consumer Surplus

The *compensating variation* associated with an economic change is the amount of money income that can be taken from or given to an individual while leaving that individual as well off as *before* the economic change. Compensating variation always involves the question: what would it take to return the individual to the prior utility level once the change has occurred? The comparative benchmark (or base case) is the initial situation. More generally, compensating variation measures the *willingness-to-pay* for a welfare gain or the *willingness-to-accept* a welfare loss.

If the economic change is a price increase, the compensating variation is the amount of money income that the consumer must receive to leave utility unaffected by the price increase (that is, to get the consumer back onto the original indifference curve). It is

therefore a measure of the consumer's *willingness-to-accept* the price increase (a welfare loss).

On the other hand, if the economic change is a price decrease, the compensating variation is the amount of money income that can be taken from the consumer while leaving utility unaffected by the price decrease. It is therefore a measure of the consumer's *willingness-to-pay* for the price decrease (a welfare gain).

The compensating variation associated with a price change can be illustrated by using a Hicksian, or compensated, demand curve along which utility (or real income) is held constant. The downward slope of the Hicksian demand curve relates only to substitution effects. Income effects are compensated away. Since income effects are positive for normal goods, the Hicksian demand curve is steeper than the Marshallian, or ordinary, demand curve, along which money income is held constant. It follows that Marshallian consumer surplus slightly *overstates willingness-to-pay* for a price decrease (compensating variation), while it slightly understates willingness-to-accept a price increase.

The *equivalent variation* associated with an economic change is the amount of money income that can be given to or taken from an individual while leaving that individual as well off as *after* the economic change. Equivalent variation always involves the question: what would it take to move the individual to the new utility level if the change did not occur? The comparative benchmark (or base case) is the new situation. More generally, equivalent variation measures the consumer's *willingness-to-pay* to avoid a welfare loss or *willingness-to-accept* forgoing a welfare gain.

If the economic change is a price increase, the equivalent variation is the amount of money income that the consumer would be willing to forgo in order to avoid the price increase (and thereby place the consumer on the same indifference curve as did the price increase). It is therefore a measure of the consumer's *willingness-to-pay* to avoid the price increase. On the other hand, if the economic change is a price decrease, the equivalent variation is the amount of money income that the consumer would need to be paid to be as well off as with the price decrease. It is therefore a measure of the consumer's *willingness-to-accept* not benefiting from the price decrease. Notice that the equivalent variation for a price increase is the compensating variation for a price decrease, and vice versa.

For normal goods, Marshallian consumer surplus slightly *overstates willingness-to-pay* to avoid a welfare loss (equivalent variation), while it slightly understates willingness-to-accept forgoing a welfare gain. It follows that, when income effects are non-negative, each of the following inequalities applies:

- (a) $WTA > CS > WTP$,
- (b) $EV > CS > CV$, for a price decrease (or welfare gain), and
- (c) $CV > CS > EV$, for a price increase (or welfare loss),

where CS = Marshallian consumer surplus, CV = compensating variation, EV = equivalent variation, WTA = willingness-to-accept, and WTP = willingness-to-pay. Thus, when income effects are non-negative, it will ordinarily be the case that

- (d) willingness-to-accept (WTA) $>$ CS $>$ willingness-to-pay (WTP).

As long as income effects are relatively small in comparison to substitution effects, Marshallian (or ordinary) consumer surplus is an adequate measure of willingness-to-pay, and, thus, of welfare effects. Indeed, if the income effect is negligible in size in relationship to the substitution effect, then the three measures of consumer surplus coincide, and $CV = CS = EV$. In most instances, it would be reasonable to assume the close approximation of Marshallian consumer surplus to willingness-to-pay and/or willingness-to-accept.

However, the pair-wise differences between Marshallian consumer surplus, willingness-to-pay and willingness-to-accept may be of importance when irreversible environmental outcomes are being considered, that is when there are few close substitutes for an environmental resource but welfare (income) effects could be substantial. In this case, *loss aversion* would render WTP (the amount that an individual would be willing to pay to avoid a loss of size x) significantly smaller than WTA (the amount that an individual would be willing to receive to compensate for a loss of size x). Because willingness-to-accept question formats can lead to open-ended answers, willingness-to-pay question formats should ordinarily be used within the contingent valuation method.

(f) Market Failure

Competitive markets, in which consumers are free to choose what they purchase, and production is not concentrated in the hands of a few, facilitate economic efficiency. When markets are functioning competitively, government intervention generally leads to market distortions and efficiency losses. Public policy interventions should largely be directed to situations in which market failure occurs.

Market failure is a situation in which private markets fail to achieve allocation (or Pareto) efficiency. There are a number of types of market failure (or distortion). These include imperfect competition, asymmetric information, positive externalities, negative externalities, public goods, inter-temporal market failures, and policy-induced market failures. We begin by considering the market failure that is associated with the provision of public goods.

(g) Classification of Goods

	Non-Excludable	Excludable
Non-Rival	Public Goods	Club and/or Toll Goods
Rival	Congested, Open Access Goods	Private Goods
Nature of Property Rights	Usufructuary	Proprietary

Market failure is associated with non-excludability. In the case of public goods, insufficient quantities would be provided by private markets due to the *free rider problem*. In the case of congested, open access resources, economic rents are dissipated, and the resource stock is over-exploited (too many harvesters chasing too few resources).

Using a bundle of sticks analogy, property rights systems are normally characterised by comprehensiveness, exclusivity, enforceability (security of tenure), duration of tenure, transferability, and benefits conferred. The *attenuation* of a property right (e.g., the taking back of a resource tenure) may or may not be associated with a legal requirement for compensation.

(h) Environmental Externalities

Market failure is also associated with *externalities*, or non-market spill-over effects (usually on third parties) that often have environmental consequences. Many environmental externalities occur because of avoidable and unavoidable gaps in our systems of property rights.

Positive externalities occur when the marginal social benefits associated with an activity exceed the marginal private benefits. Insufficient quantities of the activity are undertaken because the private sector is unable *to appropriate* this excess value, resulting in a loss of economic efficiency.

Negative externalities occur when the marginal social costs associated with an activity exceed the marginal private costs. Excessive quantities of the activity are undertaken because the private sector fails *to internalise* these excess costs, again resulting in a loss of economic efficiency.

Taxation and subsidy instruments may be used to get around the problems of internalisation and lack of appropriability. One example can be drawn from Forest Renewal BC. If forest replenishment and watershed restoration have positive non-timber benefits, subsidise these activities; whereas, if timber harvesting reduces non-timber benefits (e.g., reduces amenity and/or eco-system service values), tax this activity by increasing stumpage charges.

If environmental externalities lead to market distortions, a third measurement rule applies:

(c) the net changes in social costs (benefits) that are associated with negative (positive) environmental externalities should be added to primary social costs (benefits).

The augmented rule is, therefore, $NSB = dCS + dPS + dGR + dNEB$, where $dNEB$ is the net impact of the project on externality benefits and costs. Among projects for which $NSB > 0$, one should invest in those with the highest NSB .

(i) Unemployed Labour and Cost-Benefit Analysis

The employment of previously underemployed or unemployed labour gives rise to an increase in social surplus. Project labour cost outlays exceed the social opportunity cost of the labour hired, which would otherwise earn a lower wage. Under these circumstances, the *shadow wage rate* that should be used in cost-benefit analysis is less than the wage rate observed in the primary labour market, and can often be measured by the wage rate in a secondary (or next best alternative) labour market.

In cost-benefit analysis, the wages paid to employ labour on a project should always be treated as a cost. The fact that a project may generate employment does not get counted as a benefit. However, if project outlays involve the payment of wage rates that are higher

than those that the workers could earn under alternative circumstances, then there will be an element of producer surplus to set against the project wage bill. Gains associated with increased labour incomes are a cost offset, rather than a project benefit.

For example, when a project increases wage rates from $w(0)$ to $w(1)$, the proper measures of the *shadow price* (or opportunity cost) of project labour, and the per worker gain in producer surplus are, respectively, given by:

$$[w(1) + w(0)] / 2, \text{ and } [w(1) - w(0)] / 2.$$

When there is significant unemployment, the *reservation wage* (or social support wage), $r(0)$, may be used to replace $w(0)$ in the previous formulas. Thus, the *shadow price* (or opportunity cost) of project labour, and the per worker gain in producer surplus, respectively, become:

$$[w(1) + r(0)] / 2, \text{ and } [w(1) - r(0)] / 2.$$

When there is a dual labour market structure involving, say, both urban and rural labour markets, and labour is free to migrate between markets in search of jobs, the *expected wage* in the urban labour market may be used to *shadow price* the labour employed on urban projects. If z is urban employment, u is urban unemployment, and w is the urban wage rate, the wage expected by labourers who migrate from the rural to the urban sector, $E(w)$, equals p times w , where $p = z / (z+u)$ is the probability of finding a job in the urban sector. An average of the rural wage and the expected urban wage is used as the shadow price, but in migration equilibrium the rural wage and the expected urban wage are equal. The shadow price of labour is thus inversely related to the urban unemployment rate.

(j) Cost-Effectiveness Analysis

Cost-benefit analysis is based upon the efficiency principle, which states that net social benefits are maximised when the marginal social benefits from an allocation of resources are equal to the marginal social costs. However, it is often the case that marginal social benefits are difficult to quantify, let alone monetise. In these cases, one may have to fall back on cost-effectiveness analysis. The cost-effectiveness principle states that the least-cost means of achieving an environmental target or other goal will have been attained when the marginal costs of all possible means of achievement are equal.

There are two alternative ways of framing cost-effectiveness analysis objectives. These alternative ways are:

- (a) to choose the activity or process j which maximises the outcomes to cost ratio, $E(j)/C(j)$, subject to a budget constraint: $C(j) < C$, and
- (b) to choose the activity or process j which minimises the cost of providing a desired outcome level, that is minimise $C(j)/E(j)$, subject to $E(j) > E$.

It should be noted that cost-effectiveness is a necessary, but not sufficient, condition for economic efficiency. Satisfying the cost-benefit criterion ($NSB > 0$) requires, at a minimum, cost-effective decisions to be made.

Part Two: The Timing of Benefits and Costs: Net Present Value (NPV)

Please note: The material that is covered in Part Two of the course can be found in Chapters 6-8 and 10 of the Boardman et. al. textbook. The most important of these chapters are Chapters 6 and 10.

(a) Net Present Value (NPV)

Cost and benefit streams are often multi-faceted, time-dependent and uncertain. The net present value (NPV) criterion allows for the time-dependency of benefits and costs. Net present value may be written as follows:

$$NPV = \sum B(t)/(1+i)^t - \sum C(t)/(1+i)^t = \sum NB(t)/(1+i)^t,$$

where $B(t)$, $C(t)$, and $NB(t) = B(t) - C(t)$, are benefits, costs and net benefits which accrue at time t , and i is the interest rate (or discount rate). When costs and benefits accrue as streams over time, the cost-benefit analysis criterion, $NSB > 0$, needs to be restated as the $NPV > 0$ criterion. The efficiency rule of choosing to invest in projects with the highest NSB becomes an “invest in those projects with the highest NPV” rule.

The annuity value of the net benefit stream (A), or the *equivalent annual net benefit* (EANB), is the amount if received every year from year 1 to year T would generate the same NPV as the actual stream of net benefits from a project. The annuity value (A) represents levelised net benefits, and may be used to compare projects of different length (that is, for which the T 's differ). Thus,

$$\sum NB(t)/(1+i)^t = NPV = A [1 - 1/(1+i)^T] / (i),$$

where everything to the right of A in the last expression is called the *annuity factor*, or the present value of \$1 per annum if received every year, at year end, for T years. Notice that the annuity factor converges on $1/(i)$ as T gets very large (as in a perpetuity).

In cost-benefit analysis, one ordinarily likes to work in real terms, expressing costs and benefits in constant-dollar (inflation-adjusted) terms, and using a real interest rate (or inflation-adjusted interest rate). If x is the expected rate of inflation, then the real interest rate (r) is equal to $r = (i - x)/(1+x)$, where (i) is the nominal interest rate.

NPV issues: growing streams, declining streams, terminal values, decommissioning costs, and perpetuities. What permanent income stream is equivalent to the net cash flow from a depleting resource? What proportion of net cash flow must be saved and reinvested in other assets if consumption is to be maintained over time?

(b) The Choice of Discount Rate: Fundamental Issues

The trade-off that society makes between consumption today and savings to invest in ways which would enhance consumption tomorrow is measured by the *social discount rate*. Let this interest rate (in real terms) be r . The trade-off that society makes when it chooses to invest in one project rather than an alternative project is measured by the *social opportunity cost of capital*. Let this interest rate (in real terms) be m . If capital markets were perfect,

then m and r would be equal. However, numerous risks and uncertainties, and various tax wedges, make m significantly larger than r .

In general, if b measures the interest rate on investment grade debt instruments, then $r < b < m$, because there are tax wedges on both the lender (household consumer/saver) and the borrower (firm/investor in productive assets) side of the ledger. For example, m may be approximated by the borrower's *hurdle rate of return* less its economic depreciation rate, or by

$$m = v b + (1 - v) (b + p) / (1 - u),$$

where v = share of debt in firm capital, $(1 - v)$ = share of equity in firm capital, p = equity risk premium, and u = corporate taxation rate.

Whether one wants to use an r -type, a b -type, an m -type of interest rate, or some form of weighted-average interest rate depends upon the context. If the project's financing is expected to crowd out alternative private sector investment, and its benefits are mostly in the form of consumption, an m -type of discount rate seems to be required. If the benefits of a project accrue mostly as government revenues, then a b -type of discount rate (or the government borrowing rate) should be used. If the project has resource conservation and/or environmental restoration as its primary focus, then an r -type of interest rate would seem to be appropriate.

Finally, from the perspective of inter-generational equity, an r -type of discount rate seems to be more appropriate than an m -type of discount rate. Inter-generational equity is compatible with a positive discount rate provided that the economy has a positive rate of growth in consumption per head. One may legitimately discount streams of benefits that accrue to future generations if future generations are expected to be better off than the present generation.

(c) The Shadow Price of Capital

The *shadow price of capital* approach requires one to change all costs and benefits into equivalent units of domestic consumption and then to discount using the social discount rate, r . If q is the shadow price of capital, then

$$B^*(t) = [(1 - w(B) + w(B)q)] B(t), \text{ and}$$

$$C^*(t) = [(1 - w(C) + w(C)q)] C(t),$$

where $w(B)$ is the proportion of benefits which take the form of investment benefits rather than consumption benefits, and $w(C)$ is the proportion of costs which occur at the expense of alternative investments rather than at the expense of consumption. Then,

$$NPV = \sum [B^*(t) - C^*(t)] / (1+r)^t .$$

The trick is then to find an appropriate measure for q , the shadow price of capital. It turns out that $q > 1$, and that q increases with the size of the ratio $(m+d)/(r+d)$, where d is the overall economic depreciation rate of invested capital. In fact,

$$q = (1 - s) / [(r+d)/(m+d) - s],$$

where s is the economy's overall savings propensity.

This formula comes from the overall macro-economy. The marginal return on capital for the economy as a whole is $m+d$. If s of this return is saved and reinvested, the value of this investment is $s(m+d)q$. The value of the remaining consumption is $(1-s)(m+d)$. The present value of a perpetual stream of these returns requires discounting at the rate $r+d$. However, the present value of this stream is also the value of a unit of invested capital to the economy measured in terms of the *numeraire* consumption good. Hence,

$$q = [s(m+d)q + (1-s)(m+d)] / (r+d), \text{ or}$$

$$q = (1-s)(m+d) / [(r+d) - s(m+d)] > 1.$$

By way of example, if $s = 0.2$, $d = 0.1$, $r = 0.02$ and $m = 0.1$, then q would be 2. This would tend to be an upper bound for q because the difference between m and r is unlikely to be larger than this. Indeed, if $m = 0.08$, then q would be 1.71.

Notice that if benefits all take the form of consumption, $w(B) = 0$, and if all costs occur at the expense of private investment, $w(C) = 1$. In this case, the shadow price approach clearly penalises the project, since $B^*(t) - C^*(t) = B(t) - qC(t)$, so that costs get augmented by the shadow price of capital, $q > 1$. The alternative approach to this situation would be to use m rather than r as the *weighted average* discount rate, again penalising the project.

(d) Does the Financing Method Matter in Cost-Benefit Analysis?

A public investment involves a large initial capital expenditure, C , and produces a stream of consumption benefits over the next T years. Using the social rate of discount, r , the present value of the consumption benefits stream is B .

Society faces a marginal excess burden of raising public revenues from taxation equal to x . The shadow price of capital equals q . Public borrowing replaces private investment dollar for dollar. Public revenues raised through taxes displace investment in the proportion, w , and consumption in the proportion, $1 - w$.

(a) Using the shadow price of capital method, the net present value of the project when fully financed out of taxes is equal to

$$NPV(\text{taxes}) = -C(1+x)(1-w+wq) + B.$$

(b) Alternatively, if the project were financed fully by public borrowing, which is not expected to be repaid by taxes until the indefinite future (well past year T), the net present value would be

$$\text{NPV (borrowing)} = - Cq + B.$$

(c) If the project were wholly self-financed through user fees charged to the beneficiaries of the project, and the beneficiaries reduced their annual usage of the services provided by the project to a proportion, y , of its original level as a result of the imposition of the fee-for-service financing scheme, the net present value would be

$$\text{NPV (user fee)} = - C + yB.$$

(d) If the marginal excess burden, x , equals 0.25, the shadow price of capital, q , equals 2, the proportion, w , equals 0.20, and the proportion, y , equals 0.60, the three alternative NPVs are

$$\text{NPV (taxes)} = - C(1+0.25)(1-0.2+0.4) + B = - 1.6C + B,$$

$$\text{NPV (borrowing)} = - 2C + B, \text{ and}$$

$$\text{NPV (user fee)} = - C + 0.6B.$$

Ranked by benefit-cost ratios, $B/C \text{ (taxes)} > B/C \text{ (user fees)} > B/C \text{ (borrowing)}$. The financing method often matters in CBA.

(e) Risk and Uncertainty in Cost-Benefit Analysis

When project outcomes are uncertain, one possible approach is called the *expected surplus* approach. This approach requires one to place probabilities, either objective or subjective, on alternative possible outcomes, or *states of nature*. Suppose that there are n possible states of nature, to which the probabilities, $p(1) \dots p(n)$, can be assigned. By definition, the sum of the probabilities over all states of nature must be unity. Suppose that each state of nature is associated with a certain level of net benefits, $NB(1) \dots NB(n)$. Then the expected net benefits associated with the project are equal to

$$E(NB) = \sum p(j) NB(j).$$

The expected surplus, $E(S)$, is the net present value of the stream of net benefits, or

$$E(S) = \sum E[NB(t)] / (1+r)^t,$$

where r is a suitable discount rate. Projects in similar risk classes may be ranked by their expected surpluses.

The expected surplus approach may be used in a number of different circumstances:

- (a) if the risks associated with the set of alternative projects are all similar, or
- (b) if the decision-maker is risk neutral, or
- (c) if risks can be diversified away or insured against in an actuarially fair manner.

However, if none of these three assumptions is valid, alternative assessment methods may need to be employed. This may often be the case with collective risks, against which society is unable to diversify. One alternative method is called the *option price* approach.

An option price measures the amount that an individual decision-maker is willing to pay for an uncertain prospect prior to the realisation of contingencies, that is, prior to learning which state of nature will materialise. For a risk-neutral individual, this is the same as the expected surplus. However, for a risk-averse individual, option price will ordinarily differ from expected surplus. In general, whenever an uncertain prospect increases income risk, the option price associated with the uncertain prospect will be smaller than the expected surplus, to adjust for the costs of taking on additional risk.

An individual may be said to be *risk-averse* when he/she prefers an outcome which provides x-dollars with certainty to an uncertain prospect whose expected outcome is x dollars. For example, an individual may be indifferent between playing a *single* coin toss game that provides +\$600 for heads and -\$400 for tails and not playing the game at all. Notice that the expected outcome of this game, *if played many times over*, is \$100. If an uncertain prospect whose expected outcome is \$100 provides an individual with no increment in utility, and thus with less incremental utility than would a certain wealth gain of \$100, the individual is risk-averse.

The *certainty-equivalent* of an uncertain prospect is the amount which, if received with certainty, provides the same *level of utility* as the uncertain prospect. The individual is indifferent between taking on the uncertain prospect and the certainty-equivalent. In the previous example, the certainty-equivalent of the single coin toss game is \$0. Notice that, due to risk-aversion, the certainty-equivalent is less than the expected value of the game. The utility associated with the certainty-equivalent is the *expected utility* from playing the game. The expected utility approach to risk and uncertainty says that we should replace expected values with certainty-equivalents and, thus, use the option price concept rather than the expected surplus concept for ranking alternative projects. The option price may be written as

$$OP = \sum CE[NB(t)] / (1+r)^t,$$

where CE[NB(t)] is the certainty-equivalent of net benefits at time t.

Whenever the certainty-equivalent net benefits are less than the net benefits themselves, as would be the case for projects that increase income risk, the option price approach *reduces* net present value in relationship to the expected surplus approach in order to take account of risk-aversion. Occasionally, however, a project will actually decrease income risk, rather than increase it. In this case, the option price may exceed the expected surplus. For example, an environmental restoration project, or a project which lowers the risk of events (such as landslides) that lead to environmental degradation, may lower future income risks.

Whether or not the certainty-equivalent approach is usable depends upon whether or not it is appropriate to introduce an explicit utility function into the cost-benefit calculation. The simplest utility function that embeds the risk-aversion property is

$$\ln U(W) = (1-b) \ln W, \quad 0 < b < 1,$$

where W is wealth, and b (which measures the curvature of the utility function) is the Arrow-Pratt measure of relative risk-aversion.

The alternative is to fall back on the expected surplus approach. This approach can be used when all relevant prospects bear the same income risk, when the decision-maker is risk-neutral ($b = 0$), or when actuarially fair insurance is available. Moreover, when two alternative prospects have the same expected surplus, they may be ranked according to their degree of riskiness.

When individuals are risk-averse, they have clear incentives to create institutions allowing them to share, or pool, their risks. Risk-averse individuals will also diversify their asset portfolios, and prefer holding some part of any risky asset to holding the entire asset.

If individuals are risk-averse, and risks are not collective, there will be a viable market for insurance, provided that transactions costs, as measured by the resources required to write and administer insurance policies, are modest. The maximum price that the risk-averse individual will be willing to pay for insurance will exceed the pay-out that the insurance company will need, on average, to make to each individual who is insured against similar, but independent, risks. The social benefits of insurance exceed the costs of providing insurance against independent (non-collective) risks.

(f) Information, Learning and Quasi-Option Value

Quasi-option value refers to the expected value of the information that may be gained by delaying an irreversible decision. It is a value associated with reducing the potential for regret. One may ask the question: how large would quasi-option value have to be to affect the ranking of projects. Quasi-option value can be large for no development at the present time in the case of exogenous learning, and large for limited development in the case of endogenous learning, or learning-by-doing.

For example, a potential development project may have negative NPV if undertaken today. However, if the potential benefits grow at the rate g for the next n years, and if a technology that has the potential to reduce costs by a proportion x could become available in n years time, then the project may have a non-zero quasi-option value (QOV) today, given by the formula:

$$\text{QOV} = [B(1+g)^n - C(1-x)] / (1+r)^n.$$

Thus, if $B = \$100$, $C = \$200$, $x = 0.3$, $g = 0.02$, $r = 0.05$ and $n = 20$,

$$\text{QOV} = [100(1.49) - 200(0.70)] / 2.65 = 9 / 2.65 = \$3.40.$$

More generally, the present value of an environmental resource that is reserved for use in the future may also be called an *option value*. When used in this manner, option value refers to the maximum amount that a typical resource producing firm would be willing to pay today for the right to develop the resource tomorrow (i.e. in a finite number of years time). Thus, option value pertains to a marginal change in the availability of a specific class of environmental resource.

It should be noted that both quasi-option value and option value refer to values available today that may be generated by the potential use of natural resources in the future. They are therefore obliquely related to the ethic of resource conservation, where conservation is a public policy, involving investment of the social income, which seeks to increase the future availability of one or more natural resources either by reducing current consumption or by

increasing present expenditures on restoration activities. Conservation helps to keep natural capital intact. Option value approaches help to monetise the value that resource conservation provides to society. Alternatively, these approaches help to define the *user cost* associated with the consumption of specific natural resources today.

(g) Public Private Partnerships (P3s)

Public private partnerships are contractual arrangements between government and a private party for the provision of assets and the delivery of services that have traditionally been provided by the public sector. For a typical infrastructure project, four tasks are involved:

Task 1, defining and designing the project;

Task 2, financing the capital costs of the project;

Task 3, building the physical assets, e.g. the road, hospital, school, etc.; and

Task 4, operating and maintaining the assets in order to deliver the service.

The three characteristics of P3s are contracting out, private financing, and the bundling of tasks. Risk sharing and *ex ante* competitive bidding are both often features of P3s. For a P3 to be viable, one requires *both* a positive NPV for the private partner and a positive NPV from a social cost-benefit perspective. There is a double bottom line requirement.

Net transfers between the public and private sectors are a cost to one sector and a benefit to the other sector. They must enter the NPV for the private partner. However, the NPV for the private partner is an element of producer surplus that is contained within the overall cost-benefit criterion. Within the overall social cost-benefit criterion, net transfers will cancel out provided that (a) the private partner and the government sector use the same discount rate, and (b) the marginal excess tax burden associated with these transfers is small.

(h) Distributional Weights in Cost-Benefit Analysis

Please note that this topic is also summarised in Chapter 19 of the Boardman et. al. textbook.

The *ordinary cost-benefit analysis criterion* is

$$\sum \text{NPV} = \sum [B(t) - C(t)] / (1+r)^t.$$

Now suppose that there are m different social groups that are affected by the project or program being evaluated, and that the allocation of costs and benefits to these groups can be determined. Let $b(j,t)$ and $c(j,t)$ be the benefits and costs associated with group j , where $j=1 \dots m$, in time period t . One then has

$$B(t) - C(t) = [b(j,t) - c(j,t)].$$

Now define a set of welfare weights, $w(j)$, $j=1 \dots m$, such that

$$\sum_{j=1} w(j) = m.$$

Then the *weighted cost-benefit analysis criterion* is

$$\text{WNPV} = \sum_{t=0} \sum_{j=1} w(j) [b(j,t) - c(j,t)] / (1+r)^t = \sum_{j=1} w(j) \sum_{t=0} [b(j,t) - c(j,t)] / (1+r)^t .$$

The weighted CBA criterion will differ from the ordinary CBA criterion unless $w(j) = 1$ for all groups, or unless net benefits are the same for all groups, where group-specific net benefits are

$$\text{NPV}(j) = \sum_{t=0} [b(j,t) - c(j,t)] / (1+r)^t .$$

Various rules apply to the use of weighted CBA. These rules are

- (a) Ordinary CBA should always be done, because it measures economic efficiency, even if weighted CBA numbers are also reported.
- (b) The weights for disadvantaged groups should be no larger than the social cost per dollar transferred of making a distributional transfer to these groups through the regular tax-transfer system. If METB is the marginal excess tax burden that applies to income transfers, then $(1+\text{METB})$ should be the highest weight used.
- (c) If the project or program has high potential to make the disadvantaged worse off, and if $\text{NPV} > 0$, then weighted CBA should always be done because the project should probably not go ahead if $\text{WNPV} < 0$. In this case, the equity criterion would likely override the efficiency criterion.
- (d) It is a judgement call whether a program for which $\text{WNPV} > 0$, but $\text{NPV} < 0$ should go ahead. Much depends upon how different these NPVs are from zero.

Part Three: Environmental Valuation Issues and Cost-Benefit Analysis

Please note: The material that is covered in Part Three of the course can be found in Chapters 14-16 and 9 of the Boardman et. al. textbook. The most important of these chapters are Chapters 14 and 16.

(a) The Economy and the Environment

The Economy is embedded within the Environment, which can be viewed as an asset that provides a variety of services. The environment is both a *source of productive inputs* to the economy: raw materials, energy, fresh air, water, and amenities; and a *sink for non-market outputs*: solid waste, waste heat, air pollution, and water pollution.

If the flow of new pollutants, or the emissions load, exceeds the *absorptive capacity of the environment*, then the stock of pollutants in the environment will accumulate over time and lead to environmental damage. Local pollutants (like sulphur dioxide, oxides of nitrogen, and ground level ozone) should be distinguished from global emissions (like greenhouse gases such as carbon dioxide or methane). Point sources of environmental pollution (e.g. industrial plants and generating stations) should be distinguished from mobile sources (e.g. automobiles).

(b) Environmental Valuation

The template for environmental valuation ordinarily involves four classes of values. A forested area, for example, may shelter environmental values in each of these classes, although realisation of these values may well involve making trade-offs among them. The four classes of environmental values are outlined in the following table.

	Three Categories of Use Values		Non-use Values	
	<i>Consumption (or Harvest) Values</i>	<i>Amenity (+ In-stream) Values</i>	<i>Eco-system Service Function Values</i>	<i>Preservation (+ Option) Values</i>
Values based on Flows or stocks	Flow	Stock	Flow	Stock
Principal Source of Valuation Data	Direct Market Transactions	Indirect or Surrogate Markets	Indirect or Surrogate Markets	Questionnaire Surveys
Principal Valuation Techniques	Resource Rental Values	Travel Cost & Hedonic Pricing Methods	Avoided Cost Method	Contingent Valuation Method

All of the valuation techniques listed under use values are revealed preference methods, whereas contingent valuation is a stated preference method.

Consumption values refer to the harvesting or extraction of resource commodities, and are measured through direct market methods, using resource rental values for resource pricing. Where markets function properly, the resource rental value is the difference between the value of the harvested resource and the cost of harvesting. For example, for timber, the resource rental value is the difference between the delivered log price and the cost of harvesting and haulage. In essence, the resource rental value is the value of the tree on the stump (hence, stumpage value). Where market failure occurs (e.g., in open access fisheries), a shadow resource rental value should be estimated.

Amenity values, or recreational resource values, are measured through indirect market methods, such as the travel cost method or the hedonic pricing method. Most frequently, the travel cost method is used to estimate the willingness-to-pay value of recreational resources. For example, consider a small remote fishing lake as a recreational destination. By survey methods, one may be able to estimate how much a representative user is willing to pay to travel to the lake, and multiply this estimate by the number of potential users of the fishing lake amenity. Essentially, one goes to the market for travel services to estimate the social value of the recreational resource. In the case of water resources, amenity values ordinarily pertain to in-stream activities.

The hedonic pricing method can be used to estimate externality benefits and costs as well as being an alternative to the travel cost method for estimating amenity values. It is an indirect market method that can be used to estimate the willingness-to-pay for a desirable

characteristic or attribute, or to avoid an undesirable characteristic or attribute. Hedonic regressions usually relate a market price (such as an asset value) to a reasonably comprehensive set of attributes or characteristics, so as to remove the impact of influences on the asset value other than the particular attribute or characteristic whose impact one wishes to isolate. The technique therefore involves the estimation of a multiple regression equation, often specified in log-linear form, and ordinarily using cross-section data.

Eco-system service values are provided by the ecological system to the economic process at no explicit charge to human beings; e.g., the provision of water supplies from clean-flowing mountain streams. The avoided cost method can be used to measure (or at least provide a lower bound to) the value of an eco-system service function and/or the value of a restorative or preventative investment. For example, the net benefits from undertaking an investment in environmental restoration/preservation include the costs of cleaning up and mitigating the effects of an adverse event, such as a landslide, should such an event occur. There may be additional benefits which are not easy to measure.

In addition to the avoided cost method (and its close relative, the defensive and mitigative expenditure method), other techniques which are sometimes used to measure eco-system service function values include production function methods and, occasionally, hedonic pricing methods. Benefit transfer values (shadow prices from secondary sources) are, upon occasion, also used.

Double-counting of resource values can occur because some eco-system services functions support consumptive uses (e.g., provision of clean drinking water) and non-consumptive uses (e.g., wildlife viewing), by providing inputs for these end-products. However, other eco-system service functions (e.g., carbon sequestration, waste absorption within the environment's carrying capacity, etc.) provide values that are not counted elsewhere. In addition, where consumptive or recreational uses are given inappropriately low values (e.g., if clean water is treated as a free good), there is a residual shadow price value to eco-system services.

Preservation values may include existence value, bequest value, and option value components. Existence values are public good values that are intrinsic to some environmental resources even if these resources are not being used. For example, spotted owls possess existence value. Bequest values relate to the willingness to pay to leave environmental resources intact for future generations. There may be both existence and bequest elements in the cultural values which First Nations may place upon environmental resources. Generally speaking, there are no market-based methods that work for measuring preservation values. Preservation values are ordinarily estimated through questionnairebased stated preference methods, rather than market-based revealed preference methods. The main stated preference method is called the contingent valuation method. Option values, which are associated with preserving natural resource assets for future use, have been previously discussed.

(c) Valuation of Recreational Amenities: Travel Cost and Hedonic Pricing Methods

As previously stated, amenity values, or recreational resource values, are measured through indirect market methods, such as the travel cost method or the hedonic pricing method.

Most frequently, the *travel cost method* is used to estimate the willingness-to-pay value of recreational resources. The simplest version of the travel cost method is the zonal travel cost approach, which requires information on the number of visits to the recreational destination from a variety of different origins, or zones, which lie at different distances from the recreational destination. Since travel costs (including both out-of-pocket costs and time costs) generally increase with distance, one may estimate a *trip generating function* in which the visitation rate (number of visitors to a recreational destination, or number of participants in a recreational activity, generated by each origin zone relative to the population of that zone) is assumed to be a function of (a) travel costs from the origin zone to the recreational destination, (b) zonal income, and (c) zonal household demographics. One may then use the trip generating function as a proxy for the demand function for the amenities provided by the recreational site, from which one may estimate the consumer surplus, or net amenity value, associated with the site. (*See further below.*)

The *hedonic pricing method* can be used to estimate externality benefits and costs as well as being an alternative to the travel cost method for estimating amenity values. It is an indirect market method that can be used to estimate the willingness-to-pay for a desirable characteristic or attribute, or to avoid an undesirable characteristic or attribute. For example, if one wanted to isolate the impact of an amenity value such as the quality of a scenic view, the extent of aircraft noise, or local air quality, one might run a cross-sectional multiple regression in which housing prices (P) are regressed on (a) the distance from the central business district, (b) a measure of house size, (c) various neighbourhood characteristics, as well as (d) the variable of interest (an indicator of scenic view, aircraft noise, or air quality). Usually, the regression takes a convenient log-linear form:

$$\ln P = \beta_0 + \beta_1 \ln \text{CBD} + \beta_2 \ln \text{SIZE} + \beta_3 \ln \text{NBHD} + \beta_4 \ln V,$$

where V is the variable of interest. A large sample size is often required to reduce the problem of multi-collinearity.

The regression coefficient, β_4 , measures the partial elasticity, $d \ln P / d \ln V$, estimated after controlling for the influence of other variables on housing prices. The partial elasticity may be used to construct the marginal hedonic price of the variable of interest (be it scenic views, aircraft noise, or air quality). The marginal hedonic price, $h(V)$, is given by $dP/dV = \beta_4 P/V = h(V)$. The graph of $h(V)$ against V represents the demand curve for the characteristic or attribute of interest. Areas under the demand curve are estimates of the willingness-to-pay for this characteristic.

Sometimes a second step is attempted in order to control for differences in incomes and tastes across the households that own the properties whose prices have been used in the hedonic regression, by further regressing $h(V)$ on V and household income, etc. While the first step controlled for other housing characteristics, the second step would control for household characteristics. However, the second step is usually omitted in the hedonic pricing method.

In order to avoid biases from omitted variables, the hedonic pricing method may also be used to estimate variables such as the statistical value of human life. In this context, a typical hedonic regression equation would take the form:

$$\ln(\text{wage rate}) = \beta_0 + \beta_1 \ln(\text{fatality risk}) + \beta_2 \ln(\text{injury risk}) + \beta_3 \ln(\text{job tenure}) + \beta_4 \ln(\text{education}) + \beta_5 \ln(\text{age}).$$

Due to the problem of multi-collinearity, a relatively large sample of cross-sectional data may be required to generate reasonably precise estimates of coefficients such as β_1 .

(d) More or recreational resources and the travel cost method

As previously stated, amenity values, or recreational resource values, can be measured through indirect market methods. Most frequently, the *travel cost method* is used to estimate the willingness-to-pay (WTP) value of recreational resources. The simplest version of the travel cost method is the zonal travel cost approach, which requires information on the number of visits to (or recreational days spent at) the destination site by people coming from a variety of different zones of origin which lie at different distances from the destination site. Since travel costs (including both out-of-pocket costs and time costs) generally increase with distance, one may estimate a *trip generating function* in which the *visitation rate* is assumed to be a function of (a) travel costs from the zone of origin to the destination site, (b) zonal income, and (c) zonal household demographics. The visitation rate is specified as the number of visitors to a recreational destination (or the number of participants in a recreational activity) generated by each zone of origin, relative to the population of that zone. One may then use the trip generating function as a proxy for the demand function for the amenities provided by the recreational site, from which one estimates the consumer surplus, or net amenity value, associated with the site.

If the zonal travel cost approach is to be viable, the following data, at a minimum, are required to estimate the trip generating function, and thus the demand function for the services of a particular recreational resource site: (a) the number of visits from each zone of origin, (b) demographic information about the people living in each zone, (c) the distance in kilometres from each zone to the recreational destination, (d) the travel costs incurred per round-trip kilometre, and (e) the value of time spent travelling, or the opportunity cost of travel time. Data pertaining to the availability of substitute sites are sometimes also incorporated into the analysis.

The trip generating function which serves as a proxy for the demand function for the amenities provided by the recreational site might be specified as a cross-sectional log-linear regression equation of the form:

$$\ln VR(i) = \beta_0 + \beta_1 \ln TC(i) + \beta_2 \ln HI(i) + \beta_3 \ln YP(i) + \dots, \text{ for } i = 1 \dots m,$$

where $VR(i)$ is the visitation rate for i -th zone of origin (number of visits to the recreation site as a ratio to zonal population), $TC(i)$ is the travel cost estimate for visits from the i -th zone to the recreation site, $HI(i)$ is average household income for the i -th zone, and $YP(i)$ might be the proportion of young people in the total population of the i -th zone. Within this regression equation, zonal population has been used to normalise the number of visits from each zone in order to get around a statistical problem known as heteroscedasticity. The number of recreational days consumed at the recreational site is the number of visits multiplied by the average number of recreational days spent per visit.

For such a regression equation to provide a reliable estimate of the elasticity of demand for the services of the recreational site, which is given by the estimated numerical value of the regression coefficient, β_1 , there needs to be both a sufficient number of observations (zones of origin) and sufficient variance in travel costs among the different zones of origin.

This may be difficult to achieve, for example, in the Sea to Sky (S2S) Land and Resource Use Management Plan (LRMP) area because travel costs from various parts of the Greater Vancouver area (which generates the greatest number of recreational visits to almost any given recreational destination in the S2S area) are quite similar. In principle, of course, travel cost estimates should include a time cost element as well as a distance-related cash cost element. Nevertheless, if the zonal travel cost variable does not have sufficient variance to provide a reliable estimate of the elasticity of demand, then the individual participant response data derived from a travel cost sample survey, rather than the zonal average response data, will need to be used. This would imply the use of the individual, rather than the zonal, travel cost method.

If site characteristics are an important determinant of the value of the overall recreational experience, the travel cost method can be augmented into a hedonic pricing approach known as the hedonic travel cost method. This approach requires travel cost data that pertain to a considerable number of recreational sites, all of which have the potential to provide similar recreational services. For example, white-water streams with a variety of different characteristics can substitute for each other in the context of kayaking activity. Alternative hiking trails can provide differential access to scenic views and wildlife viewing opportunities. The following section describes the hedonic travel cost method.

(e) Hedonic travel cost methods

The problem with the application of the hedonic pricing method to the characteristics of alternative recreational sites is that one does not have direct information on site values (or prices, P). Apart from situations where the fees associated with the maintenance of licensed or leasehold tenures are based upon appraised (or assessed) land values, there are no equivalents to property prices available. However, if one has completed travel cost surveys for a number of alternative recreational sites, all of which provide the opportunity for similar adventure tourism and recreation (ATR) activities, one can integrate the data across these site-specific surveys and run a hedonic travel cost regression, usually specified in log-linear form.

The left hand side variable in the regression equation would continue to be the visitation rate to each destination site from each zone of origin, $VR(i,j)$. The right hand side variables include travel costs incurred from each zone of origin to each destination site, $TC(i,j)$, zonal average household income, $HI(i)$, zonal household demographics, $YP(i)$, and destination site characteristics that differ from one site to another, $SC1(j)$, $SC2(j)$, and $SC3(j)$, etc. Thus, $SC1(j)$, $SC2(j)$, and $SC3(j)$, etc., are site characteristics indicators. Other things being equal (including estimated travel costs), better destination sites should be associated with larger visitation rates (and more recreational days spent) than less attractive destination sites.

The resulting regression equation would have the form:

$$\ln VR(i,j) = \beta_0 + \beta_1 \ln TC(i,j) + \beta_2 \ln HI(i) + \beta_3 \ln YP(i) + \beta_4 \ln SC1(j) + \beta_5 \ln SC2(j) + \beta_6 \ln SC3(j) + \dots, \text{ for zones } i = 1 \dots m, \text{ and sites } j = 1 \dots n,$$

where the basic variables are described above. This cross-sectional regression equation has the same form as a trip generating function. However, the differing characteristics of the alternative recreational sites are also allowed to affect visitation rates. Thus, both zonal push factors (proximity effects) and destination pull factors are allowed to affect visitation rates.

It follows that the hedonic travel cost regression equation has the same basic format as the gravity model used to explain trade flow volumes among countries. In the gravity model, distance between countries normally has a negative effect on trade flow volumes, while in the hedonic travel cost model, travel costs as a function of distance are expected to have a negative effect on visitation rates, that is, β_1 is expected to be negative. Again, the estimated numerical value of the regression coefficient, β_1 , is an estimate of the elasticity of demand for the services of the specific type of recreational destination site to which the regression equation applies.

A possible configuration of the site characteristics data that are assembled into the SC variables would be data which rank order the various recreation sites by their ability to provide key amenity features, probably no more than three or four in number. The resulting matrix of site characteristics data would be of size n by k if there are k amenity features that could affect the relative popularity of visits to the alternative recreational sites, where k needs to be considerably smaller than n . For example, the rankings for the first site characteristic would constitute the variable, $SC1(j)$, and would appear as the first column of the site characteristics matrix. Separate matrices could be constructed for amenity features (or site characteristics) that support winter season and summer season ATR activities, and separate regression equations could be run using segregated winter season and summer season data. Whether or not this was done, the estimated coefficients, β_4 , β_5 , and β_6 , etc., would measure the importance of key amenity features in the explanation of the pattern of visitation rates across the various recreation sites.

The geographic area containing the alternative recreational sites for which travel cost data are combined into a single hedonic travel cost regression equation may, or may not, coincide with a given LRMP area. Within the geographic area, some of the sites may provide amenity services that are supportive of more than one ATR activity. With a suitably designed travel cost survey, it should be possible to quantify the total number of recreational days enjoyed in the pursuit of each separate ATR activity that are associated with the overall site visits to each recreational site from each point of origin. A separate left hand side variable could then be constructed for each major ATR activity, and a representative hedonic travel cost regression equation could then be estimated for each activity.

In this way, an overall hedonic travel cost methodology that facilitates the estimation of the overall elasticity of demand for the amenity services of the recreational sites within a general geographic area could be transformed into an approach that facilitates the estimation of the elasticity of demand for each major ATR activity that is pursued within

the same geographic area. This is important if what one needs to measure are the consumer surplus and producer surplus values that are associated with each ATR activity, rather than with overall recreational site amenities. The particular site characteristics variables that would be used in the separate regression equations would, in all probability, differ from one ATR activity to another. It is possible that zonal household demographics variables should also be re-specified from one ATR activity to another.

(f) Valuation of Eco-System Services: Avoided Cost Method

As stated previously, the avoided cost method can be used to measure (or at least provide a lower bound to) the value of an eco-system service function and/or the value of a restorative or preventative investment. The net benefits from undertaking an investment in environmental restoration/preservation include the costs of cleaning up and mitigating the effects of an adverse event, such as a landslide, should such an event occur. There may be additional benefits which are not easy to measure.

If the probability of an adverse event is p , then the expected loss, $E(L)$, from the adverse event is pL , where L represents the clean-up and mitigation costs that would be triggered by the occurrence of the adverse event. Because the probability of the adverse event is unknown, $E(L)$ may be hard to estimate. If $E(L)$ is under-estimated, one may fail to undertake the investment when it would have been desirable (Type One error). If $E(L)$ is over-estimated, one may undertake the investment when it turns out to have been unnecessary (Type Two error).

For example, suppose that the adverse event is a landslide, which may occur with probability p , and which, if it occurred, would destroy the eco-system services of a clean-flowing watershed. How do we place a value on the potential loss of these services? The avoided cost method would attempt to estimate $E(L) = pL$, and to use $E(L)$ as the estimate of these benefits, where L might include the costs of developing alternative water sources and/or installing water treatment facilities. This is likely to provide a lower bound estimate since it may fail to include in-stream benefits (e.g., to a fishery) along with the consumptive benefits of clean water flows. It may also fail to account for risk aversion since it uses the expected loss, $E(L)$, rather than a certainty equivalent. Indeed, if the decision-maker is risk-averse and insurance against the adverse event is unavailable, one should, in principle, replace $E(L)$ with a certainty equivalent which exceeds $E(L)$. As a result, one may want to make the investment even if its cost somewhat exceeds $E(L)$. This is the precautionary principle.

Expected wealth, $E(w)$, equals $w - pL$, where w is existing wealth. The certainty equivalent (CE) is calculated from an expected utility approach where

$$E(u) = (1-p) U(w) + p U(w-L).$$

Certainty equivalent wealth is that amount of wealth with certainty that provides the utility level $E(u)$, and is less than $E(w) = w - pL$ when the decision-maker is risk-averse. Thus, $pL = w - E(w)$ is a lower bound estimate of willingness-to-pay, and the proper measure of willingness-to-pay is $w - CE$. However, the expected utility approach is only as usable as the decision-maker's willingness to specify his/her degree of risk aversion, as embodied in a utility function. As a result, one ordinarily falls back on the expected loss approach.

(g) Preservation Values and the Contingent Valuation Method

Preservation values (or non-use values, which include existence, bequest and cultural values) can be estimated by the contingent valuation method, although this method is not confined to the measurement of non-use values. However, there is generally no way to cross-check the reliability of estimates of preservation values (as opposed to use values) based upon contingent valuation questionnaires with alternative estimates based upon revealed preference methods. Thus, if the contingent valuation method is perceived to be too unreliable and/or too expensive to pursue, preservation values have to be brought into resource valuation in a qualitative way. This could involve a multiple accounts and/or multiple attribute approach.

The contingent valuation method is often associated with a number of sources of bias: (a) hypotheticality (strategic and non-commitment) bias, (b) neutrality (context of question) bias, (c) sampling (non-response) bias, and (d) warm glow (moral satisfaction) bias. There are also issues associated with the possible double-counting of use and non-use values, and with the nature of the survey instrument. Dichotomous choice or referendum methods are preferred. However, if monetisation is eschewed, multiple accounts approaches in which preservation values enter qualitatively should be used.

Because there are several difficulties associated with the contingent valuation method, economists often fall back on qualitative analysis. For an environmental restoration project whose net present value (NPV) before considering preservation values is negative, one should ask the question: how large would the impact of the project have to be from the perspective of preservation values to tip the NPV into positive territory? If the value seems reasonable, proceed with the project; otherwise don't. For resource-using rather than conservation-related projects whose NPV before considering preservation values is positive, one should ask the question: how large would the preservation value costs have to be to tip the NPV into negative territory? If this magnitude is within reasonable bounds, do not proceed with the project. However, if the potential preservation value consequences would need to be unreasonably large to tip the NPV to negativity, the project could go ahead, although the precautionary *safe minimum standards* approach might not lead to this conclusion.

There may be other, less traditional, methods for assessing preservation values than the contingent valuation method and/or qualitative analysis. One of these methods involves option value approaches, although these approaches are limited to measuring future use values, as opposed to the existence and bequest value elements within preservation values.

(h) Benefit transfer: Shadow prices from secondary sources

Due to the cost of conducting original surveys, the benefit transfer method has become widely used in cost-benefit analysis studies. The benefit transfer method takes estimates of variables such as shadow prices, demand elasticities, and supply elasticities that have been generated in previous empirical studies and uses them in the context of the cost-benefit study being undertaken. The study site is mined for relevant estimates, and these estimates are loaned to the policy site, where cost benefit analysis work is currently underway. Benefit transfer is, therefore, a technique that tries to avoid the need to re-invent the wheel.

Benefit transfer values can only be as accurate as the initial benefit estimates. Moreover, differences between the study site and the policy site with respect to incomes, tastes, and other socio-economic factors, and with respect to geographic location, temporal factors, project characteristics, and welfare measures used, may give rise to a number of inconsistencies when benefit values are transferred. Ordinarily, the variables that are transferred from the study site to the policy site are shadow price estimates and, in particular, measures of willingness-to-pay or consumer surplus. However, estimates of functional values, such as demand and supply elasticities, are often imported from a relevant study site.

The reliability of benefit transfer exercises depends quite crucially on the comparability or similarity of the study site situation and the policy site situation. Some of the criteria that should govern the use of benefit transfer values include: (a) site similarity is normally a pre-requisite; (b) study values transferred between geographically removed sites (e.g. across continents) are unlikely to be reliable; and (c) the quantity, volume, or population numbers which may be used as multipliers of transferred shadow prices need to be consistent with these shadow prices. If these criteria are not met, serious errors can creep into the use of benefit transfer values.

(i) Externalities and Regulatory Instruments

When pollution externalities result from commodity production, market failure involves

- (a) excess production of both the commodity (goods) and the associated pollution (bads);
- (b) under-pricing of the products that are responsible for pollution;
- (c) insufficient market incentives for producers to search for ways to yield less pollution per unit of output (or to reduce emissions-intensity); and
- (d) insufficient recycling and/or reuse of the polluting substances since release into the environment is inefficiently cheap.

Solutions to negative externality problems include *internalising* the externality by

- (a) altering people's preferences towards so-called *green* goods;
- (b) applying quantitative controls and regulatory standards;
- (c) applying effluent charges, or pollution taxes;
- (d) creating transferable emissions permits while limiting their available quantity; and/or
- (e) redesigning property rights institutions.

Of these solutions, (a) may result from the activities of environmental non-governmental organisations (ENGOS), (b) involves the use of command and control instruments, (c) and (d) are referred to as market-based instruments, while (e) involves consideration of the *Coase theorem* and transactions costs.

(j) Internalising Externalities and the Polluter Pay Principle: Three Key Questions

- (a) What is the appropriate level of waste, effluent or emissions flow in relationship to the absorptive capacity of the environment?
- (b) When reductions in waste, effluent or emission flows are required, how should the responsibility for achieving this lower flow level be allocated among the various sources of the pollutant?

- (c) Which regulatory instruments are most likely to achieve appropriate levels of waste, effluent or emission flow reductions in a cost-effective and/or efficient manner?

The benefits associated with emission reductions take the form of avoided environmental damage. These benefits increase as more emission reduction (or abatement) occurs. Thus, if R is the overall amount of emission reduction, we may write the benefit function as $B(R)$. These benefits increase as more abatement occurs, so that $dB(R)/dR > 0$. However, marginal avoided damages ($dB(R)/dR$), although positive, will tend to fall as larger amounts of emission reduction occur, and especially as the residual volume of effluent becomes small. It follows that the marginal avoided damages relationship falls as emission reductions rise.

The costs associated with emission reductions may be referred to as abatement costs. These costs increase as more abatement occurs. Thus, if R is the overall amount of emission reduction, we may write the cost function as $C(R)$. These costs increase as more abatement occurs, so that $dC(R)/dR > 0$. However, marginal abatement costs ($dC(R)/dR$) will tend to increase as larger amounts of emission reduction are undertaken. It follows that the marginal abatement cost relationship rises as emission reductions rise. This relationship may be fairly flat for low levels of abatement, and become steep as increased abatement occurs. Indeed, the marginal abatement cost curve may become vertical well before complete emissions control is approached.

The maximisation of the net benefits from emission reductions requires that $NB(R) = B(R) - C(R)$ be maximised. This occurs at emission reduction level R^* , where marginal benefits are equated to marginal costs, $dNB(R)/dR = dB(R)/dR - dC(R)/dR = 0$, or $dB(R)/dR = dC(R)/dR$. The emission reduction level that maximises the net benefits from abatement activity can be seen in the standard diagram (in which the marginal avoided damages and the marginal abatement cost relationships are both graphed with emission reductions on the horizontal axis, and in which their intersection determines the efficient level of abatement activity, R^*) also to minimise the sum of abatement costs and residual damage costs.

The efficient level of effluent production is rarely, if ever, equal to zero, because to reduce pollutant flows to zero would ordinarily be too expensive in relationship to the gain from eradicating the last unit of pollutant flow. The analysis also suggests that R^* may differ from one local area to the next depending upon differences in absorptive capacity and in the costs of environmental damage. Atmospheric emissions into a crowded urban air-shed may have more extensive health effects than those emitted in sparsely populated regions. Indeed, BC Hydro uses the following cost function adjustment for emissions associated with its thermal generating plants: $1.0 + 0.9 \times (\text{population in } 100,000\text{s within } 100 \text{ km of the emission source})$. This adjustment is applied to marginal damage cost estimates (largely based on health effects) for pollutants such as oxides of nitrogen, sulphur dioxide, and micro-particulates.

(k) Command and Control Instruments

We now consider two polar forms of pollution control mechanisms, including

- (a) regulated environmental standards, or quantitative command and control instruments, which sometimes take the form of mandatory abatement technologies and

associated equipment installation (e.g., selective catalytic reduction technologies), and

(b) effluent charges (or environmental taxes), and tradable emission permits (or emission reduction credits), both of which are examples of market-based instruments.

Essentially, quantitative command and control instruments are neither cost-effective, nor efficient. They do not allocate the responsibility for emission reductions across firms in a cost-effective manner, nor do they lead to an efficient level of emission reductions, except by chance. Indeed, the cost of achieving a given level of emission reductions will be minimised if and only if marginal abatement costs are equalised for all emitting firms (or all emission sources). Cost-effectiveness can be achieved by the substitution of either effluent charges or tradable emissions permits for quantitative command and control restrictions. However, efficiency may be hard to achieve without knowledge of the shape of the marginal avoided damages relationship.

Effluent charges try to control abatement effort by controlling a key price (the per unit tax on emissions), while tradable emissions permits control the overall volume of abatement effort by limiting total emissions and letting the market determine the per unit price of emissions permits. A tradable emissions permit system requires (a) an overall target for emissions to be established, (b) a system for allocating initial target shares, and (c) the creation of a property right to these shares that is freely tradable.

The price-quantity relationship for effluent charges and tradable emissions permits goes in opposite directions, but both instruments have the potential to be cost-effective, even if not also efficient in the static sense. However, both effluent charges and tradable emission permits provide an incentive for firms to find innovative ways of reducing abatement costs. Thus, they share a *dynamic efficiency* property in a manner which command and control regulatory instruments do not share.

(I) Cost Effective Approaches to Environmental Protection

Let E and F be the pre-abatement volumes, in tonnes, of atmospheric emissions generated by firms X and Y , respectively. Let $C(A)$ and $C(B)$ be the total expenditures incurred by firms X and Y , respectively, on activities that would abate their respective emissions by A and B tonnes. For both firms X and Y , marginal abatement costs [$C'(A)$ and $C'(B)$] are positive and increasing in their respective arguments, indicating that diminishing returns set in as abatement effort increases.

Assuming that there are no free allocations of emission reduction credits, $E - A$ and $F - B$ are the emissions volumes for which purchased credits are required. If p is the unit price of emission reduction credits, then $p[E - A]$ and $p[F - B]$ are the costs incurred respectively by firms X and Y to purchase these credits.

The total emissions costs incurred are equal to

$$T(X) = p[E - A] + C(A) \text{ for firm } X, \text{ and}$$

$$T(Y) = p[F - B] + C(B) \text{ for firm } Y.$$

The total savings for firm X which are associated with not having to purchase emission

reduction credits for A tonnes of emissions are equal to $S(X) = pA - C(A)$, while the total savings for firm Y which are associated with not having to purchase emission reduction credits for B tonnes of emissions are equal to $S(Y) = pB - C(B)$. Notice that $T(X) = pE - S(X)$, and $T(Y) = pF - S(Y)$.

The problem for firm X is to choose A such that $T(X)$ is minimized and, correspondingly, $S(X)$ is maximized, while the parallel problem for firm Y is to choose B such that $T(Y)$ is minimized and, correspondingly, $S(Y)$ is maximized. The solution to these problems requires firm X to set A such that $C'(A) = p$, where $C'(A)$ represents firm X's marginal abatement costs, and firm Y to set B such that $C'(B) = p$, where $C'(B)$ represents firm Y's marginal abatement costs. Notice that, as p increases, both T and S will increase, creating an increasing incentive for investments in abatement technologies.

Thus, when firms are faced with a choice between reducing their atmospheric emissions or purchasing emission reduction credits, cost-effectiveness requires the volume of abated emissions to be determined by equating the cost of the last unit of abatement achieved (the marginal abatement cost) to the externally given price of emission reduction credits. Moreover, it is essential for cost-effectiveness that all emitting firms face the same external price of emission reduction credits. The responsibility for abatement activity is not allocated across firms X and Y in a cost-effective manner unless $C'(A) = C'(B)$, and this requires both firms to face the same market price (p) of emission reduction credits.

(m) Local Atmospheric Emissions

Nitrogen oxides (NO_x), sulphur dioxide (SO₂), volatile organic compounds (VOCs), particulate matter, and ground level ozone, etc., are all pollutants that affect the ambient quality of local air-sheds. Both mobile sources (e.g., automobiles and trucks) and point sources (e.g., thermal generating plants) contribute to the pollution of local air-sheds, and to the phenomenon of acid rain. In Canada, the sources of NO_x emissions are transportation (52.4%), industrial sources (25.2%), non-industrial fuel combustion (13.5%), and open sources (8.9%). Ground level ozone (which is produced by chemical reactions from combining NO_x, VOCs, and sunlight) is an increasing problem in many urban areas. Thermal generating plants are an important source of SO₂ emissions.

The health effects of local atmospheric emissions are dependent upon the population density of the metropolitan area, and the location of point sources of pollutants. The age of both industrial facilities and automobiles affects the amount of effluent they release. In several jurisdictions, new automobile sales must obey, on a fleet basis, low emission vehicle (LEV) restrictions.

Ambient air quality standards are ordinarily measured in milligrams per cubic metre, or in parts per million. With ambient air quality standards, the extent to which the flow of new discharges needs to be controlled will depend upon existing concentrations, which themselves depend upon weather conditions. There are both time and space dimensions to the control of local atmospheric emissions, and to the nature of control mechanisms. Competitiveness effects may also be important. Although local residents might be willing to trade-off some amount of industrial activity (and associated job potential) for improved ambient air quality, it is not clear that any such trade-off exists for reduced greenhouse gas

emissions by local industries.

Local atmospheric emissions should be distinguished from global emissions of greenhouse gases: carbon dioxide (CO₂), methane (CH₄), and nitrous oxide (N₂O), which contribute to global warming. Emissions of greenhouse gases are ordinarily measured in tones of CO₂-equivalent. Methane emissions are multiplied by 21 to obtain CO₂-equivalency, while nitrous oxide emissions are multiplied by 310. Per tonne, these emissions have more substantial and long-lasting atmospheric effects than CO₂ emissions.

(n) Uncertain Information and Governing Instrument Choice

Tradable emissions permits are generally a more flexible instrument than effluent charges in response to external data changes such as new sources of pollution, cost inflation, and technological change, which would affect the position of the marginal abatement cost relationship when aggregated across effluent sources. Directly limiting the total quantity of available credits (or permits) and allowing the market to set the price of these credits allows less room for error on the emissions side than setting the price of credits (or, equivalently, the scale of effluent charges) and letting the market determine the amount of abatement activity (and thus the residual effluent volumes) that occurs.

Uncertainty about the shape and position of the marginal avoided damages relationship not only means that efficient outcomes (as distinct from to cost-effective outcomes) are unlikely to be achieved, but also gives rise to consideration of the costs of being wrong. Under conditions of uncertainty (but within the context of a cost-effective control instrument), setting some overall quantitative standard that maintains an acceptable level of risk may be the most appropriate *second best* decision that can be made. Essentially, one uses the principle of *minimising the potential for regret*, or the *safe minimum standards* approach.

(o) Emissions Trading Systems

Command and control frameworks, including mandated control technologies and regulated emissions standards, may be contrasted with market-based instruments such as effluent charges (and other environmental tax shifting arrangements), and emissions trading programs (involving emission reduction credits and the ability to purchase offsets). Emissions trading systems (or cap and trade systems) use a currency called an emissions reduction credit. Should any point source emitter decide to control its emissions to a degree higher than necessary to fulfil its legal obligation to meet an emissions standard set by the regulatory authority, it can apply to this control authority for certification of excess control as an emissions reduction credit. To receive certification, the emissions reduction must be quantifiable, surplus, enforceable, and permanent. Once the credits are certified, other point source emitters may purchase these credits to provide offsets to any emissions that may exceed the standard.

Emissions trading reduces overall compliance costs and is thereby cost-effective. Emissions trading allows for the separation of who will pay for the control (internalization) from who will install the control equipment (ordinarily the source with lowest abatement costs). Emission reduction credits work best either where they apply within a specific local air-shed, or where the location of emissions is unimportant to overall damage costs (e.g. the

creation of acid rain for a whole region).

Both emissions trading systems and environmental tax-shifting systems involving effluent charges are consistent with the equi-marginal cost-effectiveness principle. Emissions trading systems set the overall quantitative emissions reduction target and let the market set the price of tradable emissions reduction credits. Effluent charges set the price for emissions and let the market determine the emissions reduction volume.

The issue of how the initial allocation of emission reduction credits (or emission permits) is established is important. The allocation can be based upon historic emission profiles, upon existing output volumes, through the auctioning of these credits, or by some combination of these mechanisms. The third of these methods compensates the public sector generally for the allowable level of emissions, whereas the first two are obviously less costly for the affected firms. The second method is preferable to the first because it does not penalise individual firm growth; however, it requires some form of claw-back mechanism to be put in place.

Emission trading systems require some means of determining the initial allocation of emission permit quotas and/or emission reduction credits. Various possibilities are outlined in the following table, where n = emission-intensity norm for all sources/firms, e = actual emission-intensity of source/firm, Q = firm output, and gQ = growth in firm output.

	purchased permits (existing output)	purchased permits (increased output)
all permits auctioned	eQ	$e(gQ)$
gratis permits based on predetermined history	$(e-n)Q$	$e(gQ)$
gratis payments based upon current output level	$(e-n)Q$	$(e-n)(gQ)$

The third of these systems does not penalise output growth, and would definitely be preferred by growing firms. However, to become operational, the third possibility requires a claw-back system, which reduces n over time, to be established, if overall emissions are to be controlled.

Part Four: Modelling the Urban Economy

Please note: Many of the topics covered in Part Four do not appear in the Boardman et. al. textbook, although some of them are covered in Chapter 16.

(a) The Central Business District and Residential Location Decisions

The traditional model of urban spatial structure contemplates a circular city in which work activities take place in the *central business district* (CBD), and households spread themselves over the remainder of the city's land mass. Commuting costs are assumed to be proportional to the distance between the place of residence and the CBD. In equilibrium, where there is no incentive for households to move their places of residence, household

utility must be the same regardless of household location. As a result, the rental price of housing must fall with distance from the CBD to compensate households for higher commuting costs.

The utility maximisation problem of the commuting household may be modelled as follows:

Maximise $U = U(Z, H, L)$,

where Z = consumption goods,
 H = housing services, and
 L = leisure hours,

subject to a budget constraint of the form:

Expenditure (E) = $Z + pH + tX = w(T - L - mX) = \text{Income } (Y)$,

where p = price of housing services per working day,
 X = radial distance in kilometres from residence to CBD,
 t = out-of-pocket travel cost per round-trip kilometre to CBD,
 m = travel time in hours per round-trip kilometre to CBD,
 T = total available time in hours in a standard day, and
 w = net hourly wage rate.

Notice that working hours are equal to $(T - L - mX)$, and that the price of consumption goods is taken to be the numeraire.

The first-order conditions for utility maximisation may be obtained by differentiating the Lagrangian expression:

$$F = U(Z, H, L) - q [Z + pH + tX - w(T - L - mX)],$$

with respect to Z , H , and L , obtaining $U'(Z) = q$ and, therefore, $U'(H)/U'(Z) = p$ and $U'(L)/U'(Z) = w$. Maximisation requires that, between each pair of goods (including leisure hours), marginal utility ratios be equated to relative prices.

We now look for residential locations among which the commuting household is *indifferent*, or

$$dU/dX = U'(Z) dZ/dX + U'(H) dH/dX + U'(L) dL/dX = 0.$$

Given the first order conditions, this is equivalent to looking for locations such that

$$dZ/dX + p dH/dX + w dL/dX = 0.$$

But from the budget constraint, we also know that

$$dZ/dX + H dp/dX + p dH/dX + t + w dL/dX + wm = 0.$$

By subtraction, it follows immediately that

$$H dp/dX = - (t + wm), \text{ or } dp/dX = - (t + wm) / H.$$

(b) The Rent-Gradient Equation

This is the first important result of the Mills-Muth-Alonso-Henderson model of urban spatial structure, the intensity of land use and household location decisions. The equation says that there will be a *rent gradient* which relates the price of housing services to the distance between a commuting household's place of residence and the central business district (CBD), where it is assumed that the commuter works. The slope of the rent gradient relationship is equal to $-(t + wm)/H$, where $(t + wm)$ is the full cost per round-trip kilometre of commuting to and from work. The travel time portion of this is w .

Muth reaffirms (on p. 200) the central rent-gradient result of the traditional mono-centric model. "To a first approximation, the condition for utility to be constant with a slight move in any direction is that real income be constant. The latter requires that the change in the expenditure necessary to rent a given dwelling just offset the change in commuting cost incurred, or that the percentage change in the rental of a given dwelling be numerically equal, but opposite in sign, to the ratio of the change in commuting costs to housing expenditures." This statement may be expressed by the rent gradient formula: $H dp/dX = -(t+wm)$.

Muth then reviews the properties of the mono-centric model of urban spatial structure, while also considering extensions to the radial city and other deformations of Euclidean space, arguing that one can still work with constant cost commuting contours. He also discusses the clustering of similar land uses due to both market forces and zoning regulations. He considers endogenous transport costs and congestion in the context of the mono-centric city, and argues that building more roads to relieve traffic congestion leads to excessive allocation of scarce land to roadways.

(The Muth reference can be found in chapter 13 of in Banister, Button and Nijkamp (eds.), *Environment, Land Use and Urban Policy*, Northampton, Elgar, 1999.)

(c) The Market for Housing Services: Implications for Urban Density

As explained, the price of a given quantity of housing services declines with distance from the city's central business district (CBD). This phenomenon, in which households trade off commuting costs against housing costs, is known as the rent gradient. The rent gradient phenomenon also gives rise to differences in the structural density of housing (dwelling units per hectare) across an urban area. However, differences in land scarcity will place different urban areas on different rent gradient relationships. The greater land scarcity of large cities will generate rent gradients that lie outside those of smaller cities (e.g. Tokyo vs. Nagasaki, and Vancouver vs. Regina). A city with high transportation costs will have a steeper rent gradient than a city with low transportation costs.

When coupled with a model of competitive housing supply, the Mills-Muth-Alonso-Henderson model also has the following predictions: land rent per hectare, the structural density of housing (house size to land size ratio), and population density are all decreasing functions of distance to the CBD, while dwelling size itself is an increasing function of distance to the CBD.

To illustrate, profits (V) earned by suppliers of housing services may be written as

$$V = N [p f(s) - q s - r],$$

where p = price of housing services,

s = structural density ratio, or the ratio of building input to land input in the housing services production function, $H = N f(s)$, which is assumed to exhibit no scale economies or diseconomies,

N = land input into the housing services production function,

q = unit cost of building structures, and

r = unit rental cost of land.

Profit maximisation implies $p f'(s) = q$, where $f'(s) > 0$ is the marginal productivity of structures in the production of housing services. However, the force of competition drives up land rents and thereby eliminates positive profits so that $r = p f(s) - q s$. It follows that

$$dr/dX = f(s) dp/dX + p f'(s) ds/dX - q ds/dX = f(s) dp/dX,$$

since $p f'(s) = q$. Thus, dr/dX has the same sign as dp/dX . The rental value of land falls with distance from the CBD.

Since the unit cost of building structures is not dependent upon location, $dq/dX = 0$. It follows that

$$f'(s) dp/dX + p f''(s) ds/dX = 0.$$

The force of diminishing returns implies that $f''(s) < 0$, and thus that ds/dX takes the same sign as dp/dX . The structural density of housing (building size relative to land size) falls with distance from the CBD.

There are, of course, many other variables that affect the price of housing services in addition to distance from the CBD, including housing stock composition, quality of housing, size of dwellings, proximity to transportation, proximity to other amenities, neighbourhood characteristics and land scarcity.

(d) Opportunity Cost of Travel Time (Value of Travel Time Savings)

The previous model involves the twin assumptions that all travel time is either lost wage income (work time) or lost leisure time, and that freedom of choice over working hours leads commuters to value lost leisure time at the hourly wage rate. The first assumption is probably true for *on the job* travel time. However, some forms of commute are not incompatible with work and/or leisure (for example, reading while travelling on a train or bus), and some travel time is expended for pleasure activities. In addition, workers may have little discretion over the number of hours worked.

Accordingly, as a rule of thumb, one should use the wage rate (w) for the hourly cost of business (on the job) travel, while using approximately $0.5w$ for the hourly cost of commuter time, and smaller numbers for leisure time (pleasure) trips. It should be remembered that there is also an out-of-pocket cost of travel, so that the full cost of travel is $(t + wm)$ per kilometre for business travel, and approximately $(t + 0.5wm)$ per kilometre for commuter travel.

(e) Congestion Externalities and the Private Automobile

Consider a situation where 5,000 commuters have a choice between two alternative routes between home and work. Travel by route 2 takes 30 minutes, or \$5 per commuter, regardless of the number of commuters. Travel by route 1 takes 18 minutes, or \$3 per commuter, for up to 1,200 commuters using this route; 30 minutes, or \$5 per commuter, when there are 3,600 commuters using this route; and 46.8 minutes, or \$7.80 per commuter, when there are 6,000 commuters using this route. Average costs rise as more commuters opt to take route 1 because of traffic congestion along this route.

In open access equilibrium, commuters will spread themselves across the routes so that *average* commuting costs (ACC) are equal for the two routes. This solution has 3,600 commuters using route 1, and 1,400 commuters using route 2, generating total commuting costs of \$25,000.

However, the open access equilibrium fails to minimise total commuting costs. Indeed, total commuting costs are minimised when *marginal* commuting costs (MCC) are equalised across the two routes. The MCC associated with route 2 is \$5 per commuter regardless of the number of commuters using this route. For any number of commuters (N) between 1,200 and 3,600 using route 1, the MCC associated with this route is:

$$MCC(1) = \$3 + 0.0025 (N - 1,200).$$

Notice that $MCC(1) = \$9$ when $N = 3,600$, and that the number of commuters on route 1 that is consistent with $MCC(1) = MCC(2) = \$5$ is 2,000. Thus, the cost minimising allocation of commuters is 2,000 to route 1 and 3,000 to route 2. With this allocation, total costs are equal to \$21,800, made up of \$15,000 for route 2 commuters, and \$6,800 for route 1 commuters (at an average cost of \$3.40 per commuter).

In open access equilibrium, commuters allocate themselves to routes so as to equalise average commuting costs on the two routes. Yet optimality (cost-minimisation) requires that commuters be allocated so as to equalise marginal costs on the two routes. Because route 1 is subject to congestion, marginal costs exceed average costs. Therefore, in equilibrium, too many commuters use route 1. This is the congestion externality.

An optimal road toll on route 1 would create a disincentive for commuters to use this route, and could therefore remove the congestion externality. Since the ACC on route 1 is equal to \$3.40 when $N = 2,000$, the optimal road toll on route 1 is \$1.60. If such a road toll were implemented, toll revenues would equal \$3,200, which is equal to the cost savings between the cost-minimising allocation and the open access allocation.

In summary, “The congestion externality arises because, when an additional car enters the road and causes cars to be slightly more tightly packed together, that additional car lowers the speed and raises the travel times of all cars travelling on that road. The additional traveller does not account for the costs that he imposes on other users when making his decision to travel on the road as opposed to other roads or modes of transportation, or when making his decision to travel at this time as opposed to other times. Because he does not account for the full social costs of his trip on this road at this time, his choice of this road and this time as opposed to other roads, times, or goods may be socially inefficient.”

(Henderson, *Economic Theory and the Cities*, p. 138)

“If congestion on roads is unpriced, this means that automobile travel is priced at less than marginal cost. If other goods are priced at marginal cost, this implies that automobile travel is implicitly subsidised relative to other goods and thus is over-consumed relative to other goods. There are two optimal ways to solve this problem of distortions in relative prices and consumer choices. One is to impose congestion tolls.... The other is to subsidise all other goods (particularly close substitutes such as urban transit systems) by the same relative amount as road travel, so that consumers are no longer encouraged to use roads relative to other goods.” (Henderson, p. 145)

(f) The Cost of Road Congestion in the United States

Urban Area	Daily Vehicle miles travelled* (millions)	Delay cost (US\$ millions per year)	Fuel cost	Total cost
Los Angeles	196	7790	830	8620
New York		7140	760	7900
San Francisco	58	2760	300	3060
Chicago	79	2720	280	3000
Washington	49	2690	270	2960
Detroit	57	2210	210	2420
Houston		1830	170	2000
Boston		1500	150	1650
Atlanta	46	1400	130	1530
Seattle-Everett	31	1280	140	1420
Miami	28			
San Diego	38			
San Bernadino	27			

* Daily vehicle miles travelled on freeways and major arterial roads. (Source: Tim Lomax, Texas Transportation Institute)

(g) Modal Choices, Interchanges, and Public Transport Systems

The investigation and analysis of consumer and supplier decision making in urban transportation is complicated by the large number of dimensions in which choices may be made, and the complex physical and behavioural linkages between many of these variables. Travelers may choose the *locations* of their activities, and the *routes*, *times*, and *modes* of travelling between these locations. Their choices are influenced not only by the price and income variables that would enter a neoclassical demand function; they are affected importantly by the service quality offered by available routes or modes, particularly travel times, scheduling and physical convenience, reliability, and so on. Moreover, that service

quality is itself the result of demand and supply interactions.

The traditional urban transportation planning model that is used to evaluate alternative investments in transportation infrastructure consists of a series of inter-related models:

- (a) land use forecasting models that project the location of economic activities, employment and population by urban zone;
- (b) trip generation models that provide trip origins and destinations across urban zones;
- (c) zonal interchange models that convert projected trip origins and destinations into forecasts of inter-zonal travel;
- (d) modal split models that assign inter-zonal trips to alternative transport modes (and associated routes) and identify inter-modal exchange/transition points; and
- (e) network assignment algorithms that provide assignments to public and private transportation networks.

The existing transportation system, including its service and cost characteristics, has a major influence on traveller choices that affect (c), (d) and (e) above.

In assessing transportation infrastructure projects, one generally needs to include estimates of (a) the value of travel time savings, (b) the extent to which accidents/injuries/fatalities may be reduced, (c) the impact on vehicle operating costs, (d) the impact on the environment, and (e) the construction and maintenance costs associated with the project. Reducing congestion externalities may be an important aspect of some of these components.

In measuring travel time savings, one should (a) estimate the travel time impact for the average traveller in hours per day (or portions thereof), (b) estimate the average number of travellers by type (business, commuters, others) that are affected per day, (c) price the hours appropriately for each type of traveller, and (d) convert to annual totals if other data that enter the cost-benefit framework are estimated in annual terms. For part (a), one needs to remember that distance is equal to average speed multiplied by elapsed time, and that travel time savings is measured by distance/speed under the status quo minus distance/speed when the project is in place.

(h) The Statistical Value of Life

There are several methods that can be used to measure the statistical value of life. These include:

- (a) the *forgone earnings* method (not recommended),
- (b) the contingent valuation method,
- (c) labour market studies, and
- (d) consumer goods market studies.

The forgone earnings method or *loss of lifetime income* approach used in the law of torts may be useful in settling legal claims, but since it relates to actual lives taken it is not used to estimate the value that society places on a life saved, or the cost it associates with the loss of life, when the individual involved is *randomly selected*. One is concerned with the value of a statistical life, where the individual is not specified and could be anyone within the particular society.

Contingent valuation methods may be used to survey individuals about their willingness to invest in specific ways to increase health and safety, and to prevent fatal accidents. This stated preference method may be less reliable than revealed preference methods, such as labour market studies or consumer goods market studies.

Labour market studies are based on the wage premiums paid to those who undertake risky jobs, or the differences in wage rates that are observed among jobs with different degrees of mortality risk, for example, high rise construction jobs versus other manual labour jobs with little or no risk of accidental death. Essentially, if w measures the annual wage premium paid for the risky job, and p measures the differential risk of a fatal accident (for example, the differential incidence per 100 workers per year), then w/p measures the statistical value of life. Since w is a substantial dollar value and p is a very small difference in probabilities, w/p will normally turn out to be measured in the millions. However, since wage differentials need to be controlled for influences other than mortality risk, a hedonic regression relating w to p and several other labour market characteristics, used as control variables, may need to be estimated.

Consumer goods market studies look at individual behaviour with respect to decisions concerning safety issues, including the purchase of insurance and safety-related goods. These studies measure the willingness-to-pay for safety-related devices (for example, air bags in cars) that potentially lower the probability of accidental death, and are again based upon dividing dollar amounts spent per year (w) by the reduction in the incidence or probability (p) of accidental death that results from using these devices (again taken on a yearly basis). Thus, w/p is an estimate of the statistical value of life. Hedonic techniques might also need to be used with this method.

(i) Transportation Systems and Urban Form

Newman and Kenworthy study the relationships between transportation systems and urban form in 32 cities in the United States, Australia, Canada, Europe and Asia. They find distinct contrasts among regions in urban structure and sub-city urban densities. These differences are explained by the distinction between automobile-based transportation systems and public transit systems (bus, rail), and associated facility development. They also find a distinct negative correlation between per capita petroleum consumption and urban density. (See Bannister, Button and Nijkamp, *Environment, Land Use and Urban Policy*, Chapter 28.)

Anderson, Kanaroglu and Miller discuss the relationship between urban spatial structure and energy utilisation within cities, using the following schematic:

Urban form and spatial structure (size, density, spaces) →	Transport systems and characteristics of travel (modes of transport, distances travelled, frequency of trips, vehicle occupancy) →	Energy use and environmental impact
--	--	-------------------------------------

They study the impact of journey to work (commuting) and all other (non-work) trips, and

identify two major trends: concentration of an increasing share of population and economic activity into urban areas, and the expansion of urban sprawl, thereby transforming urban form. They ask the question: does adding road infrastructure reduce congestion and vehicle emissions, or lead to more dispersed and inefficient patterns of land development (urban sprawl)?

They also study three concepts of urban form: concentric city, radial city, and multinucleated city. They note the correlation between energy efficiency and environmental quality of the local air-shed, and stress the importance of open spaces with dense vegetation within urban areas (and thus the need to intersperse high and low density districts). (See Bannister, Button and Nijkamp, *Environment, Land Use and Urban Policy*, Chapter 16.)

Breheny (in *Sustainable Development and Urban Form*, London, Pion, 1992) discusses the theme of transportation, energy use and urban form, and suggests that *decentralised concentration*, where a number of well-linked centres with high overall densities function within a larger urban area, could be energy efficient, while green spaces could be maintained between centres. Good public transportation systems between decentralised nodes (e.g. light rail transit, or LRT, systems) are essential if transport costs are to be minimised. Excessive emphasis on compactness and urban containment can lead to congestion diseconomies.

Appropriate planning policies would therefore include discouragement of dispersed, low density residential areas, or any significant development heavily dependent upon car use; some degree of concentration, though not necessarily centralisation, of activities; integration of development with public transport facilities; and the maintenance of moderately high densities along transport routes. This approach would promote the development of a number of suburban centres within larger urban areas, focussing on improved transport systems and attempting to avoid the congestion and high fuel consumption associated with a single urban core. However, efficiency does not imply rearranging our lives to minimise transport costs. Rather it implies a search for a suitable balance between transport costs and compatibly configured land uses. *Decentralised concentration* may provide such a balance.

Breheny draws several conclusions:

- (a) urban containment policies should continue to be adopted, and the decentralisation process slowed down;
- (b) extreme compact city proposals are unrealistic and undesirable;
- (c) various forms of *decentralised concentration*, based around single cities or groups of towns, may be appropriate;
- (d) inner cities must be rejuvenated, thus reducing further losses of people and jobs;
- (e) urban (or regional) greening must be promoted;
- (f) public transport must be improved both between and within all towns;
- (g) people-intensive activities must be developed around public transport nodes, along the Dutch *right business in the right place* principle;
- (h) mixed uses must be encouraged in cities, and zoning discouraged;
- (i) combined heat and power (combined cycle co-generation) systems must be promoted in new and existing developments; and

- (j) fuel taxes and other economic instruments should be used to discourage singleoccupant automobile use.

Acutt and Dodgson (Chapter 18 in Willis, Turner and Bateman (eds), *Urban Planning and Management*, Northampton, Elgar, 2001) undertake a comprehensive review of the relative effectiveness of a range of alternative economic and regulatory policy instruments, available to counteract the many adverse environmental consequences of the transport sector. Economic instruments considered encompass fuel taxes, emission taxes, variable car excise taxes, scrap bounties, road congestion charging, parking charges and public transport subsidies; the regulatory instruments surveyed include emission level regulations, road construction, traffic calming, vehicle use restrictions, parking controls, land use planning, vehicle noise regulations and safety regulations.

The effects of these various instruments are contrasted on a number of transport indicators (such as car ownership, kilometres travelled by car, kilometres travelled by public transport, and fuel consumption) and on levels of environmental costs (including atmospheric emissions, congestion effects and traffic accidents); and assessed against the feasibility of the policies. The use of economic instruments might be expected to increase economic efficiency, since this is their primary purpose, although in practice there are problems in defining the economically efficient price and tax points. Regulatory approaches are seen as cruder and less efficient economically; but can be perceived as a significant complement to economic instruments, and as a mechanism to fine-tune traffic management at the local level.

In summary, urban design philosophies and town planning policies have gained popularity in recent years as ways of shaping travel demand. Their objectives are to reduce the number of motorised trips; increase the share of non-motorised trips (trips by foot and bicycle); reduce travel distances (encourage shorter trips); and increase vehicle occupancy (trips by public transport, and ride sharing). The examples, in Singapore and London, of road congestion charges via electronic vehicle identification systems, are of considerable interest.