TECHNICAL PAPER 13

CRI

CENTRE FOR
THE STUDY OF
REGULATED
INDUSTRIES

WHAT ARE MARGINAL COSTS AND HOW TO ESTIMATE THEM?

Ralph Turvey

UNIVERSITY OF
BATH
SCHOOL OF MANAGEMENT

WHAT ARE MARGINAL COSTS AND HOW TO ESTIMATE THEM?

Professor Ralph Turvey

Desktop published by
Jan Marchant

ISBN

Preface

The CRI is pleased to publish ‘What are Marginal Costs and How to Estimate them?’ as CRI Technical Paper 13. A distinguished economist, Professor Ralph Turvey developed the thinking on marginal cost, notably in his 1969 paper in the Economic Journal, which, here, he modestly, but incorrectly, refers to as “expositions in long-forgotten works”.

Privatisation, incentive regulation and the introduction of competition has revitalised the debate on the theory and practice of measuring marginal costs, and Professor Turvey’s paper well demonstrates how important it is to be clear about matching the appropriate definition of marginal cost to the objective in mind. The paper brings together a rigorous analysis of the principles of marginal cost with comparative analysis of its application in different regulated sectors. It should, therefore, contribute to both academic and policy debate.

The CRI would welcome comments on this paper and further analytical work in the area. Comments, enquiries or manuscripts to be considered for publication should be addressed to:-

Peter Vass
Director – CRI
School of Management
University of Bath
Bath, BA2 7AY

The CRI publishes work on regulation, by a wide variety of authors and covering a range of regulatory topics and disciplines, in its International, Occasional and Technical Paper series. The purpose is to promote debate and better understanding about the regulatory framework and the processes of decision making and accountability.

The views of authors are their own, and do not necessarily represent those of the CRI.

Peter Vass
Director, CRI

March 2000

Contents

PRINCIPLES

Economists tend to think about costs in terms of static, timeless models with continuous cost functions. The real context is, however, one of businesses and systems which already exist and have accumulated a collection of assets of various vintages whose accounting cost reflects past prices, past circumstances and arbitrary conventions about depreciation. In the applied economics context, such as utility regulation, the textbook theory is of no help. This paper sets out a more useful approach, with examples partly following expositions in long-forgotten works[1].

Economic concepts

Economic vs accounting cost

Cost analysis and the allocation of cost can mean both:

- *attributing causality*, that is, asking how cost would change if output changed. This is an economic question which serves decision-making. Since these decisions relate to the future, the estimation of economic cost involves comparisons between future alternatives;
- *determining revenue requirements for fixing prices*, that is, deciding which customers should contribute how much towards covering accounting costs, including a target profit. This can be either a business or a political issue about perceived fairness.

The distinction between economic cost and accounting cost is particularly important in relation to existing assets. The economic cost of using them may be very different from their accounting cost, whether this depreciation reflects historical or replacement cost:

- if they could be rented, as with commercial premises, buses or planes, their economic cost is the rent that could be obtained;

[1] Turvey, R (1968) *Optimal pricing and investment in electricity supply*, Allen and Unwin.

The contributions of French electricity economists deserve recognition; they were made available in translation in Nelson, J [Ed] (1964) *Marginal cost pricing in practice*, Prentice Hall International.

Turvey, R. (1969) 'Marginal cost', *The Economic Journal*, Vol.79, June 1969, pp.282-99

Littlechild, S (1970) 'Marginal cost pricing and joint cost', *The Economic Journal*, Vol.80, June 1970.

Turvey, R (1971) *Economic analysis and public enterprises*, Allen and Unwin, chapters 6 and 7.

Ralph Turvey is visiting professor in regulation, London School of Economics

- if they could be sold, even if only for scrap, the economic cost of using them for a year is their operating cost plus the excess of what they could be sold for now over the discounted value of what it is guessed they could be sold for at the end of the year;
- if they would have to be expensively decommissioned, the economic cost might even be negative.

The economic costs and the values of existing assets are two totally different things. To say that an existing asset such as a railway line in a tunnel or a water main under a street may cost very little to use does not necessarily mean that it has a low value. They are defined as follows:

- *economic cost*, as just described, is the value of the sacrificed alternative (opportunity cost), that is, what would be done with an asset if it were not used for production;
- *value* is the effect upon the present worth of the enterprise’s future cash flow of an asset vanishing in a puff of smoke – ‘deprival value’ (this may equal replacement cost or may even equal depreciated historic cost). If, under this scenario, the best alternative is to build a new generating station which would burn a different fuel and, being new, would last longer; then the calculation of deprival value would not be simple.

Any excess of value over economic cost is an *economic rent*.

Marginal cost

Marginal cost is an estimate of how economic cost would change if output changed. *Marginal* means a first derivative, but in practice, because of indivisibilities in plant sizes, we are often interested in the per unit change in cost that will be caused by a substantial change in a future output, not of a one unit change. Furthermore, investment and capacity are not continuously variable, they are lumpy. Transco builds 900 mm and 1,000 mm pipelines but has never considered building 901 mm or 999 mm ones. Marginal costs involve forecasting, since they are the differences between what was and what would have been with different outputs. The consequence is that, even when the concept of marginal cost is completely agreed in principle, its estimation involves far more than calculations founded upon a set of rules. All forecasts are subject to error, including marginal cost estimates.

Given:

- an existing system, and
- plans for expanding (or contracting it) and replacing plant so that its output will meet the forecast growth (or shrinkage) of demand,

marginal cost is an engineering estimate of the effect upon the future time stream of outlays of a postulated change in the future time stream of output. There are as many marginal costs as there are conceivable postulated changes. Estimating any of them usually requires engineering and, often, operational research skills. It rarely requires accounting skills.

Marginal costs between upward and downward changes may differ. If they are estimated by comparing a postulated alternative scenario with the most probable or base case, increments and decrements are equally likely. It is then advisable to compute both and take their mean.

Unfortunately, the use of the term *increment* makes it easy to forget to examine decrements as well as increments. A further problem is that different people use *incremental cost* to mean different things, including:

- the cost saving from wholly abandoning the production of something, that is a 100% decrement in the output of a product or service[2];
- the future costs that would be saved if outputs were permanently maintained at their present level, resulting in saving the cost of all planned or forecast future capacity expansion to meet forecast demand growth.

To avoid these confusions, I shall continue to use the term 'marginal cost' even when it relates to the unit cost of a large change in an output.

Marginal cost to whom?

Marginal cost to the provider of a good or service – *marginal private cost* – can be distinguished from marginal cost to the economy as a whole – *marginal social cost*. Marginal social cost is marginal private cost plus any marginal costs imposed upon others and less any marginal benefits conferred upon others.

Marginal cost should be looked at in system terms. For marginal private costs, this refers to the utility or enterprise's own system and for marginal social costs, this refers to the economy as a whole.

The necessity to look at the utility's own system as a whole is demonstrated by the following example provided by a study of the marginal cost of water supply in a city where a number of tube wells pumped water into water towers, which fed a meshed reticulation of distribution mains. Demand growth in winter could be met by more pumping from existing wells, but providing for demand growth in summer would require additional tube wells at new sites without any need to reinforce the distribution network. The annuitised cost and annual running costs of a new well, divided by its summer output and multiplied up to allow for distribution leakage, yielded a marginal cost per cubic meter delivered in summer. However, this understated marginal private cost to the utility because the extra pumping would gradually lower the water table, increasing the utility's

[2] This is what OFTEL, for example, means by long-run incremental costs, though 'it is assumed that all assets are replaced in the long-run' which turns out to mean that these costs are estimated for a hypothetical system incorporating 'the latest available and proven technology' but with the same topology as the existing system. The result is that , for OFTEL, the long-run incremental costs and the stand-alone costs of access and conveyance are defined as being the same (OFTEL (1995) *Pricing of Telecommunications Services from 1997: Annexes to the Consultative Document*, December, Annex D). Similarly the assertion that the stand-alone cost of any set of services equals the total costs of the firm minus the incremental cost of its other services, holds only for a brand-new system built from scratch, a point concealed by the ingenuous assertion that 'a complication arises if the firm is not an efficient supplier' (Baumol, W and Sidak, J (1994) *Toward competition in telephony*, MIT Press).

future pumping costs throughout the year. The present worth of this increase thus needed to be added to the initial figure.

The differences between marginal private costs and marginal social costs are demonstrated by the following examples:

- Road use

The *private marginal cost* to a car owner of driving a mile is the petrol cost per litre divided by miles per litre, plus, say, one ten thousandth of the price of a ten thousand mile service, plus, maybe, some extra depreciation in the value of the car and an allowance for the value to the driver of his own time. *Marginal social cost* will differ, because:

- it values petrol only at factor cost;
- by adding to congestion, if the mile is driven on a busy road, a time cost is imposed upon other drivers and their passengers;
- if accident rates are proportional to the square of traffic flow, then the marginal accident rate is twice the average, but only the latter enters into marginal private cost.

Congestion and accidents are costs even if a money value cannot be put on them.

A second transport example relates to buses. Putting one more or one less bus on a route adds or subtracts from the bus operator’s costs. However, by increasing or decreasing the frequency of the service, it reduces or increases the average waiting time of passengers. Since time has a value (though estimating it is tricky), marginal social cost is less than marginal private cost.

- Taxation

A public power corporation had a choice of having a combined cycle gas turbine (CCGT) plant built on a turnkey contract and handed over to it immediately; or having a consortium build, own and operate the plant, charge a tariff for its output for twenty years, and then transfer ownership. To inform the decision, the present worth of the corporation’s payments under these two alternative methods of obtaining extra generation were compared. Since the plant would be the same in both cases, the information needed to make the comparison was limited:

- the CCGT would be used as an intermediate plant, producing at full capacity for 16 hours a day and at about one-third of capacity for 8 hours. Hence, a kWh-weighted average of the two energy costs was used for the comparison;
- the corporation might achieve lower plant availability and a higher heat rate than the consortium, yielding a smaller output. The marginal cost of producing the difference (from oil-fired plants) had to be added to corporation costs under the turnkey alternative to create comparability with the proposed tariff.

The question left open was whether the system was the corporation or the whole economy. Under the *build, own and operate* alternative, the consortium would pay the national profits tax and its quoted tariff, therefore, included an allowance for this, since its overseas' promoters were concerned with their projected net returns. While the tax under this alternative would raise the cost of energy to the corporation, it would not raise the cost to the whole economy as the tax would accrue to the government.

Marginal cost of what?

The postulated change in future output must be properly specified. Single-output systems hardly exist. Since the electricity industry supplies electricity at *n* different voltages and in *m* different areas in each of the 17,520 half-hours of the year, the number of marginal costs of electricity in any year is $n \times m \times 17{,}520$. For practical purposes, however, simplification into a smaller number of grouped outputs is necessary. In the case of electricity generation, one of them, will almost certainly be peak kW output and others will be average kWh outputs over groups of hours. All of these have marginal costs. In the case of gas transmission, peak day firm demands in a one-in-twenty winter in each Local Distribution Zone are relevant for capacity planning. Water supply is similar.

Joint costs arise when capacity used for producing output X is also used for producing output Y; and where they are not substitutes in production but, if capacity is fully utilised, are produced in fixed proportions (for example, oil & gas from a well, or east-bound and west-bound trains on a transport link).

In such cases, costs are not independent of demand, for a marginal capacity cost can be computed only as:

- for X

  marginal capacity cost *plus* marginal operating cost of X *minus* any extra revenue from Y *less* operating cost of the extra Y;

- for Y

  marginal capacity cost *plus* marginal operating cost of Y *minus* any extra revenue from X *less* operating cost of the extra X.

In these cases, marginal cost cannot be derived from a cost-minimising engineering model alone and, therefore, cost allocation includes the determination of revenue requirements. For example, the cost of a one-way trip by a liner can be allocated in the sense of deciding fares but not in a causal sense.

Marginal cost when?

There are two aspects of timing:

- when the output is produced, and
- when the decision is taken.

How much the latter precedes the former determines whether the increment or decrement can be accommodated solely using existing plant more or less intensively; or whether additional capacity must be installed to provide for an increment, or expansion plans must be postponed or plant scrapped to adjust to a decrement. I use the distinction between long and short-run as signifying whether capacity can or cannot be varied.

This is *not* the same as the textbook economics distinction, where *long-run* signifies that all inputs can be varied. A textbook long-run cost function relates to a brand-new, built from scratch system. Accounting and economic costs do not need to be distinguished in this case, so it can be said that, with economies of scale, pricing at long-run marginal cost would not cover total accounting cost. But in this paper, *long-run* simply means far enough in the future for plant additions or retirements to be made; and *short-run* means with given capacity. Thus, a long-run marginal cost cannot be estimated for the near future, whereas a short-run marginal cost can be estimated for any time in the future given the system that is then assumed to exist. Nothing at all can be said, a priori, about the relationship between these forward-looking marginal costs and revenue requirements. Nor does the concept of economies of scale have meaning as future plant additions will differ, both technologically and in cost, from existing plant.

Very often, however, we are interested not in a one-shot increment or decrement but in a time stream. The cost change will then be a time stream too, so we must discount both to a common date at the cost of capital, divide the one into the other to arrive at a present worth of the unit cost of an increment or decrement above or below the base case projection.

It is convenient to call an increment to, or decrement from, the base case which is added to, or subtracted from, the base case forecast in all years commencing from a specified year, a *permanent* increment or decrement.

If indivisibilities were not important, then under an optimal expansion plan, the short-run marginal cost would equal the long-run marginal cost (if there were a long-run marginal cost, that is if there were time to alter construction plans). This is because, with divisibility, optimality implies that the marginal cost of changing output by using existing capacity more or less intensively will equal the marginal cost of changing it by adding to or reducing capacity.

When discussing marginal cost as a basis for pricing, it may not be appropriate to concentrate on short-run marginal costs in the near future. The effect upon costs of a one-off, one-year increment or decrement in load is only relevant when consumers will take a one-year decision. In many cases, however, the major decisions taken by consumers will be about *whether* to consume; while the plant and equipment that they decide to install is planned to be used for a period of years. Once they become consumers, however, their consumption in each year of this period may be a function of the charges they pay. If the responsiveness of their investment decision to the level and shape of the charges is greater than that of their annual consumption decisions, optimal resource allocation requires that the marginal costs underlying the charges relate to an increment or decrement of equal duration. A long-term contract is then appropriate. In its absence, a tariff expressed as an annual fixed charge plus monthly, quarterly or annual variable charges will often be regarded as a prediction of the charges to be paid over the whole life of their investment, thus conveying a long-term message.

A simple constructed example

This model relates to a public utility which supplies its output at two levels each year – in a number of peak days and a number of off-peak days. To provide a numerical example, it is assumed that in the current year there are 50 days of peak demand of 1,000 units, which is forecast to grow at 3% a year; and 315 off-peak days with a demand of 500 units, forecast to grow at an annual rate of 3.5%. It is also assumed that total capacity should always remain at a minimum of 5% greater than peak demand in order to provide security. This condition is just met in the current year, the existing plant having a capacity of 1,050. Its annual fixed operating and maintenance cost is set at £60 in the example and its variable cost at £0.05 per unit of output.

New capacity can be commissioned in any of the next twenty years (the period covered by the demand forecast). The objective should be to minimise the present worth of the costs of meeting it, subject to the security constraint. It is assumed that the cost of capital is 7% and the calculation is made at constant prices.

There are five plants which could be built and commissioned to provide the necessary expansion in capacity, as shown in **Table 1.** Note that A, B and C have lower variable operating cost than the existing plant. The five are as follows:

Table 1: Specifications of generating plants

Plant	Daily capacity	Capital cost	Expected life	Annual fixed Operating cost	Variable Operating cost
A	200	£40,000	50	£110	£0.02
B	400	£50,000	30	£65	£0.03
C	420	£72,000	50	£70	£0.04
D	450	£70,000	50	£75	£0.06
E	350	£40,000	50	£65	£0.065

Finding the optimal expansion plan turns out to be a complex programming problem. The solution is to add to the existing plant by commissioning plant B in year 1, plant A in year 10 and plant E in year 16. The costs for each year, starting with the commissioning year, until the end of the twenty-year period include:

- *variable operating costs of all four plants*. These are dispatched in merit order to meet peak and off-peak demand each year, daily output of each plant being multiplied first by the number of peak, respectively off-peak, days and then by that plant's variable operating cost;
- *annual fixed operating cost of all four plants;*
- *for the three new plants, annuitised capital costs*. Treating the annuitised capital cost of a plant as an annual cost approximates including their capital costs in their entirety in the year of commissioning and crediting their residual values at the end of the twenty years. It implicitly computes residual values as the present worths of their annuitised capital costs over the remainder of their lives.

The minimised present worth of costs incurred over all the years 1 to 20 is £160,407. The discounted amount of total outputs over the same period is 2,950,185 units.

The marginal cost for any defined increment or decrement is computed by adding or subtracting it from the demand forecast to obtain a revised forecast and then reoptimising. The marginal cost equals the difference between the two present worths of cost divided by the difference between the two discounted amounts of output units.

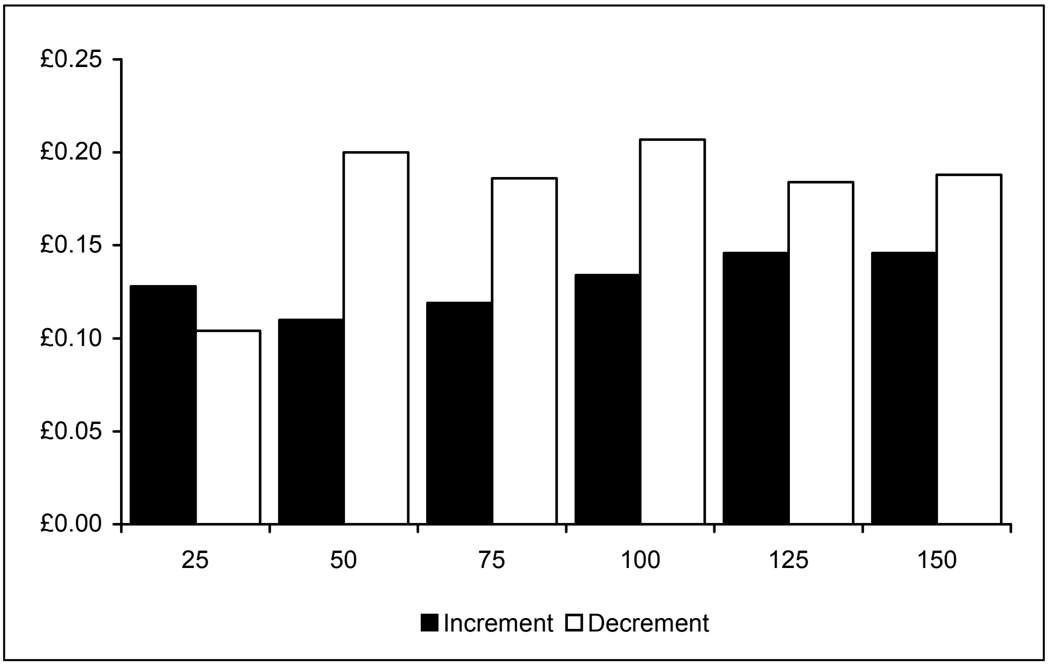
With a permanent increment of 100 starting in year 1, the order in which plants are to be commissioned remains B, A, E; but A is brought forward one year to year 9 and E two years to year 14. In addition, plant C is commissioned in year 20. The present worth of costs and discounted amount of output become £167,044 and 2,999,689 respectively. Marginal cost is thus calculated to be £0.134. The output of the existing plant falls when plants B, A and C are commissioned since thay have a lower variable operating cost. Although plant E has to be commissioned in year 14 in order to maintain the required capacity margin of 5%, it does not start to produce until year 15.

If the permanent increment of 100 is assumed to start only in year 6, the order in which plants are to be commissioned does change, becoming A, B, E, C.

With a permanent decrement of 100 starting in year 1, the order in which plants are commissioned again remains B, A, E, but all three are postponed. Marginal cost in this case is £0.207.

Marginal costs vary with the size of the increment or decrement as **Figure 1** shows, again for permanent increments or decrements starting in year 1.

Figure 1: The relationship of marginal cost to the size of the increment or decrement



The marginal costs of permanent increments and decrements of 100 rise, though not monotonically as their starting year is assumed to be later and later. The marginal costs of single-year increments and decrements equal only the unit variable cost of the existing plant except in years when excess capacity in the base case is less than 100 or would be without the decrement. In these years the marginal cost of such one-year increments or decrements is very high.

Even in this very simple model, the interrelationships are fairly complicated. They are all the more so in more realistic models.

A different approach

Thus far the discussion has run in terms of postulating increments or decrements to the demand forecasts and estimating the change in future costs necessary to adapt to them.

An alternative procedure in practice is to start with a change in capacity costs and enquire what increments or decrements in outputs the change in capacity would allow or require. It is often done by estimating the costs of bringing forward or postponing the next proposed addition to (or scrapping of) capacity and dividing by the increment or decrement in future outputs that would then be possible while maintaining an unchanged quality of service. It yields a marginal cost for an increment or decrement lasting for only one or a few years. This may be simpler. It has the great advantage of avoiding the complexities of multi-period optimisation of the whole time stream of future planned capacity extension and system operation. It may be justified:

- because there is no alternative in the absence of such explicit optimised long-term planning;
- as yielding an answer that is consistent with it when such planning is done.

However, it cannot be assumed that this will always be the case, since the optimal adjustment of plans to a postulated increment or decrement lasting only one or a few years may be more complex than simply advancing or retarding the construction of one new lump of capacity. It can be demonstrated, using the simple model expounded above, that there are cases where adjustment to a postulated increment or decrement changes the optimal order in which such new lumps are added to the system.

Probability

When there is quantifiable uncertainty, a probabilistic estimate of marginal costs is appropriate. In general, the cost corresponding to the expectation of stochastic variables is not the same as the expected cost computed as a probability-weighted average of the costs corresponding to each possible value of the stochastic variables.

Examples of the need for probabilistic estimates arise, in particular, in respect of the provision of services where the rate at which customers arrive and the service time required to service each customer are variable and uncertain. Ports furnish a striking example, since the process of berthing, unloading and loading a ship can be delayed when berths are occupied by previous arrivals. This imposes a time cost upon the owner of a ship that is delayed. Variable delays of this type also impose a cost upon freight consignees if they have to hold larger stocks to cope with the uncertainty of delivery dates. For the analysis of costs in such circumstances, the application of queuing theory is absolutely essential[3].

Estimating the operating cost of electricity generation, where plant availability is uncertain, provides another example. In **Table 2**, the joint probabilities for each possible combination of plant availabilities, which sum to unity, are multiplied by the corresponding marginal costs and summed to yield the probability weighted average of marginal cost at each of a number of load levels.

[3] Jansson, J and Schneerson, D. *Port Economics.*

Table 2: The probability weighted average marginal cost at different load levels

Plant	MW Capacity	Marginal cost	Outage rate	Availability	Value of lost load
A	60	3	0.2	0.8	
B	50	3.5	0.1	0.9	*10*
C	20	4	0.1	0.9	

Plants available	A,B,C	A,B	A,C	B,C	A	B	C	0	Probability
Their capacity	130	110	80	70	60	50	20	0	weighted
Joint probability	0.648	0.072	0.072	0.162	0.008	0.018	0.018	0.002	mean
MW Load									
5	**3**	**3**	**3**	**3.5**	**3**	**3.5**	**4**	*10*	**3.1**
15	**3**	**3**	**3**	**3.5**	**3**	**3.5**	**4**	*10*	**3.1**
25	**3**	**3**	**3**	**3.5**	**3**	**3.5**	*10*	*10*	**3.2**
35	**3**	**3**	**3**	**3.5**	**3**	**3.5**	*10*	*10*	**3.2**
45	**3**	**3**	**3**	**3.5**	**3**	**3.5**	*10*	*10*	**3.2**
55	**3**	**3**	**3**	**4**	**3**	*10*	*10*	*10*	**3.4**
65	**3.5**	**3.5**	**4**	**4**	*10*	*10*	*10*	*10*	**3.9**
75	**3.5**	**3.5**	**4**	*10*	*10*	*10*	*10*	*10*	**4.9**
85	**3.5**	**3.5**	*10*	*10*	*10*	*10*	*10*	*10*	**5.3**
95	**3.5**	**3.5**	*10*	*10*	*10*	*10*	*10*	*10*	**5.3**
105	**3.5**	**3.5**	*10*	*10*	*10*	*10*	*10*	*10*	**5.3**
115	**4**	*10*	*10*	*10*	*10*	*10*	*10*	*10*	**6.1**
125	**4**	*10*	*10*	*10*	*10*	*10*	*10*	*10*	**6.1**
135	*10*	*10*	*10*	*10*	*10*	*10*	*10*	*10*	**10.0**

When the availability of capacity is probabilistic and/or when future output is probabilistic (because of forecasting errors and weather variability), there is a probability that capacity will be insufficient, leading to failure to supply. The example above is limited to uncertainty on the supply side

One method of dealing with uncertainty in planning future system expansion is as follows. The present worth of future costs which is minimised in determining the optimal expansion plan includes the cost of supply failures. The probability of such failures is then a result of the optimisation, not an input to its computation. The higher the penalty cost attributed to supply failure, the more future capacity will be planned and the smaller will then be the probability of failures. This penalty cost can either reflect a judgement about the (negative) value of power cuts, hosepipe bans and so on, or the amount of any compensation payable by the enterprise in the event of failures.

An alternative method of dealing with uncertainty is to calculate marginal costs without including any explicit valuation of the penalty cost of failure in a probabilistic estimate of marginal costs. Instead, the cost-minimising optimisations of alternative expansion plans can be subject to a security constraint. There are two ways of doing this:

- the size of the postulated increment or decrement of forecast demand can be adjusted so that the probability of failure of supply with it is the same as in the base case of the existing demand forecast and expansion plan. This enables marginal costs to be estimated with the risk of failure of supply impounded in ceteris paribus;
- a rule of thumb provides the constraint. Examples include:
    1) that capacity must exceed requirements in an average winter by the amount of extra demand expected to occur in one winter out of twenty.
    2) the system must be able to withstand the simultaneous outage of any two power lines.
    3) capacity must exceed forecast peak demand by 15%.

CASE STUDIES

Electricity

This section demonstrates, through a series of case studies, the range of methods used to calculate marginal costs in the different regulated industry sectors. Examples are provided for electricity generation in thermal, hydro and mixed hydro-thermal systems; electricity transmission; electricity distribution; water supply; gas transmission; and the short- and long-run marginal costs in the railway sector.

Electricity generation

A thermal system

Forward-looking long term plans are developed in most electricity generating systems, since the optimal choice of what new capacity to build next is not independent of subsequent choices to be made later. Analysis and planning usually extends twenty or thirty years into the future, requiring load forecasts for the whole period covered. The work uses suites of computer programs developed by electrical engineers, such as WASP, from the International Atomic Energy Agency and EGEAS, developed by Stone and Webster for the Electric Power Research Institute.

In order to estimate marginal costs by difference, the Israeli Electric Corporation has used EGEAS to produce optimal system expansion plans for alternative load forecasts stretching twenty years into the future[4]. The optimal plans are computed by using dynamic programming to minimise total discounted costs, with probabilistic estimation of production costs and of the amount of loss of load.

Forecasts of future hourly load curves are used to define nine groups of hours – peak, shoulder and off-peak, in summer, winter and spring/autumn. Starting with a base case, a load increment is added to, or a load decrement is subtracted from, one of these groups. The difference in the discounted amount of energy provided is then divided into the difference between the discounted cost of the base plan and the discounted cost of the optimal plan for the altered load forecast to yield a marginal cost for that group of hours.

Upward marginal costs turned out to exceed downward marginal costs. The difference consisted almost entirely of operating costs, since the plan adjustments all involved either simply bringing forward or postponing the installation of new gas turbines. Their very much higher fuel costs would thus be incurred earlier or later respectively.

[4] Porat, Y, Rotlevi, I and Turvey, R (1997) ‘Long-run marginal electricity generation costs in Israel’ *Energy Policy* Vol.25 no. 4 pp. 401-411.

There are two problems that arise with such computations:

- the first relates to the value of lost load. To deal with this, the first of the two methods described above to calculate marginal costs without including any explicit valuation of the penalty cost of failure was used. The sizes of the postulated increments and decrements were adjusted so that the amount of unsupplied energy was the same as in the base case and therefore did not have to be valued;
- the second is the need to avoid a bias of choice against capital expenditure made near the end of the twenty year period (assuming that that is when the analysis terminates). This was dealt with by supposing that the load remains unchanged for an 'end period' of another twenty years, so that during this 'end period' no new investment is required but only operating costs are incurred. This allows trade-offs between operating costs and capital costs to be taken account of, since residual values discounted back for as long as forty years can safely be disregarded.

For all these calculations it was necessary to specify the capital, operating and maintenance costs (at constant prices) of candidate new generating plants, as well as their reliability and operating characteristics.

The optimal plan differences necessary to accommodate the postulated increments and decrements involved only a change in the size and timing of new Open Cycle Gas Turbines. Since documentation of this case study, however, it appears that a new Combined Cycle Gas Turbine plant might be built, an alternative not contemplated. This would alter the base plan and could lead to a different set of marginal cost results.

A separate calculation related to an increment and a decrement, of 147 MW, extending over all hours of the year. This case was examined by means of a more sophisticated calculation. Operating costs were computed for selected years by using a probabilistic chronological simulation which took account of constraints such as minimum up and down times, ramp rates and pumped storage constraints. This yielded marginal operating costs for the redispatches necessary to accommodate the increment and decrement. Marginal capacity costs, obtained from the simpler optimisation, were spread over the hours of the year proportionately to the probability of lost load. Slightly different results were obtained, primarily because of the attribution of more of the marginal capacity costs to winter peak and shoulder hours.

A simple analysis

Such complex tools may be dispensed with because their demanding data requirements cannot be met. Also, where it is clear what new plant is to be commissioned next, but the timing of its construction can still be altered, they are not necessary. Long-run marginal capacity cost can then be computed as what, for an increment, the French call a 'coût d'anticipation'.

Consider the case of an increment above the forecast peak period load. Its short-run marginal cost can be ascertained by simulating despatch and comparing operating costs, *plus* expected loss of load multiplied by the value of lost load, with and without the increment.

Long-run marginal cost, now including a capacity cost, can be ascertained by supposing the commissioning of the next plant to be brought forward in time so that a higher peak load can be met without any diminution of the probability of lost load. If plant commissioning is brought forward from 2001 to 2000 to meet a postulated higher level of peak load in 2000, there will then be an addition to system cost of one year's fixed operation and maintenance cost plus one year's annuitised capital cost of the new plant (this is only an approximation, for it neglects two points: firstly, bringing construction forward may sacrifice some expected technical progress and, secondly, it will also bring forward future expenditure on replacement of the new plant, replacement of the replacement and so on) [5].

If the new plant is a gas turbine, with high fuel and variable maintenance and operating costs, marginal energy cost in peak hours will then equal its unit fuel and variable maintenance costs.

But if the new plant is, for example, a base or intermediate load plant, its marginal energy costs will fall well below those of much of the existing plant. In consequence it will be despatched to run for much of the year, displacing some generation by some of the existing plant. To ascertain the resulting savings in fuel, variable maintenance, and operation and maintenance (O & M) costs it is necessary to simulate system operation in the year 2000 with and without the new plant. These savings are then deducted from the annuitised capital, fixed O & M costs, to obtain the net increase in system capacity costs. Dividing this by the size of the postulated increment in peak load then yields long-run marginal capacity cost per peak kW in the year 2000. A complication which can arise is that the new plant's expected availability may be lower in its first years of operation than subsequently – teething problems being a normal phenomenon in the early operation of new plants. The calculation of the fuel savings which would result from bringing its commissioning forward then has to cover more than one year.

The following simple example of the computation of marginal cost upwards, set out in **Table 3**, illustrates these points:

[5] For discussion of these complications, see Turvey, R (1968) *Optimal pricing and investment in electricity supply*, Allen and Unwin, p39-42.

Table 3: Marginal cost of peak HV supply: A simple imaginary example

Capital cost of new 500MW plant				$350,000,000	
Capital cost of associated transmission				$4,500,000	
Total capital cost				$354,500,000	
Annuitised over 25 years @ 7%				$28,429,746	
Annual fixed O & M				$750,000	
Total annual cost					$29,179,746
	Year	Base case	Brought forward 1 year	Difference	
System	2001	$123,000,000	$119,000,000	$4,000,000	
Running	2002	$137,500,000	$135,000,000	$2,500,000	
Costs	2003	$145,600,000	$145,000,000	$600,000	
Present worth of fuel savings in 2001				$6,411,693	
Total annual cost net of fuel savings					$22,768,053
Divide by 500,000					
equals marginal generation capacity cost (net of fuel savings) per kW				$45.54	
X average peak transmission loss multiplier			*1.02*		
*equals m*arginal generation capacity cost of HV supply					$46.45
Capital cost per peak kW of required transmission reinforcement				$55.00	
Annuitised over 30 years @ 7% *equals*				$4.14	
Multiply by transmission diversity factor			*1.03*		
equals marginal transmission capacity cost at HV					$4.27
Marginal transmission & generation capacity cost at HV per kW					**$50.71**
Divided by number of hours of potential peak			*200.00*		
equals marginal transmission & generation capacity cost at HV per kWh					$0.25
*plus m*arginal energy cost in potential peak hours					$0.12
Marginal capacity and energy cost in potential peak hours					**$0.37**

Simulating operation

As the examples show, simulations of the operation of an electric power system, with a given system configuration and with given assumptions about fuel prices, are carried out for two reasons. Firstly, to estimate a year's total operating cost as part of the process of finding an optimal plan for the future growth of the system over a long period. Secondly, to estimate marginal operating costs, that is short-run marginal generating costs at different load levels and thus at different times of day and year. In either case, the following steps are necessary:

1) Forecast the level and time pattern of required generation for the 8,760 hours of the year. This may be done in a complex way or simply by converting the forecast annual kWh into the 8,760 hourly forecasts by assuming that the load shape will resemble that of the last full year.
2) Simulate the way the system would be despatched, taking into account:

- the start-up and shut-down cost of each generating set, the rate at which it can be brought up and shut down, its fuel cost and heat rate (thermal efficiency) varying with load;
- scheduling of planned outages for maintenance;
- the need to have some generating sets running at less than their capacity to provide a spinning reserve which can meet unexpected surges in demand or outage of a generating set or a transmission link
- any transmission constraints;
- any must-run plants', for example combined heat and power (CHP) plants, reactive power demands.

3) Many simulations are required for each year – a simulation for each combination of possible forced outages of the generating sets. The probabilities of these different system states may be convoluted with the probabilities of different loads to allow for uncertainty in the forecast and the variability of weather round the seasonal means assumed in making the forecast. Each set of results is multiplied by its probability and summed to obtain expected values.

4) Some combinations of outages and/or weather will create demands in excess of capacity, leading to voltage reductions or power cuts (lost load). The amount of lost load multiplied by its probability summed over all the probabilistic outage and load combinations is the expected lost load (unserved energy). It can be multiplied by the estimated value of lost load, to yield curtailment or outage cost. If this is done, it can be included in total costs for the purpose of optimisation and in calculating the marginal social cost of electricity demand as:

   - the probability that an extra kWh can be supplied × marginal operating cost + the probability that it cannot × the value of lost load.

   Estimation of this value is by no means easy[6].

5) The existence of hydro plant adds additional uncertainties because the amount of water available can fluctuate from year to year, as river flow records will reveal. If the hydro capacity has a large storage reservoir, selecting the optimal timing of the use of the water for generation, to maximise the expected fuel cost of the thermal generation which it displaces, is a complex matter. One reason is that the height of the water in the reservoir, as well as the quantity of water turbined at any point of time, affects the rate at which energy is generated. Dynamic programming is a

[6] Kariuki, K and Allen, R (1996) 'Factors affecting customer outage costs due to electric service interruptions' in *IEE Proceedings – Generation, Transmission, Distribution*, Vol.143, No.6; and earlier papers by the same authors listed in their references. See also Munasinghe, M (1979) *The economics of power system reliability and planning*, published for the World Bank, Johns Hopkins University Press.

tool used to solve these problems. There is a vast technical literature on the subject, both for pure hydro and for mixed hydro-thermal systems. It should be noted that the function served by hydro plant can change through time if the share of thermal plant in the system grows. It may be worthwhile installing more generation capacity at the hydro sites for use in peak periods and even to use the sites for pumped storage.

A predominantly hydro system

Now consider the particular case of a hydro system which had only one thermal plant, its function being to provide energy when water was short.

The system was dimensioned to meet the expected firm energy consumption of its consumers in a dry year, as determined by 35 years' stream-flow records. This required the successive construction of new dams and generating plant, with associated transmission, at a pace determined by the expected growth in consumption. The result of this policy was that:

- generating capacity always exceeded maximum demand. In consequence, the marginal capacity cost of peak consumption was zero;
- in most years, water availability was more than sufficient to meet the expected firm energy consumption of its consumers. The surplus could, therefore, be sold at low prices to interruptible consumers or neighbouring utilities. Deducting the expected revenue from these sales from the capital and operating cost of new hydro plant yielded the net addition to (or saving of) cost from providing the necessary generating capability for a higher (or lower) firm load than assumed in the base case.

Given capacity, the short-run marginal cost of firm energy for any year could be computed. It was a function of water availability if known, or, if not, of the frequency distribution of stream flows, and of a probability distribution of the amount of firm consumption. It equalled the sum of:

- the probability that an altered sale of firm energy could be accommodated by selling more or less interruptible energy multiplied by its price;
- the probability that it could be accommodated by decreasing or increasing the output from the thermal plant multiplied by its marginal operating cost;
- the probability that it would result in less or more lost load multiplied by the value of lost load.

Twenty-five simulations of the operation of reservoirs and generation over 35 years (one base case, twelve one-month increments and twelve one-month decrements in firm load) allowed calculation of monthly short-run marginal costs of firm energy. Averaging the difference in costs from the base case for each month over the 35 years, and dividing by the size of the postulated increment and decrement, revealed monthly short-run marginal costs in the spring to be 1.2 times the annual average. The reason for this was that reservoir levels were low in spring, with a resulting loss of head, while consumers of surplus energy were then willing to pay a higher price. In the summer, on the other hand, reservoirs were full, consumption was at a lower level and the demand for surplus energy was less. Had reservoir capacity been greater, making a larger spring drawdown possible, with more refilling in summer, the spring-summer marginal cost difference would have been smaller.

Within each month, there was no difference between night and day or weekday and weekend in the marginal generation cost of firm energy. Marginal transmission and distribution losses, on the other hand, being sensitive to load levels, caused some night/day and weekday/weekend differentiation in the marginal costs of energy delivered to consumers.

Mixed hydro-thermal systems

Mixed hydro-thermal systems vary enormously in their marginal cost structures, depending on amongst other circumstances:

- the yearly pattern of consumption in relation to that of water inflow to the reservoirs;
- the storage capacity of reservoirs;
- the capacity of run of river plant, and
- the balance between thermal and hydro capacity.

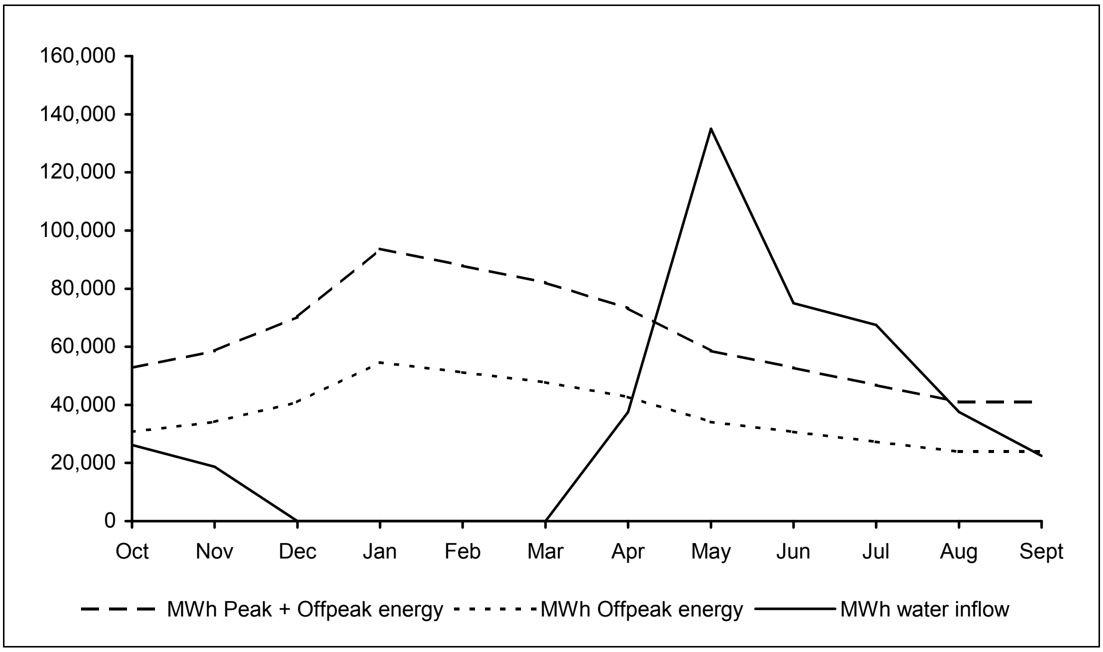
- Optimising operation

One major feature of some mixed hydro-thermal systems is, as in the case just described, that the marginal generation cost of firm energy does not vary between peak and Off-peak hours. This can be explained by means of a simple numeric example.

It is deterministic, relates to a single year and measures water in MWh thus avoiding complications of the sort described above. It assumes that the system is energy-constrained and not capacity-constrained, there being an excess of both thermal and hydro generating capacity.

When the water year starts, in October, the reservoir is assumed to be full. It is required that the reservoir should be full again at the end of the year. **Figure 2** shows the water inflow and electricity demand over the year.

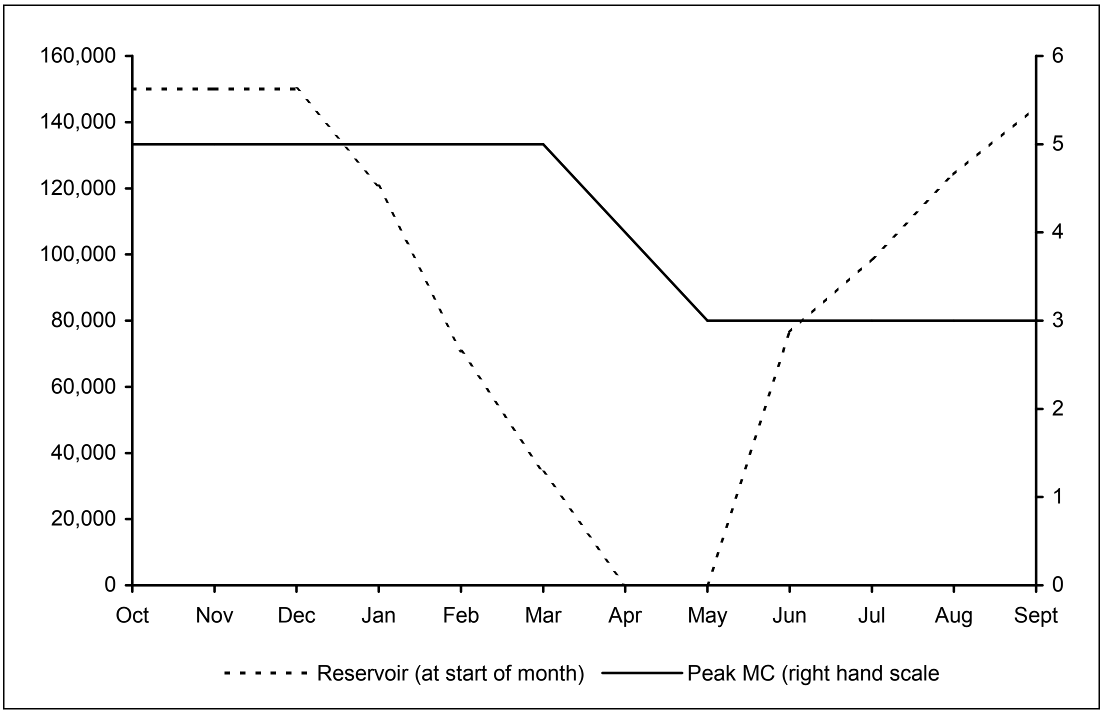
Figure 2: Water inflow and energy demand patterns



Water inflow is low in October and November and then falls to zero because precipitation falls as snow. Inflow to the reservoir only resumes with the spring melt in April. Thus, from October through March, thermal output is necessary to meet the excess of the amount of energy to be supplied over what can be produced by using all the October and November inflows and emptying the reservoir. This thermal output will be produced at minimum cost if it is produced at a constant rate throughout these months, the marginal thermal plant having a marginal cost of 5. All the variation in energy required between peak and Off-peak and between these months will be taken up by varying hydro output.

From May onwards, the water inflow suffices both to refill the reservoir and to produce sufficient hydro energy to allow thermal output to be reduced. Again, the minimum cost solution is to run thermal plant at a steady, but lower, rate. The plant with a marginal cost of 5 is now not required and the lower marginal cost of 3 of a higher merit plant sets the marginal cost. Thus because the timing of use of the available water in each half of the year is flexible, peak and Off-peak marginal costs are both 5 throughout the first six months and 3 throughout the last five. **Figure 3** illustrates the relationship between peak marginal cost and water availability.

Figure 3: Water availability and peak marginal cost



It is evident that different time patterns in water inflow and demand can lead to very different results, particularly if, as was not the case in this example, thermal or hydro capacity is limited.

- Optimising investment

A similar variety of cases can exist if long-run rather than short-run marginal costs are considered[7]. To give an example, again an extremely simplified one, consider a power system where:

- there is an excess of 12-hour daytime over 12-hour night-time demand, both of which are fairly constant throughout the year;
- this excess is provided by hydro generation, thermal plant meeting most, but not all of the base load;
- all the available water is used, reservoir capacity being sufficient to accommodate seasonal variations in inflow and
- the thermal plant capacity margin is no greater than required to provide security.

Under these circumstances, any increase in energy output will require the installation of new thermal plant. Consider an increase in daytime load averaging 1,000 MW. The cheapest way of providing it will be to add sufficient new thermal capacity to generate only an additional 500 MW but to run it on

[7] Turvey, R and Anderson, D (1977) *Electricity Economics*, Johns Hopkins University Press for the World Bank, Chapter 15.

base load. Daytime hydro generation will then have to be raised by 500MW and night-time hydro generation diminished by the same amount. Five hundred (500) MW of additional hydro generating capacity will have to be installed to cover the increase in peak demand. Thus the long-run marginal cost is the cost of building and running 500 MW of new base load thermal plant, plus the cost of adding 500 MW of generating capacity at the hydro stations.

An actual case, of Thailand, analysed in of *Electricity Economics* by Turvey and Anderson[8] was slightly different. There was a daytime peak, a daytime plateau and a light load period at night. Thermal plants, running all the time, provided a similar pattern of output, the excess of the load over thermal output in all three periods being met from a large hydro plant. Since there was no spillage, any energy increment or decrement would have had to be met by varying thermal output. Thus an increment would be provided by additional thermal output during the low load period. An increment occurring at other times of the day would be met by extra hydro generation, offset by a corresponding reduction during the low load period. Thus upward short-run marginal energy cost was the same for all the twenty-four hours of the day; indeed, for the whole year, since, although reservoir levels varied seasonally, the situation was the same throughout.

- The Blue Nile case

A totally different situation was analysed in *Electricity Economics* by Turvey and Anderson for the Blue Nile grid in Sudan[9]. The system comprised diesels, a gas turbine, steam turbines and two dual purpose hydro schemes, providing both electricity and water for irrigation. Water releases from the two dams for each twenty-four hour period were determined exclusively by the irrigation authorities. The despatchers, therefore, could only choose whether or not to turbine the water and the timing within the twenty-four hours. The pattern of Blue Nile river flows divides the year into three phases:

1) The *dry season*, when releases exceed inflows. During this phase, reservoir levels fell, gradually reducing the maximum power output of the turbines. The releases were, however, sufficient to allow the continuous operation of the turbines, except perhaps in a dry year (year to year fluctuations in river flows have been very considerable).
2) The *flood season,* when flows could be 100 times the dry season flows and carry masses of silt which could silt up the reservoirs. To avoid this, reservoir levels were kept as low as possible and all the inflow sluiced or turbined. The resulting low heads and higher water levels at the tailraces reduced the maximum power output of the turbines by some 30%.
3) The *wet season*, after the peak of the flood, both allowed the reservoirs to be refilled and allowed more water to be released than needed for full output from the turbines. The turbines could operate at full capacity once the reservoir was filled.

[8] *Op. cit.* Chapter 3

[9] *Op. cit.* Chapter 5.

In these circumstances, thermal supplementation of hydro output was necessary to provide:

- capacity and energy in the flood season;
- energy at the end of the dry season or, in a dry year, for more of the dry season;
- peak hour capacity almost every day of the year.

At other times the marginal cost of generation was zero. The result was a set of marginal cost estimates displaying a fairly complicated time pattern, even assuming mean expected river flows.

Electricity transmission

Whereas *probabilistic* methods are commonly used for estimating generation costs and the probability of loss of load, a *deterministic* approach is usual for transmission. Planning and evaluation of transmission considers one or a few system load conditions, usually peak load flows. It looks for overload or voltage violations for specific outage contingencies, that is to say for each of a large number of alternative credible outages of one or more components of the transmission system. The probabilities of each such contingency are not assessed. If they were, comparisons of the mean expectation of outage costs with the cost of the transmission system would be used to dimension the transmission system. Instead, a rule of thumb, such as ensuring the ability to withstand any one or any two credible outages is generally applied[10].

For large systems such as that of the National Grid Company, estimations of marginal costs require comparisons of costs over a period of years under alternative scenarios concerning load growth and the siting of new power stations, two independent sources of uncertainty. The number of possible scenarios is enormous, as probabilities cannot be attached to them, and because preparation of a plan for each requires consideration of a whole series of alternative credible outages. Computations of a set of expected marginal transmission costs between each pair of points (or even regions), therefore, would be extremely complex.

Long-run marginal cost pricing was considered and rejected in the National Grid Company’s 1992 Use of System Charges Review[11]. Three problems were judged to rule out such complex estimates:

[10] At least for small systems, it is now becoming possible to simulate reliability and estimate the probability of lost energy using data on load flow and records of circuit outage rates to do this for the existing system configuration and for alternative transmission projects. Such calculations will produce estimates of outage frequency and duration, of the number of customers interrupted and MW of load interruption. "Development of reliability targets for planning transmission facilities using probabilistic techniques - A utility approach" in *IEEE Trans on Power Systems* vol.12 no.2 May 1997 pp.704-9

[11] National Grid Company (1992) *Transmission Use of System Charges Review.*

1) It is difficult to establish a realistic base scenario. The reference scenario (called SYS) used by the National Grid Company in its Seven Year Statements, covers a much shorter period. Even though it uses the best demand forecasts available, the generation assumptions will probably not be fulfilled. These are that:

   - all the plants for which connection has been contracted will be built (from September 1992 to December 1995 connection contracts for 7,729 MW of new plant were terminated) (Monopolies and Merger Commission 1996);

   - no further applications to connect will be made;

   - there will be no plant closures (for which only six months' notice is required).

   The National Grid Company does construct plausible scenarios which forecast future generation projects and closures, but, as OFFER put it 'they are, by their nature, commercially confidential as they are influenced by information supplied informally to the National Grid Company by users of the system and by the National Grid Company's own judgements' [12].

2) Results would be very sensitive to the size of demand and generation increments (or decrements) assumed. For example, an increase of 300 MW in generation at a node may save investment, whereas an increase of 350 MW could give rise to major system reinforcements.

3) Detailed system planning studies would be necessary, examining changes in generation and demand at up to 400 nodes,

As an alternative, the National Grid Company calculates what it calls *Investment Cost Related Prices* using a simple transport model. This uses the following data:

- the annuitised gross asset value plus maintenance cost of lines, cables and substations per MW km of its system;
- next year's forecast peak demand at each Grid Supply Point;
- the registered capacity of each generating plant linked to its system;
- the distances along existing routes between these nodes,

and minimises the cost of the sum of (peak flow x distance x cost per MW km) for all routes. Instead of merit order operation it is assumed that input from each generator is its capacity scaled down by a uniform factor to make total generation equal the sum of peak demands at Grid Supply Points. The resulting dual for each node is the marginal cost of an increase of 1 MW in its net input or output.

[12] OFFER (1996)

This simple transport model of an imaginary system ignores physical laws relating to flows in electrical networks. It also neglects security. The National Grid Company does have an extension to it, the SECULF model, which brings in the cost of providing security. In essence it requires that no circuit must overload even if there are outages of any one or two circuits. Some 100,000 outage contingencies have to be considered one by one, a DC load flow approximation being used. But the possibility of using this for estimating nodal marginal costs was dismissed with curious brevity in the one sentence: ‘The nodal prices from the SECULF model are sensitive to a number of network dependent features and they are not considered suitable as a basis for the zonal tariff.’[13]

Nevertheless, it seems probable that such an analysis would yield excesses of marginal costs of an imaginary new system over those derived from the simple transport model which were not uniform over all nodes. One would expect them to be greatest in net import zones. However, the present system of a uniform addition to the marginal costs derived from the simple transport model implausibly implies that the marginal costs of adding security in net export zones would be the same as in net import zones. OFFER too is ‘not satisfied that the present treatment of security-related costs is justified’ being concerned that the security charge is ‘not properly distance-related’[14]. The importance of geographical differences in marginal costs follows from the propositions in the National Grid Company’s 1999 Seven Year Statement that in the Humberside and North Wales zones, very little additional generation could be accepted without transmission reinforcement to the North to Midlands boundary, whereas considerable new generation could be accepted without such need in the Southern zones.

In fact, in 1992, the National Grid Company did estimate long-run marginal costs properly, by comparing the differences in investment costs between two scenarios and a base case with the differences in revenue yielded by the Investment Cost Related Prices for transport in the imaginary system[15]. Under a high demand growth scenario, with new generation mainly in the North and with closures concentrated in the South, incremental revenue would fall significantly short of incremental investment cost. Under low demand growth and with new generation concentrated in the South, revenue would follow capital expenditure closely. It seems odd that the rejected method should have been used to argue that the much simpler adopted method was acceptable because the ‘high case observation’ was ‘balanced by the low case position’.

It has to be recognised, however, that computation of one-at-a-time long-run marginal costs for a whole series of combinations of export and import zones would necessarily be difficult as well as laborious. It is normally forgotten that marginal costs should be estimated by examining both increments and

[13] National Grid Company (1992) *Transmission Use of System Charges Review*, p.54.

[14] OFFER (1996)

[15] National Grid Company (1992)

decrements – the effect upon system expansion plans of the two are rarely symmetrical. If the base scenario already included the increase or decrease in load for which a long-run marginal cost was to be computed, the computation would require cost comparison with the base scenario minus the proposed reinforcement, but possibly plus some alternative hypothetical system reinforcement. If the National Grid Company expected that a second load or generation increment or decrement to the same part of the network would be requested within a year or two, the optimal system reinforcement might be larger or smaller than the first change alone warranted. Thus the application of long-run marginal cost pricing would be complex.

Electricity distribution

Distribution costs are much more difficult to deal with than generation costs. Their planning is usually decentralised as computer aided network design for optimising reinforcement can be used only for subsystems[16]. Data on maximum demand at different points in a distribution network are practically never available – metering at all but the largest substations is rare and loads on individual distribution transformers are unknown. Reinforcements are triggered by the construction of new buildings, the expectation of load growth from existing consumers or, more retrospectively, by complaints about low voltage or unreliability and by inspection of the state of individual parts of the network. The cost analyst therefore has to invent a series of estimating procedures in order to use whatever data can be obtained in each particular case in order to:

- gross up generation costs to allow for losses, and
- estimate marginal capacity costs in transmission and distribution.

Here is an example, in which no sophisticated analysis of transmission losses and costs is possible. The estimates become less and less reliable as one proceeds down the rows of each of table 4 and table 5. **Table 4** shows loss multipliers, that is ratios by which marginal generation cost has to be multiplied to obtain marginal cost per kWh delivered. The figures, which are perhaps untypically high, are rounded.

[16] Lakervi, E and Holmes, E J (1989) *Electricity distribution network design*, Peter Peregrine for the IEE, Chapter14.

Table 4: Transmission and distribution loss multipliers

	Annual average energy	Annual average energy cumulative downstream	Annual average energy cumulative downstream	System peak load	System peak load cumulative upstream
			↓		↑
	Input ÷ output	Input ÷ output	Marginal	Input ÷ output	Input ÷ output
EHV transmission lines	1.02	1.02	1.03	1.03	1.25
EHV/HV substations	1.01	1.03	1.05	1.02	1.22
HV transmission lines	1.02	1.04	1.09	1.03	1.19
HV/primary substations	1.02	1.06	1.13	1.03	1.16
Primary distribution lines	1.02	1.08	1.17	1.03	1.12
Distribution transformers	1.03	1.11	1.24	1.05	1.10
Secondary distribution	1.01	1.12	1.27	1.04	1.04

Except for transmission, which is usually continuously metered, peak losses have to be estimated by using a formula such as:

> Peak loss % = Annual average energy loss % x ($0.7LF^2+0.3LF$),
> where LF is Load Factor.
>
> *Note that*, with a different system configuration, where peaking plant was located in load centres while base load plant was more distant, the peak ratio of transmission input to output would be lower than the annual average ratio, not higher.

The implication of these figures is that, as shown in the third column in Table 4, a 1 MWh increase in annual consumption at low voltage requires an additional 1.27 MWh of generation, given transmission and distribution capacity. But if this capacity is expanded proportionately with the load, resulting in average loss percentages remaining approximately constant, an additional MW of low voltage demand at time of system peak will require 1.25 MW additional peak generation and, for example, additional primary substation capacity of 1.16 MW (see far right column in Table 4).

Table 5 computes the marginal transmission and distribution capacity costs per MW of low voltage consumers' maximum demands, based on the assumption that capacity will continue to grow and that future relationships between capacity and the load forecast made or two years previously will mirror the experience of the past few years as recorded in the second column. Since maximum demands do not all occur at the same time, capacity at each stage needs to change by less than the sum of the changes in the separate maximum demands from the stage downstream.

Table 5: Marginal transmission and distribution capacity costs of low voltage consumers' maximum demands (per MW)

	Marginal km or MVA per MW of own maximum demand	Unit cost annuitised over useful life plus annual maintenance cost	Marginal cost per MW of own maximum demand	Coincidence factor ie maximum of total demand ÷ Σ maximum demands	System peak load Cumulative upstream input ÷ output	Components of marginal cost per MW of LV maximum demands
EHV transmission lines	0.35	14,500	5,075	0.95	1.25	6,027
EHV/HV substations	2.7	1,350	3,645	0.95	1.22	4,225
HV transmission lines	0.4	15,000	6,000	0.9	1.19	6,426
HV/primary substations	1.8	1,600	2,880	0.9	1.16	3,007
Primary distribution lines	0.001	17,400	17	0.8	1.12	16
Distribution transformers	1.3	1,980	2,574	0.8	1.1	2,265
Secondary distribution	0 (for existing consumers)				Sum	21,966

Different parts of the system (for example those serving industrial or commercial areas and those serving residential or rural areas) may have different load shapes and different coincidence factors. Furthermore, the shares of forecast load growth from new connections and from load growth on the part of existing consumers may differ. Load growth from new connections may require much greater additions at medium and low voltage. In such cases, the calculations relating to the lower voltage levels may be carried out separately for the different areas and categories of consumer to yield different marginal capacity cost estimates, some relating to the number of consumers rather than to maximum demands.

Gas transmission

British Gas Transco calculates *long-run marginal costs* for the purpose of setting prices for the use of the National Transmission System. This is the high pressure network that transports gas from entry points to local transmission and distribution systems and large volume consumers. The capital expenditures on the National Transmission System constitute only some 18% of Transco's projected gross investment. Investment in Local Transmission System and Distribution, for which marginal cost calculations are not made, add up to almost as much. Investment in meters and replacement make up most of the balance.

Here I am concerned only with the methodology used in the cost analysis of the National Transmission System[17], not with the way the results are used for price setting. The calculations are made on the basis of a ten year development plan. The development plan reflects:

- forecast changes in input points for beach supplies;
- the development required by inauguration of the European Interconnector;
- planned major plant replacement;
- the forecast growth in 1 in 20 peak day firm demand.

This demand is forecast yearly for each of the many offtakes to Local Distribution Zones and to large consumers, mainly power stations. The matching pattern of supplies at each delivery point is also forecast. Flow patterns in the system are then analysed by use of a computer software package to identify constraints which will need to be removed. This can be done by uprating pipeline pressures, building new pipelines, uprating compressors, installing new compressors or providing additional offtakes. The options are examined to select a sequence which represents an efficient and economic solution whilst maintaining system security. The resulting plan takes the form of a set of committed projects for completion by the end of next year and a set of planned projects for the rest of the ten year period. Capital expenditure is forecast year by year at today's standard construction cost.

The development of an optimal or near-optimal plan for such a system has to take account of intertemporal relationships. The constraints that analysis reveals for any year will depend upon how the system was reinforced in preceding years. A larger reinforcement than is required for any one year may prove cheaper than a series of smaller reinforcements undertaken over several years. Such interdependencies could be dealt with by one huge multi-period optimisation model if such a model, comparable with those used for planning electricity generation, were available for gas transmission. But this is not the case. The complex software named Falcon used for analysing the daily dynamics of the system, which takes account of twenty-five factors affecting gas

[17] Transco (1998) *Transportation Ten Year Statement 1998*, Appendix 11. (*Gas transportation charges from 1st October 1999),* Transco 1999, pp 32-38, also provides an explanation.

flow, is a single-period model (as is the new software developed for cost analysis, named Transcost).

Calculations have been made for all years of the ten year period. Each annual calculation takes the forecasts for that year and the physical configuration of the system as given by the ten year plan. First, flows and pressures within the system are calculated and then the nature and costs of pipeline and/or compressor reinforcement necessary to accommodate an extra 100 million cubic feet peak day flow are calculated iteratively, the length of required additional pipeline being calculated to the nearest kilometre. This is done separately in turn from each of six input points to each of 127 offtakes. For each such combination, a calculation is made using standardised investment costs for pipelines, compressors and regulators plus project management costs for each year as if there had been no increment in the preceding years of that period.

Leaving out operating costs (which are simply set at 1.5% of capital cost), marginal costs for each such combination are estimated as:

$$\frac{\sum_{y=1}^{y=10} d_y \left(a_{20} C_y\right)}{\sum_{y=1}^{y=10} d_y I}$$

C_y is the cost of the investment in year y required by a peak increment in that year of I (set at the same amount in all years). Multiplying by a_{20} transforms this capital value into a twenty year annuity – twenty years is the anticipated life of new pipeline assets. d_y is the discount factor that yields the present worth of a year y magnitude.

This is an annuitised weighted mean of unit incremental peak costs where the weights are the discount factors d_y, as is readily apparent if the expression is transformed into:

$$a_{20} \cdot \frac{\sum_{y=1}^{y=10} d_y \frac{C_y}{I}}{\sum_{y=1}^{y=10} d_y}$$

It is interesting to compare this with what may be considered more relevant for setting prices now, namely the marginal cost of one increment lasting for ten years instead of ten alternative one-year increments. In this case, the model would be used to estimate in each of the ten years the cost of necessary reinforcements on the assumptions that:

- each year's peak load to be met is the load forecast of the ten year plan plus the enduring increment;
- each year starts with a physical configuration of the system which is in accordance with the ten year plan plus all the reinforcements already made in previous years to accommodate the increment in demand.

Let the sequence of investments computed as necessary to meet this be denoted K_y. With a useful life of 20 years, each of these additions to capacity will have a residual value at the end of the ten-year period which should be offset against its capital cost. This can be allowed for, as explained earlier, by treating the cost during the ten years as the annuitised cost each year of all the investment undertaken up to and including that year. Thus the cost in all years *y* to *10* of the year *y* investment will be $a_{20} \times K_y$. Adding up and discounting back to the present for all investment to date over all ten years gives:

$$\sum_{y=1}^{y=10}\left[\sum_{j=y}^{j=10} d_j \cdot a_{20} \cdot K_y\right]$$

as the present worth of costs. This is the sum of the present worths of a set of annuities which commence in those of the years 1 to 10 when there is investment and each of which continues until the end of year 10. Dividing by the present worth of the ten-year increment $\sum_{y=1}^{y=10} d_y I$ yields a long-run marginal cost of:

$$a_{20} \cdot \frac{\sum_{y=1}^{y=10}\left[\sum_{j=y}^{j=10} d_j \cdot \frac{K_y}{I}\right]}{\sum_{y=1}^{y=10} d_y}$$

where K_y will presumably be less than C_y in years 2 to 10.

This marginal cost concept, unlike Transco’s, does allow for backward-looking intertemporal interdependence in the sense that the physical configuration of the system assumed for each year includes the incremental investments of previous years.

But neither concept allows for full optimisation. This would take account of forward-looking intertemporal interdependence as well, by recognising that the optimal choice of reinforcement in any year will partly depend upon the options for future reinforcements. Investing more than the minimum required now may save future investment whose discounted cost is greater. Thus the series of investments K_y determined by the sort of calculation described above may be larger than ideally necessary, so that the marginal cost of a ten-year increment that would be derived from a fully optimised multi-period planning might be less.

The two cost concepts for which the model can be used answer different questions. Transco enquires, what would be the marginal cost on average over the ten years, of ten one-shot one-year increments of demand. It treats capital costs as if the reinforcements could each be rented for only one year, at a rent equal to their annuitised values, since they are assumed not to exist in succeeding years; and it computes a curiously weighted mean rather than a simple mean of the ten costs. The alternative method enquires what would be the cost, expressed as ten equal annual amounts, with one ten-year increment. Both, however, are ‘long-run’ in the sense that they suppose that enough notice of the increments to be given for Transco to be able to reinforce the system in time to provide for the increments.

Whether applied as in Transco's calculations or as with the alternative marginal cost concept presented here, the approach only estimates the cost of meeting increments in the load. Applying it to select components of the ten year development plan which could be cancelled or postponed to deal with a decrement in the load would be more complicated. Yet if the forecasts. used in the base plan are the best forecasts that can be made, downward revisions are as likely as upward revisions. Marginal cost ought therefore to be computed as the mean of marginal cost upwards and marginal cost downwards.

Water supply

Resources

In a document on the economic level of leakage[18] Yorkshire Water take the sensible position that this should be examined by including schemes to reduce leakage as well as schemes to provide new resources when choosing the investment programme that minimises the present worth of the sum of operating and capital costs. Their method of identifying the least cost solution is exactly in accord with the approach that I have suggested in the first part of this paper. This section describes some features of their analysis that reflect the particular characteristics of water supply.

For each candidate future scheme, the main data required are as follows:

- total eventual yield and the ramp-up profile to this yield
- capital cost, and replacement capital cost at the end of its life, for each of its different components such as land, civil works, mechanical and electrical equipment
- the timing of capital expenditure over the construction period
- the life of each component
- its fixed annual operating cost and its variable operating cost per megalitre
- its fixed annual impact and its variable impact on system operating cost, that is the pumping and treatment costs of the existing system
- its fixed annual and variable environmental/external costs
- the first possible year of use of the scheme
- schemes, if any, which must be built before it.

Using a selected discount rate, the present worth of the cost of each scheme could then be divided by the present worth of its full capacity yield and the schemes ranked in increasing order of this unit cost. Schemes would then be selected and timed according to this order so as to reduce the projected supply-demand deficit to zero in each future year.

This would be too simple, however. The resulting programme is unlikely to be optimal because of the lumpy nature of some schemes. Consequently, integer programming is used to compute the optimal programme, namely that which meets the forecast required demand over the next forty years while minimising the present value of all costs. Capacity is selected to match forecast dry weather peak demand, while operating costs are in effect computed for average demand with average weather by means of the introduction of a utilisation assumption. Capital components whose life expires within the forty years are assumed to be

[18] Yorkshire Water Services Limited (1997) *Establishing the economic level of leakage*. The text that follows is taken from the Appendix including Appendix 9.

renewed; components *x%* of whose life extends beyond that period have *x%* of their cost credited as a residual value at the end of it.

The calculations relate to single-valued forecasts of average daily demand, ignoring seasonal variations in demand and desired 'headroom', that is any required reserve margin of capacity. The need for a reserve margin and seasonal variations or peak demands could, if necessary, be dealt with in the same way as in the simple constructed example presented above, though this would increase the complexity of the programme.

The leakage report does note that the approach could be improved, stating that:

> 'The treatment of system effects could be made more comprehensive, by, for example, covering the degree to which sets of options allow valuable changes to reservoir control curves to be implemented without reducing security.'
>
> 'Rather than treating supply security as a parameter…supply security could be treated as a variable to be optimised jointly with everything else.'

One way of achieving this would require a probabilistic analysis. A penalty value would be imputed for shortfalls, for example, for hosepipe bans, and the present worth of costs including the penalties would be minimised, with shortfalls treated as a 'source' and costed at their penalty value multiplied by their probability.

The operational reasonableness of the best programme is checked using a model of the network to determine whether transmission capacities will allow forecast demands to be met at the requisite level of service (this is defined, for example, in terms of the probability of rota cuts). The results may indicate a need to provide extra transmission or treatment capacity. If so, appropriate extra costs will be added in. To find the marginal costs it is simply necessary to add to (or subtract from) the future demands and re-solve for the optimum plans.

Distribution

Now consider marginal distribution capacity costs. Investment is lumpy in each separate part of a network, in that new mains are sized to provide for several years' growth in demand. But if marginal costs are to be estimated for the company as a whole, or at any rate for large supply areas, an aggregative and thus smoother relationship between demand and growth in the length of mains is relevant. Since detailed plans for expansion of the network, in contrast with plans for sources, are not made for more than a year or two ahead, the relationship can only be quantified by first analysing past patterns of growth and then extrapolating and possibly adjusting them to allow for any foreseen differences between past and future circumstances.

The case relates to what, many years later, became Thames Water, and the analysis was carried out separately for each of three areas. Here, only one of them is considered.

The first step was to weather-correct past data on annual average daily supply. In principle, the analysis could alternatively have been conducted in terms of maximum annual hourly demands (or maximum daily demands when service reservoirs balance out diurnal fluctuations), with an allowance for peaks exceeding the seasonal norm during extra dry weather. In contrast with electricity, however, expressing marginal cost as a function of instantaneous maximum demands is not interesting.

The length of mains of each diameter existing at the end of the accounting year was then regressed against weather-corrected annual average daily supply, using thirteen years of data. Mains sizes whose total length was either negligible or practically constant were omitted. Mains larger than 36” were also omitted, being parts of new supply schemes rather than part of the distribution network. The total length of 3”, 5” and 7” mains actually fell, for they were in effect replaced by 4”, 6” and 8” mains, so regressions were calculated for 3”& 4”, 5”& 6” and 7”& 8” mains. The marginal regression coefficients related to extra yards of main per extra thousand gallons a day. The fixed terms were nearly all positive, indicating a less than proportionate increase in mains length, which is to be expected when demand increases but the area supplied does not. The one exception was for 18” mains, because it became a preferred size, 15”mains becoming non-preferred. Similarly 12” became preferred and 10” non-preferred. Allowance for leakage had to be made, frankly by guessing, and up to date costs per yard laid, including reinstatement were ascertained.

The results were as follows in **Table 6**:

Table 6: Marginal water distribution capital cost

Main size	Marginal length (yards)	£ Mainlaying cost/yard	£ Marginal cost
3” & 4”	3.43	18.6	63.8
5” & 6”	3.22	21.8	70.2
7” & 8”	0.57	26.6	15.2
10”	0.24	31.0	7.4
12”	1.06	35.2	37.3
15”	0.14	45.4	6.4
18”	0.44	55.9	24.6
24”	0.45	78.0	35.1
30”	0.30	108.7	32.6
36”	0.35	135.1	47.3
		$\sum$	339.9
		Adjust for leakage	377.6
		Add cost of service reservoirs, £35 per ‘000 gallons	35.0
		$\sum$	412.6
		Divide by 365 to obtain cost per ‘000 gallons	1.13
		Annuitise over 80 years at 10%	**1.3 pence**

Railway marginal costs

Short-run track costs

Electricity consumption

With electric traction, and in the absence of electricity meters in locomotives, the electricity consumption involved in a round trip has to be estimated as a function of the route, of distance travelled, the characteristics of the locomotive and train and the average speed. To convert these estimates into monetary terms, account has to be taken of both regional and time-of-day variations in electricity charges.

Track maintenance and replacement

Estimating the effects of running more or fewer trains on track maintenance and replacement requires engineering parameters for each of a number of track/sleeper types for each of a number of track quality categories for each of a number of locomotive, carriage and wagon types. They are necessary in order to estimate the marginal effect of an incremental or decremental effect per mile traversed upon five activities which must be undertaken in order to maintain set standards. These five are:

- rail maintenance, that is, replacement or spot welding of individual rails
- track geometry maintenance, that is, tamping or stoneblowing
- rail renewal
- sleeper renewal
- ballast renewal.

Estimating the effect of the passages of vehicles of different types upon the amount of maintenance required and the frequency with which renewals have to be made involves the application of engineering measurement, experimentation and experience to formulate and quantify a large number of complex engineering functional relationships. Thus, to give but one example, the total rail failure rate on continuous welded rail is the sum of five different kinds of failure rates which are functions of such variables as number of axles, wheel force, and wheel/rail contact stress.

Unit costs have to be estimated by which the physical effects on maintenance and renewal can be multiplied in order to end up with marginal cost estimates. Annuitising renewal costs over target life yields the cost incurred or saved by a one-year bringing forward or postponement of renewal.

Short-run congestion costs

Nature of marginal costs

The effects upon delays and cancellations of existing trains of adding or subtracting a train running subject to a standard distribution of delays. The costs of these delays are borne by train operators (operating costs) and by passengers. To estimate the latter, three steps are required:

- estimating delays to trains
- converting into delays to passengers
- valuing passenger delays.

Delays to trains arise because one more or one less train will reduce or increase the gap between trains, thus reducing or increasing the time available for recovery without disrupting the running of other trains. They differ from one route section to another because of differences in network reliability and the number and timing of existing trains, which varies by time of day and week, and because of differences in the number of passengers on the existing trains whose punctuality is affected.

The effects of adding or removing a train depend partly upon whether there is some flexibility in timetabling. If timetable adjustments to existing trains are feasible and there is some freedom regarding the specific timing of any new train, its effect upon punctuality will be less than if it and existing trains are ‘hard-wired’, for example with specific departure time. In the case of withdrawal of a train from the timetable, the resulting improvements in punctuality will be greater if there is flexibility. However the methods used to estimate congestion effects do not allow for such flexibility, but relate to the effects of a specific train added to or removed from a specific timetable.

Estimating train delay effects

The effects of additional or withdrawn trains on a route segment or segments of up to 250 miles in length can be estimated using the proprietary MERIT model. This is a relatively detailed Monte Carlo simulation covering 100 days of operation. Once calibrated on observed data, (the actual times trains reach a large number of points are monitored) the model can be used to analyse the effect of a new timetable (or of an infrastructure alteration). Calibration and the subsequent runs require:

- a specification of the infrastructure, including the broad layout of the route section, number of stations, number of signals, route speed and temporary engineering restrictions;
- a detailed working timetable for the route section;
- frequency distributions for each of a number of types of incidents, including track-based incidents, such as power supply problems, points and signal failures and vandalism, and train-based incidents, such as locomotive

breakdowns and late starts;

- delay distributions for each type of incident;
- control rules for junctions when decisions about priorities have to be made.

The model provides estimates of the punctuality of trains at key points along the route segment.

Another, less precise, simulation model, PSP, uses the actual performance of trains in the current timetable rather than, as with MERIT, frequency distributions of incidents. For each minute of the day in turn, for each of a sample number of days, it adds a new train of a specified type, with specified timing and routing. It models conflicts with the actual timing of existing trains and records the resulting delays both to itself and to those existing trains. Thus it can be used to estimate the delay impact of an additional train and to find potential new train paths with minimal marginal delay costs. Sample runs for a few routes show that the delays vary from minute to minute within the peak hour 17.00 to 18.00 and, looking at the whole of the twelve hours from 07.00 to19.00, that there is no simple peak/off-peak dichotomy.

Passenger delay effects

To get from delays to trains to the delays inflicted upon the passengers using them, data on passenger flows are required. Complete centralised records obtained monthly of all ticket sales by origin and destination, route and class of ticket, if they exist, though necessary, are not sufficient. Survey data on individual train loadings will also be required.

Putting a value on a minute of delay for an average passenger in excess of some threshold can best be regarded as a policy decision. The higher the value chosen, the higher will be the capital and operating costs which the policy-maker deems worthwhile to incur in order to provide punctual services. The choice of the value, particularly if it is manifested in the form of fines levied upon train operators for unpunctuality, provides them with incentives to search for a minimum cost solution.

In order for the policy-maker to structure the decision about the value of delays, it can be helpful to start with estimates of the value of usual travel time as revealed by the choices passengers make between alternative services with different travel-time/ticket-price combinations. Then a judgement can be made about (i) the value of delays expressed as a multiple of the value of usual travel-time and (ii) the minimum delay which shall thus be valued. For example, a delay to long distance passengers of more than ten minutes and a delay to commuters of more than three minutes might both be judged to impose a cost upon them of three times the estimated value per minute of their respective usual travel times.

Long-run costs

Railway network enhancements are planned to relieve congested parts of the network. The most important kinds of enhancements consist of track upgrades, realignments and recants, signalling upgrades, additional tracks and crossovers, bridge improvements and power supply upgrades. These are lumpy investments, so require the kind of long-run marginal cost analysis where, instead of postulating a change in output and ascertaining the cost of the required change in capacity, a feasible change in capacity is costed and the change in output which it will allow is ascertained. As part of their investment planning process, railway administrations periodically produce a set of candidate enhancement projects, each with its approximate cost estimate.

The effects of enhancements can be estimated using analytical tools such as the Planning Timetable Generator, being developed by consultants. This allows non-marginal timetable changes to be planned, using advanced stochastic optimisation heuristics. It is intended to replace the extremely laborious manual creation of timetables for planning purposes. Given:

- a description of the layout of the infrastructure on a route
- specified frequency of trains and their stopping patterns
- requirements for:
  1) even-interval departures
  2) clockface departures
  3) limits on stock turnaround times
  4) need to make particular connections.

It generates a timetable which minimises a weighted sum of divergences from these requirements. Hence it can be used to predict the effect upon capacity of candidate infrastructure enhancements, with the new timetables it produces being examined using MERIT.

The results of an enhancement for the part of the network where it would relieve congestion may be any or all of: (i) more trains per day, (ii) higher speeds (reduced timetabled journey times) and (iii) improved reliability (lower delays and fewer cancellations arising from incidents). With a small percentage of enhancements there are tradeoffs between these three dimensions of output. In such cases, the gain in output can only be expressed as a single figure if equivalences between them have been determined. However, in a majority of candidate projects the gains are predominantly or entirely one-dimensional. But there is another complication, namely that the gains from different candidate enhancement projects lying along a major route are not always simply additive. For example, for just one part of a route the ability to run more trains per day may not be very valuable, but it may be very large for the route as a whole: the benefit from several projects on that route together exceeding the sum of their separate benefits.

Even if the ranking of projects in order of benefits per pound of capital cost requires subjective judgement, it is clear that marginal costs will rise with the volume of enhancement expenditure. This is simply because some congestion-reducing enhancements will be easier, quicker and simpler to undertake than others.

TRAPS TO AVOID

This section identifies those traps which must be avoided when calculating marginal costs.

Treating depreciation as a cost component

The calculation of marginal cost does not involve the use of any accounting measure of depreciation. Depreciation spreads out the time stream of capital outlays at historical or replacement cost into annual chunks according to a conventional formula so that annual income can be computed. Like other accounting conventions, this is necessary for company reporting and for the purposes of taxation, despite the arbitrariness of the choice of formula and of the use of conventional assumptions about asset lives.

There is, however, a different and more meaningful concept of annual depreciation which is consistent with the analysis of marginal costs[19]. It is the year to year decline in the ‘deprival value’ of an asset. This is the increase in the present worth of future costs which would result were the asset in question to disappear in a ‘puff of smoke’. The calculation of this, the difference between the present worth of all future costs of meeting the output plan with the asset and what it would be without it, is, of course, just like the calculation of marginal costs. But it is unnecessary for that calculation.

Modern equivalent asset value

Some writers attempt to reconcile accounting costs and economic costs by using the concept of Modern Equivalent Asset Valuation. Such a valuation could be one of three things:

- ascertaining and summing the cost of replacing each item of equipment separately and adding them up;
- taking the historical costs one by one for each item of equipment, multiplying them by some price index and adding them up;
- estimating the cost of constructing a brand new system capable of producing the same outputs as the existing system.

None of these produces anything relevant to decision making. This involves the *future* costs that might be incurred from alternatives that are practical possibilities, not cost sums which would never be incurred.

Consider for example an MEA valuation of Clapham railway station. Even the single question of the value of the land it occupies defies rational examination. The cost of adding to the number of trains that can pass through it could only be

[19] As explained in chapter 6 of my *The economics of public enterprise* (Allen & Unwin), 1971.

ascertained by designing and costing the necessary works, whose only element in common with these fantasies would be the unit costs of some items of equipment.

Stand alone cost not relevant unless a plausible possibility

Consider the proposition that:

> Stand Alone Cost of a subset of services = Total Cost minus the Incremental cost of all other services.
>
> *where*, incremental cost means the cost saving that would result from not producing the other services.

This proposition lacks all practical interest because it only holds for imaginary new systems built from scratch. Consider working it out in practice for an *existing* system.

What would be the Stand Alone Cost of a new track, signalling and stations that would provide for the same volume of traffic between Basingstoke and Waterloo as is currently provided by Railtrack? What would its alignment be? How much would acquisition of the right of way cost? Would it have a third rail or an overhead power supply?

Remember that any accounting measure of total cost for Railtrack will differ from the total cost of a new built-from-scratch system, even if the depreciation component of the former values assets at their Modern Equivalent Asset value, whatever that is. The reason is that the operating cost of the actual system will differ from what would be the operating cost of such a new redesigned whole system.

The concept of scale economy only applies to forward-looking cost as a function of the capacity of any built from scratch *new* component, set of components or whole *new* system. Nothing whatever can be said a priori about the relation between:

- the forward-looking marginal or incremental (or decremental) cost of a service or set of services to be provided (or withdrawn) on an existing system;
- the total accounting cost of that system, built over the last 150 years under long-vanished circumstances, technology and expectations.

The following is an American example of a Stand Alone cost calculation[20]. The Interstate Commerce Commission applied a competitive entry test by comparing:

[20] ICC Report (1990) *Coal Trading Corp. et al versus The Baltimore & Ohio Railroad Company et al*, Appendix A.

- the revenues required to amortise the stand-alone capital and cover the operating costs of hypothetical railways, with
- the charges paid to the Baltimore & Ohio Railroad Company by the coal shipper complainants.

The Commission's Rail Costing Section produced a report on issues that were raised by the parties on appeal. The following description of some of them illustrates the problems of estimating stand-alone cost and of using it as the measure of what costs would be in the absence of barriers to entry.

1) *Stand-alone capacity*. Cost depends upon capacity – should this be for the incumbent's current output level or allow for future growth or decline?
2) *Barriers to entry.* Should the following costs be disregarded as barriers to entry: (1) the extra costs of rapid construction or the extra time required to avoid them? (2) the excess of land cost over the market values of the land required, reflecting the difficulty of assembling land for a continuous right of way? (3) the cost of the bridges now required to avoid level-crossings, the Baltimore & Ohio Railroad Company not having had to provide them when it was built? 'Yes,' said the ICC.
3) *Engineering issues.* Was single track with passing sidings sufficient for the traffic or would double track be required? Was wagon transit and detention time correctly estimated? Was a 20' roadbed good enough or should the 23'3" recommended by the American Railroad Engineering Association be assumed?
4) *Cost estimates.* Was leasing locomotives really cheaper than purchasing them? Was the sub-ballast cost quotation used acceptable (there were very many such unit-cost data issues)? Was 5% contingency allowance acceptable, or had contingency allowances been included in the individual cost items?
5) *Cost of capital*. What was the appropriate cost of equity and debt finance?
6) *Residual value.* If the computation covers, say, 20 years, what asset value is attributed to the assets at the end of that period?

It requires a great leap of faith to suppose that this approach has any relevance whatever to optimal pricing or to optimal resource allocation (or even to equity) today.

OFTEL's Long-Run Incremental Cost[21]

What OFTEL means by long-run incremental costs, appears from its statement that 'it is assumed that all assets are replaced in the long-run.' This turns out to mean that these costs are estimated for a hypothetical system incorporating 'the latest available and proven technology' but with the same topology as the existing system. On the 'scorched node' assumption, used for UK telecomms,

[21] Set out in *Pricing of Telecommunications Services from 1997* and *Annexes to the Consultative Document,* December 1995 Annex D.

access and conveyance costs for an inland Public Switched Telephone Network have been estimated both:

- by subtracting the value of assets used for other purposes, such as ISDN and virtual private networks, the top-down model;
- by adding up the costs, at today's prices, of the components it would require – the bottom-up model.

With due allowance for components which serve more than one purpose, that is, for common costs, the *bottom-up* stand-alone cost estimate should, on common assumptions, equal the sum of the *top-down* 'long-run incremental cost'[22] estimates for access and conveyance. The most important of these common assumptions relate to traffic, system topology, utilisation levels and routing factors and the way that annual equivalents of capital expenditures are derived.

What has been done, therefore, is to estimate the costs of a hypothetical new system confined to existing sites and routes. This is *entirely different* from estimating:

- potential new entrant costs (a new entrant would neither wish nor be able to occupy these sites and routes);
- marginal costs as defined and explained above (though both would use some of the same plant and equipment prices in their calculations).

The estimates pay no heed at all to plans and expectations concerning the future optimal expansion and operation of the system. No attention at all is paid to examining how system development and hence costs would vary according to traffic projections, the essence of marginal cost analysis. The estimates relate to the annualised costs of a hypothetical system existing at a point of time. Its hypothetical nature is illustrated, to take but one example, by the apparent need to discuss whether, in this system, there would be fewer bores in some ducts than actually exist.

Perhaps the use of this rather weird cost concept can be explained by the following factors:

- the baneful influence of the irrelevant constructs of academic economists;
- accountants can understand it;
- it is much easier to apply than the marginal cost concept, which requires more work and which involves facing all the uncertainties that affect forward planning for an undetermined future;
- its application in the particular case of telecommunications is less likely to threaten incumbent profits.

[22] 'Incremental cost' is curiously named because it means the *decrement* in cost that would result from permanently suspending production of an output or group of outputs.

CONCLUSIONS

The theoretical constructs of economics texts are of little use; the platitude that increasing returns to scale cause marginal to fall below average costs being one example, since it relates only to brand new built from scratch systems.

Marginal costs depend not only upon the timing of a postulated change in output but also upon the timing of the decision to adapt to it.

Marginal costs are forecasts, and forecasts are rarely accurate. However, all decisions are founded upon uncertain expectations about the future effects of current choices.

These forecasts cannot be made without the collaboration of engineers.

CRI ADVISORY COMMITTEE