Research Papers in Environmental and Spatial Analysis No. 111

Production and cost functions for utilities in an urban context: problems of specification and estimation

Hugh Wenban-Smith*

Published May 2006
ISBN 0 7530 2004 1

***Department of Geography and Environment
London School of Economics
Houghton Street WC2 2AE London, UK
Email: h.b.wenban-smith@lse.ac.uk

Abstract

In the utility regulation literature, the use of cost functions to assess economies of scale (and other economic effects) in network industries such as electricity supply is well-established. However, in applying the methods developed for electricity generation and distribution to water supply some problems have come to light. The source of the difficulty is that there has been a tendency in the electricity case to assume that production and distribution are separable. This assumption is not generally appropriate for water supply. This paper surveys the relevant literature and concludes that a proper assessment of scale effects in water supply requires use of either a composite production function or separate production/cost functions for water production and water distribution. If data limitations preclude these approaches, simpler methods using an aggregate production function or unit costs may still be feasible but the results may not be very satisfactory.
[Key words: urban infrastructure, production/cost functions, economies of scale, electricity supply, water supply.]

Introduction

Utilities (electricity and water supply, telecomms, etc) constitute an important part of the urban infrastructure. Gaining an understanding of the cost characteristics of utilities can therefore make a useful contribution to urban economics. A key feature of the situation is that utility services are delivered through networks so that the economics of distribution need to be considered in conjunction with the economics of production[1]. While the author has obtained some suggestive results for urban water supply using a simplistic approach based on unit costs – Wenban-Smith (2006) – it would be desirable to deploy more powerful methodologies based on production or cost functions. Such methodologies are well-established in the utility regulation literature and, on the face of it, it should not be too difficult to adapt them to analyse urban infrastructure. However, in seeking do this, some unexpected difficulties were encountered. The purpose of this note is to draw attention to these problems and to suggest how they might be overcome.

Schmalensee's approach

Schmalensee (1978) seems to have been among the first to give systematic consideration to the economics of distribution through a network. He was concerned that " … diagrammatic discussions of utility regulation often employ everywhere declining long-run average cost curves … [but] … When services are to be delivered to customers located at many points, cost must in general depend on the entire *distribution* of demands over space." To analyse the implications of this observation, Schmalensee constructs a simple model in which utility services are distributed to a circular urban area from a central point (the model considers only distribution costs, ignoring production). Demand per unit area, or *demand density*, is assumed to be a bounded non-negative function, $q(r)$, of the distance r from the centre. Total demand for

[1] It is another example of the general problem of location and pricing in a spatial economy, as discussed by Fujita & Thisse (2002, Ch.2).

services by those customers living between *r* and *r+δr* is *2πrq(r)δr* and the total service flow across the circle of radius *r* is given by:

$$Q(r) = \int_r^R 2\pi r q(r) dr, \qquad 0 \geq r \geq R \qquad \ldots\ldots\ldots \qquad (1)$$

The long run cost of transmitting a total service flow *Q* a small distance across a circle of radius *r* is *c(r,Q)δr*. This transmission cost function completely summarises the relevant technology (thereby abstracting, as Schmalensee remarks, from "a host of engineering problems and choices that confront actual utilities in real urban areas"). The total cost of distributing utility services in the area that would be incurred by a single firm can then be obtained as:

$$TC = \int_0^R c[r, Q(r)] dr \qquad \ldots\ldots\ldots \qquad (2)$$

Schmalensee then shows that global strict concavity of the transmission cost function, *c*, is a sufficient condition for natural monopoly in distribution (distribution cost minimized when all distribution is carried out by one firm, implying economies of scale with respect to volume distributed) and also derives certain necessary conditions. For present purposes, we simply note that whether the transmission cost function is concave or not is an empirical matter, and there appears not to be any reason why it should necessarily be so. It would seem to be necessary to examine some actual networks to learn more. Unfortunately, Schmalensee's specification does not easily lend itself to empirical investigation as the cost function *c(r,Q)* is not readily observable.

Use of cost functions in analysis of electricity supply

a. Nerlove (1963)

In a pioneering study, Nerlove (1963) analysed the production costs of 145 US electricity generating companies. According to Greene (2003, p.125)[2], this was among the first major applications of statistical cost analysis, and also the first to show how the fundamental theory of duality between production and cost functions could be used to frame an econometric model. The focus of the paper was the measurement of economies of scale in electricity generation, for which purpose Nerlove used a Cobb-Douglas production function, specified as:

$$Q = \alpha_0 K^{\alpha_K} L^{\alpha_L} F^{\alpha_F} e^{\varepsilon_i} \qquad \ldots\ldots\ldots\ldots\ldots \qquad (3)$$

where *Q* is output and the inputs are capital (*K*), labour (*L*) and fuel (*F*) and ε_i is an error term to capture unmeasured differences across firms. In this formulation, economies of scale would be indicated by the sum of the coefficients on *K, L* and *F* being greater than 1.

Because rates were set by state commissions and firms were required to meet the demand forthcoming at the regulated rates, Nerlove argued that output (as well as factor prices) could be viewed as exogenous to the firm. Hence the firm's objective could be taken as cost minimization subject to the production function, which leads to the cost function:

$$\ln C = \beta_0 + \beta_q \ln Q + \beta_K \ln P_K + \beta_L \ln P_L + \beta_F \ln P_F + u_i \qquad \ldots\ldots\ldots\ldots \qquad (4)$$

This can be estimated subject to the restriction $\beta_K + \beta_L + \beta_F = 1$. Economies of scale will be indicated by $\beta_q = 1/(\alpha_K + \alpha_L + \alpha_F) < 1$.

[2] The exposition here follows Greene closely.

Nerlove's results were consistent with economies of scale in electricity generation but these appeared to diminish as the size of firm increased. An amended specification including a term in $(\ln Q)^2$ improved the fit, implying a U-shaped cost curve such that economies of scale would be exhausted somewhere in the middle of the range of outputs for Nerlove's sample of firms.

Nerlove's work was updated by Christensen & Greene (1976), using the same data but a translog functional form, and simultaneously estimating the factor demands and the cost function. Their results were broadly similar to Nerlove's. They also redid the study using a sample of 123 firms from 1970, again with similar results. In the latter sample, however, Greene reports (p.127), "it appeared that many firms had expanded rapidly enough to exhaust the available economies of scale."

From the perspective of the present research, while this important work laid the methodological foundations for most subsequent investigation of electricity supply costs, it is noteworthy that it left out consideration of the possible influence of distribution costs on the results. Nerlove was aware of the issue but said (p.169) " … the problem of transmission and its effects on returns to scale has not been incorporated in the analysis, which relates only to the *production* of electricity." However, in a prescient, and subsequently somewhat overlooked Appendix to his article, he worked out that " … because of transmission losses and the expenses of maintaining and operating an extensive transmission network, a firm may operate a number of plants at outputs in the range of increasing returns to scale and yet be in the region of decreasing returns when considered as a unit."

b. Roberts (1986)

Roberts (1986) follows the practice pioneered by Christensen & Greene of specifying a cost function in flexible (translog) form, together with cost share equations, thereby avoiding importing unnecessary restrictions via the assumption of a specific production function.

Roberts' starting point is a transformation function for electricity production and delivery represented by:

$$T(K_G, M_G, E_P, K_D, M_D, Q) = 0 \qquad \cdots\cdots\cdots \qquad (5)$$

where Q is electricity supplied, K_G and K_D are generating capital and distribution capital respectively, E_P is purchased electricity, M_G and M_D are generating materials and distribution materials respectively[3].

He then argues (p.379) that "empirical analysis of this production process can be simplified, without greatly restricting the aspects of interest, by assuming that production occurs in two stages. First, the generation inputs and purchased power are used to produce the quantity of KwHs which the firm will supply. Second, these KwHs are then combined with transmission and distribution inputs to produce deliveries …" i.e. the transformation function can be written as:

$$T\{E_I(K_G, M_G, E_P), K_D, M_D, Q\} = 0 \qquad \cdots\cdots\cdots \qquad (6)$$

Roberts continues (p.379-80) "… the firm can now be viewed as making its input decisions in two stages. First, it chooses quantities K_G, M_G, and E_P to minimize the cost of producing the KwH input, E_I. This gives rise to a cost function for the KwH input …"

$$C_I(P_{KG}, P_{MG}, P_{EP}, E_I) \qquad \cdots\cdots\cdots. \qquad (7)$$

Then in the second stage, the firm chooses E_I and the other inputs to minimize the cost of producing deliveries. And (p.380) "Because these deliveries are geographically dispersed, the characteristics of the firm's service area, particularly its size in square miles (A) and number of customers (N), can affect the cost-minimising choice of … inputs. Since the firm is required to serve all customers within its specified

[3] To simplify the exposition, some arguments (e.g. fuel purchases) included in Roberts' specification have been omitted here and output is not sub-divided into bulk and retail sales.

service area, these two characteristics act as exogenous constraints." The firm's total cost of supplying electricity can then be represented by:

$$C(P_I, P_{KD}, P_{MD}, Q, A, N) \qquad \ldots\ldots\ldots.. \qquad (8)$$

Among the various advantages Roberts reasonably claims for this cost model are that it enables three distinct measures of economies of scale to be identified, viz:

1. $R_Q = \frac{1}{\varepsilon_Q}$, where $\varepsilon_Q = \frac{\partial \ln C}{\partial \ln Q}$, applicable when there is an increased demand for power from a fixed number of customers in a fixed service area, called "*economies of output density*" by Roberts;
2. $R_{CD} = \frac{1}{\varepsilon_Q + \varepsilon_N}$, where $\varepsilon_N = \frac{\partial \ln C}{\partial \ln N}$, applicable when more power is delivered to a fixed service area as it becomes more densely populated, while output per customer remains fixed, called "*economies of customer density*";
3. $R_S = \frac{1}{\varepsilon_Q + \varepsilon_N + \varepsilon_A}$, where $\varepsilon_A = \frac{\partial \ln C}{\partial \ln A}$, applicable when the size of the service area increases while holding customer density and output per customer constant, called "*economies of size*".

Roberts' work does indeed throw interesting new light on the economics of the distribution stage of electricity supply but, as will be argued below, the first stage cost-minimisation assumption behind (7) is open to question.

c. Thompson (1997)

The same issue emerges more strongly in the later study by Thompson (1997) of cost efficiency in the electric utility industry. Thompson's work seems to have been motivated by concern whether a regulator-driven trend towards separating vertically integrated electric utilities into a power generation unit and one or more regulated power delivery (transmission and distribution) units was economically justified. The paper explicitly presents itself as a development of Roberts' work.

Thus Thompson proceeds directly to postulate a total power procurement and delivery cost model of the form:

$$TC_D(w_E, w_{LT}, w_{LD}, w_{KT}, w_{KD}, Y_H, Y_L, S, N, t) \qquad \ldots\ldots\ldots\ldots \qquad (9)$$

Thompson comments (p.288) that this specification "contains the implicit assumption that the generation function of the vertically integrated firm is characterized by a linearly homogeneous production process. This implies constant unit costs for generated power …" and he cites "recent evidence" that "average long-run power supply costs may be constant for power supplied by the majority of electric utility firms".

Thompson goes on to note that hypotheses concerning the ability to separately analyze the vertically integrated electric utility as independent power supply, transmission and distribution service providers can be tested using this cost model by comparing it with one incorporating separability, such as:

$$TC_D\{C_S(w_E), C_D(w_{LD}, w_{KD}, Y_L, S, N), C_T(w_{LT}, w_{KT}, Y_L, Y_H)\} \ldots\ldots\ldots \qquad (10)$$

Here the cost of power supply (C_S) is dependent only on the market price of power – this follows from Thompson's assumption of constant unit costs for generated power; distribution costs (C_D) are assumed to be a function of distribution labour and capital prices, low voltage service volumes, the number of

customers and service territory characteristics; and the cost of transmission service (C_T) is a function of its own capital and labour input prices and both low and high voltage service volume.

Thompson adopts a translog form of the cost function to estimate his models using a sample of all major investor-owned electric utilities in the US for the years 1977, 1982, 1987 and 1992. This gave a sample of 83 firms for 1977 and 1982, and 85 firms for 1987 and 1992.

Among the findings, Thompson reports (p.293): "The *economies of output density*[4] are substantial, and rise considerably over the study period. On average, a 1 percent proportional increase in power sales … all else the same, increases total costs by 0.70 per cent. This results in the average cost of this activity decreasing by 0.30 per cent." One puzzle here is that as this finding is presumably (in a cross section analysis) mainly a reflection of economies of scale in power generation, it is not clear how this can be reconciled with the assumption, noted above, of constant unit costs in power generation. He also reports (p.293) that "*economies of customer density*[4], measuring the impact on costs of a proportional increase in sales volume and the number of customers … are small." Taken with the previous result, this implies diseconomies of customer numbers. The further effect of size of service area is found by Thompson to be very small but as with customer numbers, it implies a further diseconomy, leading overall to *returns to size* (R_S)[4] not significantly different from 1.

On the question of separability, Thompson calculates log likelihood values for the unrestricted model and for two restricted versions. He observes (p.294):

> "It can easily be seen that the hypothesis of separability of either the distribution system or power supply from the remaining utility services is strongly rejected in each of the time periods. This finding supports the comprehensive approach to electric utility cost analysis. It would appear that an inter-stage production technology and the beneficial use of common inputs is illustrative of the vertically integrated electric utility. These findings imply that the sum of the costs of the divested production stages would exceed the total cost of vertically integrated firm service."

However, Thompson's specification does not enable him to test whether separability might also be rejected because economies of scale in electricity production get traded off against diseconomies in distribution.

[4] As defined by Roberts – see p.4 above.

Is the assumption of cost minimization at the first (production) stage acceptable?

It is assumed by Roberts (and Thompson) that electricity production is separable (in the formal economic sense)[5] from electricity distribution. This is what enables them to assume that the costs of electricity generation (the production stage) are minimized prior to being input into the distribution stage – and hence to represent the input electricity in the cost function by a single price[6]. However, if there are scale economies in the production stage but diseconomies of scale in distribution, this assumption is inappropriate. Transfering attention from electricity to water supply, the point can be simply illustrated by reference to the diagrams in **Figure 1** below:

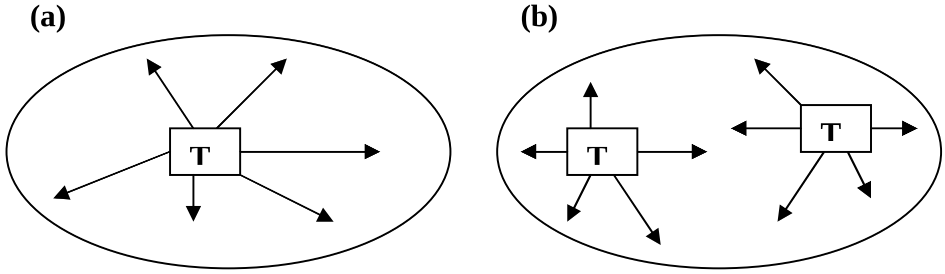

Figure 1: Water supply: Should this area be served by (a) one treatment works or (b) two (or more) treatment works?

In diagram (a), water is distributed over the whole service area from a single treatment works: This is the solution that would be chosen if economies of scale in production were the only consideration, and is the solution implied if separability is assumed. However, if there are sufficiently large diseconomies of scale in distribution, the combined costs of production and distribution may be minimized by opting for two (or more) treatment works, as in diagram (b), because the higher costs of production in smaller works may be more than offset by savings in distribution costs – particularly if, for example, the works are located near urban settlements and the rest of the service area is only sparsely populated. Of course, whether this is the case or not is an empirical matter but my own preliminary findings – see Wenban-Smith (2006) - suggest that the issue may be of more than academic interest, for water supply in the UK at least, and therefore that an important element of the situation may be missed if one proceeds to try to estimate scale economies in water supply with a cost function specification incorporating Roberts' assumption of separability.

Possible ways out

a. Estimate an aggregate production/cost function

Evidently, some care is needed in developing a production or cost function specification for estimating scale economies in water supply.

One possibility is to abandon the attempt to analyse the stages of production separately. This is the approach taken by Torres & Morrison Paul (2006)[7]. They propose a short run transformation (production) function of the general form:

[5] See Chambers (1988) pp.41-48 on separability in production functions and pp.110-119 on separability in cost functions.
[6] A similar assumption is made by Duncombe & Yinger (1993) in their two stage specification of a cost function for fire protection.
[7] The authors have informed me that they had originally hoped to carry out the separate analysis but it proved too difficult, mainly it seems because of collinearities in the data.

$t(Y, V, \overline{X}, Z)$ …………….. (11)

where Y is a vector of outputs (retail and wholesale water), V is a vector of variable inputs (e.g. labor, electricity and purchased water, whether used in production or distribution), $\overline{X}$ is a vector of quasi-fixed inputs (e.g. storage and treatment capacity) and Z is a vector of technical/environmental characteristics. This leads to the short run cost function:

$VC(Y, P, \overline{X}, Z)$ ……………… (12)

where P is a vector of the variable input prices. The authors comment (p.106): "In essence, this cost function describes the input use of water utilities producing at the frontier of the production possibility set, given short run capital (quasi-fixed) input constraints and assuming that firms choose the cheapest combination of variable inputs to produce the observed Y". From this short run cost function, the vector of cost-minimising variable input levels is captured by the vector of derivatives of the cost function with respect to the input prices.

One innovation in this study is to make output endogenous (whereas most studies in this field take output to be endogenous, i.e. water companies are obliged to supply whatever is demanded [8]) by adding to the system of equations, the identity:

$Y_f = GS + X_{Pw} - Y_w - loss$ …………….. (13)

where Y_f and Y_w are retail and wholesale water respectively, GS is groundwater plus surface water extracted, and X_{Pw} is purchased water.

Of particular interest here is Torres & Morrison Paul's treatment of output density, which is similar to Roberts' described earlier. They remark (p.108) that " … output density … depends on three main variables: output, number of customers and service area size. A standard measure of scale economies … actually measures volume … economies … - the cost impact of an increase in output given the existing network. A full measure of economies of scale or size requires recognising that increasing 'scale' involves also expansion of the network, and thus depends on a balance of cost associated with water volume, connections and distance." The implications become clearer when the various measures of scale economies are defined.

Economies of volume scale are defined as :

$$\varepsilon_{CY} = \frac{\partial VC(Y,...)}{\partial Y_w}\frac{Y_w}{VC(Y,...)} + \frac{\partial VC(Y,...)}{\partial Y_f}\frac{Y_f}{VC(Y,...)} \quad \ldots\ldots\ldots \quad (14)$$

This is the inverse of Roberts' R_Q. The double term is necessitated by the decision to treat retail and wholesale water as multiple products. Related to this is a definition of economies of scope.

Economies of vertical network expansion measure the combined effect of higher volume and more customers, with the demand per customer and the size of the service area held constant, and are defined as:

$$\varepsilon_{CYN} = \varepsilon_{CY} + \varepsilon_{CN}$$

where

$$\varepsilon_{CN} = \frac{\partial VC(Y, P, \overline{X}, Z)}{\partial N}\frac{N}{VC(Y,...)} \quad \ldots\ldots\ldots \quad (15)$$

Here N is number of customer connections, which is a component of Z. This is the inverse of Roberts' R_{CD}.

[8]In contrast, Saal & Parker (2005, pp.5-6) for example state: "Considering that water companies have a statutory obligation to meet demand for water and sewerage services, it is appropriate to assume that outputs are exogenous and inputs are endogenous rather than the other way round." It seems that Torres & Morrison Paul's alternative approach did not have a big impact on the results but it did correct some regularity conditions (CJ Morrison Paul, personal communication)

Economies of horizontal network expansion (or *spatial density*) then measure the combined effect of higher volume and larger service area, with numbers of customers held constant, and are defined as:

$$\varepsilon_{CYS} = \varepsilon_{CY} + \varepsilon_{CS}$$

where

$$\varepsilon_{CS} = \frac{\partial VC(Y,P,\overline{X},Z)}{\partial Sa}\frac{Sa}{VC(Y,...)} \qquad \ldots\ldots\ldots.. \qquad (16)$$

and here *Sa* is service area size, also a component of *Z*. This is not a measure used by Roberts.

Finally, *economies of size* (p.111) "prevail if a combined measure of volume, customer density, and spatial density economies, constructed by adding the cost effects from marginal increases in both customer numbers and service area size to economies of volume … falls short of one." That is, if

$$\varepsilon_{Size} = \varepsilon_{CY} + \varepsilon_{CN} + \varepsilon_{CS} < 1$$

This is the inverse of Roberts' R_S.

Although water treatment and water distribution have not been analysed separately in Torres & Morrison Paul's model, volume economies can be seen as likely to arise mainly at the treatment stage while economies (or diseconomies) linked to customer numbers or service area relate primarily to the distribution stage. Their approach can thus be seen as going some way towards isolating the different economics of production from those of distribution. This is an important step forward if there are indeed "potentially significant cost trade-offs involving water production and network size". However, because their specification does not distinguish between inputs to the production stage and inputs to the distribution stage, there must remain some uncertainty about the size of these effects.

Torres & Morrison Paul's article is thus a useful contribution to the literature, bringing out more clearly than before the effect of demand density on costs. There is however another possible qualification as regards their measurement of the effect of size of service area. Torres & Morrison Paul considered including length of pipes in the vector of quasi-fixed inputs but decided against when they found that pipeline length was strongly correlated with service area size. Therefore, as only variable costs are modeled, it is not clear how the extra (capital) costs of the longer pipes required by larger service areas can be reflected in ε_{CS}, which may therefore be underestimated. On this, Torres & Morrison Paul comment (p.8, Footnote 13) " … if [pipeline length is] included as a level the estimates are not robust due to multi-collinearity. If included as a ratio (pipeline length per customer), network size is in some sense controlled for, causing the ε_{CN} estimates to have a downward, and the ε_{CS} estimates an upward trend over the size of firms." The question here is whether the short run specification of the production/cost function used can adequately represent differences in the capital invested in systems of different sizes and densities.

b. Estimate separate production (or cost) functions for each stage

On the face of it, the problems identified above might be avoided by estimating a production or cost function for each stage of water supply separately. However, this route is not without its own problems, as explained below.

i. Production stage

The standard procedure would be to start by postulating a production function of the general form:

$$Q_T = f(K_T, L_T, N_T) \qquad \ldots\ldots\ldots.. \qquad (17)$$

where Q_T is quantity of water produced for input into the distribution system, K_T is (a vector of) capital employed in water treatment, L_T is (a vector of) treatment operating costs and N_T is a vector of

environmental factors (such as type of water) likely to affect treatment costs. From this production function, assuming cost minimization, a cost function can be derived of the general form:

$$C_T = C(Q_T, p_{KT}, p_{LT}, N_T) \quad \ldots\ldots\ldots\ldots.. \quad (18)$$

or, if capital is taken to be quasi-fixed (as in Torres & Morrison Paul):

$$C_T = C(Q_T, \overline{K_T}, p_{LT}, N_T). \quad \ldots\ldots\ldots.. \quad (19)$$

A possible problem here is that the specification strictly relates to individual plants so can only be implemented if plant level data is available. In the US, although many water utilities appear to operate at rather small scale with only one treatment works, there is little data available on capital inputs. In the UK, on the other hand, more data is available at company level but most water companies are rather large, and operate large numbers of plants, with very limited plant level data publicly available (not including plant level costs).

ii. Distribution stage

It may be rather stretching the concept of a production function to apply it to distribution but supposing the concept is accepted, it could be postulated to have the general form:

$$Q_D = f(Q_T, K_D, L_D, N_D) \quad \ldots\ldots\ldots\ldots\ldots.. \quad (20)$$

where K_D , etc are the distribution equivalents of the treatment variables. Even before proceeding to the derivation of a cost function, there are some problems to address:

- First, just as the treatment production function needs to be related to an appropriate unit of production, the relevant unit for distribution needs to be defined. Typically the distribution system for each community (village, town or city) is more or less self contained so that each such self contained distribution system is probably the appropriate unit for analysis. In the US, this is often compatible with the production unit, facilitating data collection and analysis. In the UK, however, each company serves a large number of communities and information on the geography and costs of each system is not easy to assemble.
- Secondly, there is a question about how distribution output (Q_D) should be measured. Volume of water is arguably inadequate as it does not reflect the transport of water from works to customer, which is the essence of what the distribution system is "producing". A composite measure, such as Ml x km, would seem to be superior but it may be difficult to construct such a measure.
- Thirdly, there is likely to be strong collinearity between Q_T and Q_D as the difference between them (in volume terms) is leakage which, in UK at least, is fairly uniform at about 20% for all companies, except one. Use of a composite measure of distribution output might reduce this problem.
- The treatment of leakage is another issue. Should it be treated as a cost, or, as some authors have done, as a separate output?

Subject to resolution of these problems, it would in principle then be possible to proceed to derive a distribution cost function of the general form:

$$C_D = C(Q_D, p_{QT}, p_{KD} or \overline{K_D}, p_{LD}, N_D) \quad \ldots\ldots\ldots\ldots. \quad (21)$$

There remains the problem of how to price the treated water input into the distribution process. If, as is commonly supposed, there are economies of scale in water treatment, the price should vary according to the volume input.

c. Estimate a composite production function

As an alternative to two separate production functions, one might think of constructing a production function which includes all the separate inputs to production and distribution in a single function, such as:

$$Q_D = f(K_T, L_T, N_T, K_D, L_D, N_D) \qquad \ldots\ldots\ldots\ldots \qquad (22)$$

As with the separate production functions discussed above, implementation will require suitably disaggregated data, with the added complication that production units may not align with distribution areas. There is also greater likelihood of unacceptably high collinearity between variables. Such problems are even more probable if a cost function were to be derived from this composite production function. In particular, the price of capital for production is likely to be identical to the price of capital for distribution, as is the price of labour for each stage, rendering their separate effects unidentifiable.

Conclusions

In the light of this examination of the issues, one may conclude that for the purpose of investigating economies of scale in urban water supply (or other infrastructure services):

1. A possible starting point would be a composite production function like (22) above, provided appropriate data is available, and there is not excessive collinearity between variables. However, it is very likely that estimating a cost function based on this production function would not be feasible because of collinearity in the prices.
2. A better prospect would be to start from separate production functions like (17) and (20) and/or their related cost functions (18) or (19) and (21). There would still be a number of practical issues to resolve, as noted above.
3. If neither of the above approaches can be successfully implemented, the aggregate production function used by Torres & Morrison Paul remains as a possible approach although it may not fully expose the different economics of production and distribution.
4. Failing any of the above, a simpler approach based on unit costs may still be feasible although this will not reveal potentially important economic effects (such as factor substitution).
5. Whatever approach is adopted, regard should be had to the different dimensions of economies of scale in distribution as developed by Roberts, and extended by Torres & Morrison Paul. An approach which focused just on volume economies would miss important aspects of the situation.

References

Chambers R G (1988) *Applied Production Analysis* CUP

Christensen L and W H Greene (1976) "Economies of scale in US electric power generation" *Journal of Political Economy*, 84:655-76

Duncombe W and J Yinger (1993) "An analysis of returns to scale in public production, with an application to fire protection" *Journal of Public Economics*, 52:49-72

Fujita M and J-P Thisse (2002) *The Economics of Agglomeration* CUP

Greene W H (2003) *Econometric Analysis (5th Edition)* Prentice Hall (Pearson Education International)

Nerlove M (1963) "Returns to scale in electricity supply" in C Christ (Ed) *Measurement in Economics*, Stanford University Press

Roberts M J (1986) "Economies of density and size in the production and delivery of electric power" *Land Economics* 62:378-87

Saal DS and D Parker (2005) "Assessing the performance of water operations in the English and Welsh water industry: A panel input distance function approach" *Aston Business School Research Paper RP0502*.

Schmalensee R (1978) "A note on economies of scale and natural monopoly in the distribution of public utility services" *Bell Journal of Economics*, 9:270-6

Thompson H G (1997) "Cost efficiency in power procurement and delivery service in the electric utility industry" *Land Economics*, 73:287-96

Torres M and C J Morrison Paul (2006) "Driving forces for consolidation or fragmentation of the US water utility industry: A cost function approach with endogenous output" *Journal of Urban Economics* 59:104-120

Wenban-Smith H B (2006) *Urban infrastructure: Density matters, not just size* (London School of Economics Department of Geography & Environment Working Paper No. 104)