ECONOMIES OF SCALE, DISTRIBUTION COSTS AND DENSITY EFFECTS IN URBAN WATER SUPPLY

A spatial analysis of the role of infrastructure in urban agglomeration

Thesis submitted for the degree of Doctor of Philosophy
(Ph.D)

by

Hugh Boyd WENBAN-SMITH

LONDON SCHOOL OF ECONOMICS
AND POLITICAL SCIENCE

Declaration

This thesis is my own work, apart from referenced quotations and assistance specifically acknowledged.

(Signed): ……………………………………………………….

(H B WENBAN-SMITH)

(Date): …………………………………..

ECONOMIES OF SCALE, DISTRIBUTION COSTS AND DENSITY EFFECTS IN URBAN WATER SUPPLY: A spatial analysis of the role of infrastructure in urban agglomeration

Abstract

Economies of scale in infrastructure are a recognised factor in urban agglomeration. Less recognised is the effect of distribution or access costs. Infrastructure can be classified as: (a) Area-type (e.g. utilities); or (b) Point-type (e.g. hospitals). The former involves distribution costs, the latter access costs. Taking water supply as an example of Area-type infrastructure, the interaction between production costs and distribution costs at settlement level is investigated using data from England & Wales and the USA.

Plant level economies of scale in water production are confirmed, and quantified.

Water distribution costs are analysed using a new measure of water distribution output (which combines volume and distance), and modelling distribution areas as monocentric settlements. Unit distribution costs are shown to be characterised by scale economies with respect to volume but diseconomies with respect to average distance to properties. It follows that higher settlement densities reduce unit distribution costs, while lower densities raise them.

The interaction with production costs then means that (a) higher urban density ("*Densification*") is characterised by economies of scale in both production and distribution; (b) more spread out settlement ("*Dispersion*") leads to diseconomies in distribution; (c) "*Suburbanisation*" (expansion into lower density peripheral areas) lies in between, with roughly constant returns to scale, taking production and distribution together; and (d) *"Constant density"* expansion leads to small economies of scale. Keeping (per capita) water supply costs low thus appears to depend as much on density as size.

Tentative generalisation suggests similar effects with other Area-type infrastructure (sewerage, electricity supply, telecommunications); and with Point-type infrastructure (such as hospitals), viewing access costs as distribution costs in reverse. It follows that the presumption in urban economics that such services are always characterised by economies of scale and therefore conducive to agglomeration may not be correct.

ECONOMIES OF SCALE, DISTRIBUTION COSTS AND DENSITY EFFECTS IN URBAN WATER SUPPLY: A spatial analysis of the role of infrastructure in urban agglomeration

Contents

Appendices

List of Tables and Figures

Chapter I

Chapter II

Chapter III

Chapter IV

Chapter V

Chapter VI

I. OVERVIEW: MOTIVATION, METHODOLOGY AND KEY FINDINGS

1. Motivation

Infrastructure is the Cinderella of urban economics. The accumulated investment in urban infrastructure is absolutely massive[1]; yet it is almost invisible in the literature. While the part played in urban agglomeration by thick labour markets, economies of scale in manufacturing, specialisation, technological spill-overs and consumption externalities have all recently attracted considerable attention, infrastructure has rather been taken for granted, providing a backdrop to the urban drama but not, seemingly, playing an active part.

Insofar as infrastructure has attracted attention, the predominant proposition is that it is characterised by economies of scale. Thus McDonald (1997), discussing urbanisation economies in his standard text remarks (pp.40-41): "Economies of scale exist in the provision of inputs that are not specific to a particular industry. An important example is the general urban infrastructure." Similarly, Fujita (1989, p.135) observes that "… the provision of many *public services and facilities* (such as schools, hospitals, utilities, and highways) typically exhibits the characteristic of economies of scale." If this is the case, one would expect infrastructure to make a large positive contribution to urban agglomeration economies. However, the evidence for such an effect is not strong. Although some studies of urbanisation economies have found a positive effect, others have not (Eberts & McMillen (1999, pp.1460-1491) provide a review of the evidence) and there is a tendency in the theoretical literature to downplay the role of scale economies in agglomeration (Duranton & Puga (2004)).

The aim of this thesis then is to take a close look at the micro-economics of one example of urban infrastructure – water supply – with the aim of arriving at a better understanding of its contribution to agglomeration economies, and in the hope that this will throw some light on the role of infrastructure more generally. The core of the argument is that it is insufficient to focus just on economies of scale. Urban areas have a spatial aspect so that it is necessary also to take into account the costs of accessing facilities or distributing services, sometimes over considerable distances. As the analysis

[1] No plausible estimate of value to substantiate this assertion could be found – a further indication perhaps of the relative neglect of this topic.

of commuting costs by Arnott (1979) suggests[2], these activities may well be characterised by scale diseconomies. If then there is a trade-off between economies of scale in production and diseconomies in distribution (or access), this will weaken the influence of scale economies on agglomeration, perhaps helping to explain the inconclusive evidence on this point. The results obtained lend some support to this line of argument but also, and perhaps more importantly, they draw attention to the role of density, with high densities reinforcing scale economies but low densities adding to costs. This suggests that the contribution of infrastructure to agglomeration, whether large or small, may be due as much to density effects as to scale effects.

2. Research strategy

Infrastructure, widely defined[3], can be viewed as belonging to one of two categories: (a) Area-type, where the product of a facility needs to be distributed to consumers over a defined area (e.g. utilities); or (b) Point-type, where the services of a facility can only be consumed by users in its catchment area making their way to it (e.g. hospitals). The former involves, in addition to production costs, distribution costs; the latter, access costs. Whereas production can generally be expected to exhibit economies of scale, this is not necessarily the case for distribution (or access). There is a spatial aspect to distribution (and access): more output means either larger service/catchment areas (with greater distances and so, potentially, higher costs) or higher densities (with savings, perhaps, from greater proximity but also the risk of higher congestion costs). There may in consequence be a trade-off between economies of scale in production and diseconomies in distribution/access.

To investigate this question, urban water supply is taken as a case study. It is an example of Area-type infrastructure; the production technology is characterised by economies of scale and is not very complicated; and water distribution costs are high

[2] He shows average commuting cost to be an increasing function of city size by considering a circular city of uniform population density, where all commuting is to a central business district and transport cost is proportional to distance. Total commuting costs are then given by:

$$TCC = \int_0^R r.2\pi r.dr = \frac{2}{3}\pi.R^3 = \frac{2}{3\sqrt{\pi}}N^{3/2}$$

Where R is the radius of the city and N is its population, i.e. aggregate commuting costs increase more than proportionately with population, and average commuting cost is an increasing function of N.

[3] There has been a recent tendency for the term "infrastructure" to be reserved for transport infrastructure. The wider definition adopted here is discussed in **Chapter II, section 1**. The application of the analysis to transport is discussed in **Chapter VII**.

relative to production costs. The effects of interest should therefore be particularly evident in this case. The research question can then be summarised as:

<u>The Research Question</u>
Viewing water supply as a type of urban infrastructure, what is the interaction between economies of scale, distribution costs and density effects at settlement level?

The question is of interest in its own right as there is controversy about whether there really are economies of scale in this case[4]. But what is learnt about water supply should also shed light on the role of other types of infrastructure in urban agglomeration.

The availability of suitable data is often a critical factor in research which aims to quantify effects of this kind. In this case, use has been made of data from three primary sources:

a. **Ofwat (2003a): The annual June Returns made to Ofwat by water companies in England & Wales.** While there is a great deal of information in these returns, a difficulty is that the water companies are rather large, each serving numerous settlements. However, with some ingenuity, it has proved possible to use this data to infer some settlement level effects (particularly in water production);

b. **AWWA (1996): A 1996 survey of its members by the American Water Works Association (AWWA).** This is more suitable for our purposes in that most US water utilities are quite small, often serving a single community. However, the information is less full than the Ofwat data (in particular, there is a lack of information on capital costs);

c. **Not referenced: A large amount of highly disaggregated internal information, provided for the purposes of this research by one of the larger companies reporting to Ofwat.** This information has been particularly helpful in elucidating scale effects at local level in water distribution but has also provided corroboration for findings on water production from other sources. As the company does not wish to be publicly identified, it has been given the pseudonym Britannia Water Company (BWC) in this report.

[4] An Ofwat Press Notice in 2004 was headlined "There is no evidence of general economies of scale in the water industry". (Ofwat PN 01/04, 14 January 2004)

3. Methodology

The aim of the empirical work reported in **Chapters IV**, **V** and **VI** is to use data on water supply to throw light on the interaction between economies of scale, distribution costs and density effects at settlement level. The methodologies used, which are discussed in **Chapter III**, build on the approaches found in the utilities literature surveyed in **Appendix B**, which are mainly those of industrial economics. The topic however straddles a number of branches of economics. Initially interest was sparked by urban economics (e.g. Fujita (1989)). Other relevant fields include transport economics, the economics of public goods, the theory of public facility location (e.g. Love *et al* (1988)) and the urban planning literature on “sprawl”. More recently, the emergence of “The New Economic Geography” (e.g. Fujita *et al* (1999)) has given new life to the study of spatial economics, particularly the interaction between economies of scale and transport costs. Some of the literature in all these fields has been consulted in carrying out this research.

Generally, the utilities literature points to the use of cost functions as the way into assessing scale effects and this has been taken as a starting point. The objectives of this research however are different from those in the mainstream utilities literature in that the focus is on settlement (rather than company) level effects; and water supply is viewed as an example of urban infrastructure, rather than as a branch of manufacturing industry. Moreover, the aim is only to arrive at a reasonable characterisation of the main effects, rather than a precise estimate for any particular town or company.

The device of treating capital in the water industry as “quasi-fixed”, pioneered by Garcia & Thomas (2001), and since widely adopted, has been followed. This leads to the use of a “short term” cost function, in which operating costs are the independent variable while the fixed capital becomes in effect a control variable. In fact, in the case of water distribution, it has been taken a little further, with a Leontief-type production function for this activity postulated, when fixed capital can be dropped from the relationship.

Nerlove (1963) in his pioneering application of cost functions looked just at the production stage of electricity supply but subsequent work has often included the distribution stage as well. The problems introduced by this extension have attracted little comment. In the literature surveyed, water production and distribution have not been analysed separately (and this seems to be the case for electricity supply also). In **Chapter III** reasons why the methods currently in use may not fully expose distribution effects are put forward (these include non-separability, multi-collinearity and inadequate representation of the spatial aspect of distribution). In consequence, any trade-off between economies of scale in production and diseconomies in distribution will remain obscure. Yet it is these effects that matter in the urban context. That is why it has been seen as necessary to look separately at the distribution side, which is where the spatial aspect comes into play.

In developing the analysis of water distribution, an important innovation is the introduction of a new measure of distribution output. This aims to reflect the distance over which water has been piped as well as its quantity, rather as tonne-kms or passenger-miles are used in transport studies. This measure is derived by modelling distribution areas as monocentric settlements, so as to approximate the average distance to properties (called φ) which is then multiplied by total consumption. The same output measure is appropriate when production and distribution are combined; and it would appear to be worthy of consideration, *mutatis mutandis*, in studies of other utilities when distribution as well as production costs are under consideration.

As other authors have noted (e.g. Stone & Webster Consultants (2004)), the inclusion of distribution means that there is more than one dimension of scale to consider. We draw particular attention to:

a. ε_W – the elasticity of distribution (or total) costs with respect to consumption per property;

b. ε_N – the elasticity of distribution (or total) costs with respect to numbers of properties;

c. ε_A – the elasticity of distribution (or total) costs with respect to area served.

A brief comment is appropriate here on why panel data have not been used (the point is more fully covered in **Chapter IV, section 1 (e)**). Although several years of Ofwat data are available, the companies reporting to Ofwat are mostly too large for present purposes and examination of 6 years data for the smaller water companies (the WOCs) suggested that the year to year variation in key variables might be so small as to render the results unreliable. The AWWA and BWC data is only available for a single year. In any case, the fixed effects that are removed by panel methods (e.g. size of service area) are precisely what is of interest in this research. Therefore, the analysis has relied on cross-section analysis (for 2002/3 in the case of Ofwat data).

4. Key findings

Against this background, the ways in which this research has advanced knowledge can be summarised.

a. What is new in this thesis

- Previous work in urban economics has tended to assume economies of scale in urban infrastructure without considering the effect of distribution or access costs;
- Water supply has not previously been used as a model for urban infrastructure, despite the advantage of simple technology and reasonably accessible data;
- Ofwat data for England & Wales does not seem to have been previously used to examine the economics of water supply at settlement level, company level studies being the norm; but settlement level analysis helps understanding of what is intrinsically a spatial industry;
- There has been little previous recognition of the need to investigate water production and water distribution separately if their different characteristics are to be fully exposed;
- New estimates of plant level economies of scale in water production have been produced;
- An innovative non-linear specification has been developed to estimate scale economies for WTWs and boreholes separately by exploiting information on plant numbers and sizes in the Ofwat June Returns;
- A new measure of water distribution output has been developed which recognises the spatial dimension of water supply;
- The monocentric urban model has been adapted to capture in a compact way the different spatial characteristics of distribution areas;
- It has been found that there are scale economies in water distribution with respect to volume but diseconomies with respect to average distance to properties (and these effects have been quantified);
- It has been demonstrated that the various cost elasticities to be derived from these results are conditional on the scenario under consideration (the scenarios

are (a) *densification*, (b) *dispersion*, (c) *suburbanisation*, and (d) *constant density*);

- It is suggested that the ideas and methods developed in this thesis may be applicable, with due care, to a range of other types of infrastructure;
- An implication of the results is that density as well as size needs to be taken into account in studies of urbanisation economies: measuring city size by population alone risks missing density effects (unless density happens to be correlated with size)[5].

b. Water production

It is conventional wisdom that there are economies of scale in water production. The evidence in **Chapter IV** confirms that this is indeed the case for water treatment works (WTWs), even when water acquisition is included. Here, by exploiting the Ofwat data on numbers and sizes of works for each company, it proved possible to develop a method to estimate plant level economies of scale for boreholes and WTWs simultaneously, despite the absence of separate cost information on these two types of supply. The results obtained in **Chapter IV** for this and other cases are summarized in **Table 1.1**. Here the estimates are expressed as returns to scale so that a value greater than 1 indicates economies of scale. It may be seen that although the estimates vary, there is a consistent finding of economies of scale for WTWs; for boreholes however the values are not significantly greater than one.

Data source	**No of cases**	**Speci-fication**	**WTWs**		**Boreholes**	
			Operating costs	**Total costs**	**Operating costs**	**Total costs**
Ofwat companies (see Table 4.9)						
All Cos	21	(4.20)	1.56**	1.28*	1.04	1.27
(S.E.)			*(0.16)*	*(0.14)*	*(0.16)*	*(0.25)*
AWWA (see Table 4.7)						
TreatSW	145	(4.12)	1.25***	n.a.	-	-
(S.E.)			*(0.05)*			
TreatSWN	115	(4.12)	1.37***	n.a.	-	-
(S.E.)			*(0.06)*			
TreatGW	161	(4.12)	-	-	1.10*	n.a.
(S.E.)					*(0.05)*	
BWC (see Table 4.4)						
WTWs	15	(4.10)	1.28***	n.a.	n.a.	n.a.
(S.E.)			*(0.06)*			

Table 1.1: Estimated plant level returns to scale in water production (Significance levels, relative to 1: * = 1%, ** = 5%, * = 10%)**

[5] And in the presence of congestion, the relevant concept may be "effective density".

It is important however to recognize that these are *plant level* findings. When two or more works are operated by a company (for example, because the size of works is limited by the capacity of the water sources; or because the communities it serves are small and/or widely separated), these scale economies will be less evident. The benefits of large scale production can therefore only be reaped where circumstances permit the operation of large WTWs, typically where there is a large population and access to high capacity water resources. Birmingham, for example, which has a population of over 1 million and access to water from the Elan Valley, is mostly supplied by a single large WTW (the Frankley works) leading to relatively low water supply costs for that city.

In **Chapter VI**, these estimates of returns to scale for WTWs and boreholes are deployed in conjunction with estimates of scale effects in water distribution to explore the implications for various urban configurations. WTWs (for which economies of scale are quite large) allow a productive exploration of the trade-off between production economies and distribution diseconomies. This is not the case for borehole supplies, where the evidence for scale economies is weak (their costs apparently depending mainly on factors other than scale), serving as a reminder that even in the case of water supply, it cannot be taken for granted that there are always economies of scale in production.

c. Water distribution

Water distribution costs are at least as significant as water production costs. In the case of BWC for example, although distribution operating costs are about the same as production operating costs, distribution capital costs are about twice as large. Scale effects in distribution therefore merit careful attention.

It has already been noted that the concept of scale in water distribution has more than one dimension. As Schmalensee (1978, p.271) has remarked: “When services are delivered to customers located at many points, cost must in general depend on the entire distribution of demands over space.” The modeling and empirical estimation in **Chapter V** indicates that two aspects are particularly important: the volume of water distributed and some measure of the size of the service area. The volume of water consumed is the product of numbers of properties and usage per property but if usage per property does not vary much from place to place, estimation of the volume scale effect will be much the same whether volume or numbers of properties is used. For the service area measure, the more obvious possibilities include the actual area and length of mains. As the former will often include areas of unserviced land, the latter is preferable. However, better still would be a measure which can capture the spatial distribution of properties and this is what our measure φ aims to do. As explained earlier, it is derived by treating service areas as monocentric settlements of a size determined by the observed length of mains and property density for each area. This produces a measure of the average distance to properties, which can be applied flexibly to a wide range of actual situations. Although an approximation, it provides a versatile tool with which to represent the spatial aspect of distribution.

Armed with this tool, the reasonably clear results summarized in **Table 1.2** below are obtained. Generally, it was found that there are scale economies in distribution with respect to volume consumed (with a coefficient of about 0.4 – a value less than 1 indicating volume scale economies) but diseconomies with respect to average distance to properties (with a coefficient of about 1 – a value greater than 0 indicating distance diseconomies) – see equation (5.19) in **Chapter V**. The implications for distribution costs then depend on how these influences balance out. However, as explained in **Chapter V, section 2(c)**, the relevant elasticities cannot be directly inferred from these coefficients. Further analysis is required to separate the pure volume effect (due to variations in consumption per property) from spatial effects (due to variations in the number and location of properties).

A monocentric settlement can be approximately characterized by four parameters: d_0, its density at the centre; λ, the rate at which density declines away from the centre; N, its population; and R, its radius. To assess distribution cost elasticites, four monocentric settlement scenarios with different spatial characteristics are set up[6]. These are:

(a) ***Densification***[7]: Number of properties (N) varies, while settlement radius (R) is held constant (λ also therefore varying);

(b) ***Dispersion***: Coefficient of dispersion (λ) varies, holding number of properties (N) constant (R also therefore varying);

(c) ***Suburbanisation***: Number of properties (N) varies, holding λ constant (R also therefore varying);

[6] Central density (d_0) is taken to be 30 properties/Ha in all cases.

[7] It is recognised that this term has acquired particular policy connotations in the urban planning context; here it is simply adopted as a convenient descriptive label.

(d) ***Constant density***: Number of properties (*N*) varies, holding density (*N/A*) constant (when both λ and *R* vary).

Distribution cost elasticities are then evaluated for an average BWC urban district (*N* = 18,000), an average WOC (*N* = 200,000) and an average US retail only water utility (*N* = 50,000). The outcome of this further analysis is summarised in **Table 1.2**, which brings together results for the quantity elasticity ε_w (elasticity of cost with respect to consumption per property) with the spatial elasticities from **Table 5.6** in **Chapter V**. For ε_w, $\varepsilon_{N/\bar{R}}$, $\varepsilon_{N/\bar{\lambda}}$, and $\varepsilon_{N/\bar{D}}$ a value less than 1 indicates scale economies, a value greater than 1 scale diseconomies; for $\varepsilon_{A/\bar{N}}$ any value greater than 0 indicates diseconomies. A standard error is only available for ε_w; in the other cases, the range of values obtained by calculation across the sample is indicated. For a fuller discussion, see **Chapter V**, **section 6**.

	Average BWC urban district	**Average WOC**	**Average US retail utility**[a]
No. of properties	18,000	200,000	50,000
1. Quantity effect			
ε_w *(S.E.)*	0.43 *(0.23)*	-0.21 *(0.34)*	0.37 *(0.10)*
2. Spatial effects			
(a) Densification			
$\varepsilon_{N/\bar{R}}$ *(range)*	0.73 *(0.80 – 0.70)*	0.81 *(0.83 – 0.75)*	0.68 *(0.71 – 0.69)*
(b) Dispersion[b]			
$\varepsilon_{A/\bar{N}}$ *(range)*	0.18 *(0.21 – 0.07)*	0.19 *(0.22 – 0.07)*	0.17 *(0.20 – 0.06)*
(c) Suburbanisation			
$\varepsilon_{N/\bar{\lambda}}$ *(range)*	1.03 *(0.97 – 1.45)*	1.32 *(1.19 – 1.45)*	1.07 *(1.00 – 1.16)*
$\varepsilon_{A/\bar{\lambda}}$ *(range)*	0.63 *(0.70 – 0.17)*	0.51 *(0.73 – 0.37)*	0.58 *(0.69 – 0.47)*
(d) Constant density			
$\varepsilon_{N/\bar{D}} = \varepsilon_{A/\bar{D}}$ *(range)*	0.91 *(0.92 – 0.90)*	1.02 *(1.02 – 1.07)*	0.92 *(0.92 – 0.98)*

[a] In this case the volume variable was ln*QDI*.

[b] For this elasticity, a value > 0 implies diseconomies; for the others a value >1.

Table 1.2: Comparison of distribution cost elasticities across three data sets

It can be seen that the findings for BWC urban districts and US retail utilities are consistent: There are quite large scale economies with respect to consumption per property in water distribution (returns to scale = 1/ $\varepsilon_w \approx 1/0.4 = 2.5$). Among the spatial elasticities, *densification* and *constant density* expansion are also characterised by scale economies (returns to scale = $1/\varepsilon_{N/\bar{R}} \approx 1/0.7 = 1.4$ and $1/\varepsilon_{N/\bar{D}} \approx 1/0.9 = 1.1$ respectively). On the other hand there are diseconomies associated with *dispersion* and *suburbanisation*. The WOC results are in reasonable agreement as regards *densification* and *dispersion* but show higher diseconomies for *suburbanisation* and (small) diseconomies for *constant density* – possibly a reflection of the relatively large size of the WOCs so that there are a number of subsidiary settlements around the main centre. (The reason for the negative value for ε_w for the WOCs, albeit with a large standard error, has not been determined[8].)

d. Production and distribution combined

At the outset, it had been anticipated that while economies of scale in water production would be confirmed, diseconomies would be found in water distribution. It would follow that in urban water supply systems, a trade-off between these effects would be at work, qualifying the popular view that infrastructure services, such as water supply, are characterized only by economies of scale. In fact, a more complicated story has emerged. Generally, it has been found that there are volume economies of scale in water distribution as well as in water production but that density effects also need to be taken into account, with low density adding substantially to distribution costs. An important feature of the situation, conditioning these results, is that water suppliers generally have to take the size and location of the settlements they serve as given. They are not able to pursue cost savings by organizing the merger or relocation of small towns or awkwardly located customers; and it is doubtful whether even in the longer term, differential water supply costs have much effect on the evolution of settlement patterns.

The results for water production and water distribution can be brought together using the same settlement scenarios: i.e (a) *densification*; (b) *dispersion*; (c) *suburbanization*; and (d) *constant density*, assuming a single large WTW of the appropriate size and constant consumption per property (see **Table 6.4** in **Chapter VI**). Now, as numbers of properties are increased in each scenario (leading to higher volumes, given constant usage per property), the key difference is how density is affected.

- With (a) *densification*, because the urban boundary does not change as property numbers increase, density increases in parallel, so that volume economies

[8] One possibility is that it could be due to low distribution costs for large industrial supplies.

predominate in distribution as well as production. For example, unit water supply costs for a town doubled in size to 50,000 properties occupying 2,250 Ha (density 22.2 properties/Ha) will, according to these calculations, be 16.2% lower than for a town of 25,000 properties occupying the same area (density 11.1 properties/Ha), about half of the reduction coming from lower unit water production costs and half from lower unit distribution costs.

- With (b) *dispersion*, the number of properties does not increase, so that there is no volume effect, but the more dispersed pattern of settlement means lower density and an increasing average distance to properties, and hence higher distribution costs. For example, unit water supply costs for a town of 18,000 properties spread out over 2,090 Ha (density 8.6 properties/Ha) will be 10.8% higher than for a town of 18,000 properties occupying only 735 Ha (density 24.5 properties/Ha), all due to a 23.4% increase in unit distribution costs.
- With (c) *suburbanization*, the number of properties increases but because the increase is into less dense peripheral areas, average density falls and average distance to properties increases, albeit to a lesser extent than with (b). In this case, volume economies (in both production and distribution) are more or less balanced by average distance diseconomies. For example, unit supply costs for a town which has grown to 50,000 properties occupying over 20,000 Ha (density 2.4 properties/Ha) will be much the same as for the same town when it was only 15,000 properties occupying 985 Ha (density 15.2 properties/Ha) with the 25% reduction in unit production cost due to higher volume largely offset by a similar increase in unit distribution cost (the distance effect outweighing the volume effect in distribution here).
- With (d) *constant density*, the number of properties increases in line with the increase in area so that density is unchanged although the average distance to properties does increase. In this case, volume economies (in both production and distribution) outweigh the average distance effect. For example, unit supply costs for a town of 50,000 properties occupying 5,000 Ha (density 10 properties/Ha) will be 16.7% lower than for a town of 15,000 properties occupying 1,500 Ha (also 10 properties/Ha), about three-quarters of the reduction coming from lower unit production costs and one quarter from lower unit distribution costs.

e. Wider implications of the research

These examples are enough to illustrate the range of effects that might be observed, but which are particularly relevant when thinking about urban infrastructure? In studies of agglomeration, it is common to use population as the measure of size[9]. One lesson from these examples is that it may not be sufficient to look at numbers alone. Whereas increase in size through densification would, it seems, bring economies of scale (in water supply at least), with a positive influence on agglomeration, as would (to a lesser extent) constant density increase, increase in size through suburbanization would be roughly neutral in cost terms. To get the full picture, it would appear necessary to take density explicitly into account as well as size. Moreover, it would be misleading to regard urban areas of similar size, as measured by population, as equivalent from an agglomeration perspective, if they have very different densities. As the 'dispersion' example suggests, lower density towns or cities are likely to have higher distribution (and access) costs. Put differently, agglomeration by densification would have real cost advantages (at least up to the point where congestion costs become appreciable) whereas suburbanization would not.

Another way to look at the matter is to compare water supply costs as between a small town and a large one. Even if they have the same density, the 'constant density' calculations point to lower costs in the larger town. If this effect generalizes to other types of infrastructure, it suggests an important reason why large settlements might over time prosper more than small ones; and if the larger one is also denser, the advantage becomes greater still. A related point arises when an area is occupied by several small settlements rather than one large one. If each settlement operates its own water production facilities, it risks a double cost penalty, on the production side from smaller plant size and on the distribution side from greater dispersion. Of course, infrastructure costs are not the only consideration but if, for example, people have a preference for suburban living, these calculations indicate that there is likely to be a cost penalty (whether or not this is visited on suburbanites through tariffs and connection charges).

It has not been possible in **Chapter VII** to go beyond some pointers to the application of our water supply findings to a wider range of urban infrastructure. It is likely that distribution costs are less significant in the case of other utilities, although capital

[9] "The urban area population is the standard measure of urban size in studies of urbanisation economies." Eberts & McMillen (1999, p.1481). Although urban areas will by definition probably have relatively high densities, there can still be considerable variation in density between one urban area and another.

investment in distribution systems is important. While in general lower distribution costs can be expected to favour agglomeration by extending the area that can be economically served, high capital costs will still require that settlements be dense as well as relatively large if the necessary investments are to be viable. At the same time, we have pointed to some developments, such as small sewage treatment works and local power generation, which may help small settlements. The scope for application to Point-type infrastructure, such as hospitals, appears good. There has been a tendency to disregard access costs in these cases but the methods we have developed for water distribution costs could readily be applied – the effect, it appears, given that health authorities (like water companies) have to take the existing pattern of settlement as given, would probably be to moderate enthusiasm for over-large facilities.

Application to transport is less obvious. While there are some suggestive similarities, notably when the spatial aspect of transport networks is under consideration, transport also raises issues which go beyond those examined in this thesis. An important instance is congestion, which hardly arises in the case of water supply[10] but is of considerable importance in transport. At the same time, the role of density in facilitating the provision of low cost, high capacity transit has parallels in water supply, as does the difficulty of maintaining viable public transport where density is low, for reasons entirely analogous to those applying to water distribution, i.e. higher infrastructure requirements and longer distances per unit of output.

What is clear is that economies of scale in production are not the only factor at work. The spatial aspect with its impact on distribution and access costs is also important. In this research, we have tried to bring this aspect into focus by considering four contrasting urban growth scenarios, characterised as (a) *densification*, (b) *dispersion*, (c) *suburbanisation*, and (d) *constant density*. The results have been discussed in **Chapter VI**. The general conclusion emerging from this work is that scale effects in infrastructure may depend as much on density as on size *per se*. High density settlement has the potential to permit both large scale production and low cost distribution; on the other hand, low density adds to distribution (or access) costs. It follows that the general presumption in urban economics that such services are always characterised by economies of scale and therefore conducive to agglomeration may not be correct. This suggests that there should be more direct consideration of density effects in studies of

[10] The drop in pressure which can occur at times of peak demand for water is perhaps the nearest equivalent.

urbanisation economies (by including density as an independent variable, or both population and area, or by using some measure of sprawl as a proxy for density).

f. Limitations of the research

- The absence of any price effects in the cost functions used means that some effects may have been missed (e.g. greater use of more capital intensive or automated technologies in areas where labour costs are relatively high). However, this limitation may not be too serious in an industry where technology is fairly standard and when the analysis is a single year cross section.
- Treatment of the demand for water has been largely by-passed in this thesis. In the case of water supply this can be defended on the grounds that there is a legal obligation to supply and that consumption is very insensitive to price effects, particularly where supply is unmetered. In effect demand has been assumed to be exogenous, both in its locational and its quantitative aspects. In extending the results to other infrastructure services, this stance would be less easy to defend.
- More generally, more case studies using actual costs for particular areas would also be desirable, to explore in more detail how the combined costs of production and distribution are optimised in practice (or how other factors, such as security of supply and water quality considerations, lead to arrangements that are not strictly cost-minimising).
- Although congestion is potentially an important negative factor in urban agglomeration, little evidence of this has been found in the case of water supply. It is likely to be more significant in the case of other infrastructure, particularly transport.
- The consideration of the wider implications of this research in **Chapter VII** has only scratched the surface. There is much scope for further research, particularly into the part played by density in the economics of agglomeration.

5. Outline of thesis

The structure of the thesis is as follows: **Chapter II** defines infrastructure, briefly reviews the literature on infrastructure and agglomeration, as well as that on sprawl and the cost of public services, and explains the choice of water supply as a case study. **Chapter III** then, drawing on the literature reviewed in **Appendix B**, discusses the methodological implications of the different characteristics of water production and water distribution and hence the justification for examining them separately in the empirical work described in **Chapter IV** (water production) and **Chapter V** (water distribution). **Chapter VI** then brings these results together showing how economies of volume scale in both production and distribution can be offset to a greater or lesser extent by spatial costs. Comparisons are made with results obtained by other researchers using other methods. Finally, **Chapter VII** draws together the conclusions on the interaction between economies of scale, distribution costs and density effects in urban water supply and then considers how far these conclusions may be generalisable to other types of infrastructure.

6. Acknowledgements

I would like to take this opportunity to acknowledge the support and encouragement of my supervisors at the London School of Economics, Prof. Paul Cheshire and Steve Gibbons; I have also benefited from the advice and teaching of other LSE staff, particularly Henry Overman, Christian Hilber and Christine Whitehead. At an early stage in the research, I had a most useful discussion with Scott Reid, then with Stone & Webster Consultants now at ICF Consulting, which helped me to identify much of the relevant literature and gain an appreciation of areas likely to be fruitful to research. At Ofwat, Mark Hann, Helen Twelves and Jackie Cranmer (Librarian) helped me to locate data and answered queries; and at AWWA, Kurt Keeley (Data Manager) kindly provided me with a copy of their 1996 survey data on disk. I am particularly grateful to my contact at “BWC”, who wishes to remain anonymous, but without whose cooperation in the form of large amounts of internal data it would not have been possible to obtain many of the results reported in this thesis. Finally, thanks to my wife for her discreet support and encouragement.

The use of Ofwat data in this research is authorised by HMSO licence No. C2007000459.

II. INFRASTRUCTURE AND THE URBAN ECONOMY

"Cities are the summation and densest expressions of infrastructure" Herman and Ausubel (1988, p.1).

1. Defining infrastructure

While the importance of infrastructure in urban development is generally recognized, the treatment of infrastructure in texts on the economics of urbanization tends to be perfunctory and there is some fuzziness about what is actually meant by infrastructure. Recently, there has been a tendency for the term to be applied just to transport infrastructure, particularly roads. However, if the aim is to understand the role of infrastructure in urban development, a wider definition is appropriate.

"Infrastructure is the term applied to large-scale engineering systems and includes a variety of public works, such as roads, bridges and sewer systems, as well as privately managed utilities such as electric power and telephone service" is Herman and Ausubel's attempt at a definition. Later in the same volume Beckmann (1988, p.98) offers "Infrastructures are basic to all economic life. The urban infrastructure is one of the most diverse and complex. To name only the most important components, it includes streets and public transportation; water supply and sewage removal; police and fire protection; judicial, educational and health facilities; and parks and other recreational facilities." This is better in that it recognizes the variety of types of infrastructure and it also distinguishes between urban and other infrastructure. However, the inclusion of "police and fire protection" apparently widens the scope of the term to include operational personnel as well as the buildings from which they operate.

Infrastructure can be seen as lying somewhere in a spectrum that ranges from climate and geography at one end to the services and operating systems associated with infrastructure at the other. Within this spectrum, a distinction can usefully be made between amenities and infrastructure. Gyourko & Tracey (1991, p.775) suggest "A pure amenity is a non-produced good such as weather quality that has no explicit price". Brueckner *et al* (1999, p.94) make a similar distinction when they say "Natural amenities are generated by an area's topographical features, including rivers, hills, coastlines, etc." On this basis, natural features such as rivers, lakes and favourable sites count as amenities rather than infrastructure even though they might (for example) reduce the need for water services infrastructure or provide natural fortification and sound foundations for buildings. This seems more straightforward than having a category of "natural infrastructure", while recognizing that infrastructure costs may differ between sites because of such factors.

Bartik and Smith (1987, p.1210-11) widen the term amenity to include not only intangible features of a place such as air quality and "the charm of a historic neighbourhood" but also at least some public services, such as education and police services; and this is the way the term is generally used in the hedonic pricing literature. In Cheshire & Sheppard's 1995 article "On the price of land and the value of amenities" the term amenities embraces such characteristics as "the character of neighbouring houses and households, localized traffic effects and the quality of the micro-environment and local public goods such as schools" (p.247). This extends the term to cover public and other services, which may or may not be associated with physical infrastructure.

In the end, there is probably no "right" definition for either infrastructure or amenities, much depending on the context in which the terms are being used[11]. In the present context, there is advantage in reserving the term infrastructure for structures and facilities that are the result of human intervention, creating something physical that was not there before. This definition excludes amenities and services but is still wide enough to embrace the movement of soil to create embankments or cuttings as well as the erection of buildings and the laying of lines. This is similar to the position taken by Biehl (1986, p.87): "The difference between infrastructure and other potentiality factors, such as the location of the region or its natural resource endowment, is that the service bundles inherent in infrastructure have been 'artificially' created through investment, whereas location and natural resources are 'naturally' given." It enables attention to be focused on specific well-defined facilities, with identifiable costs, whose existence is the result of deliberate decisions by public or private entities. Unlike (say) the weather, the amount of infrastructure of this kind is a matter of choice.

[11] Indeed, the suggestion was put to me by Prof. Cheshire that: "Maybe one should not think of infrastructure at all, rather a set of services necessary for cities, which are complements to private and public consumption and can be produced from investment in collective goods, the natural environment and even organization/application of knowledge."

2. Infrastructure and agglomeration

It might be thought that something as basic as the economics of infrastructure would already have been thoroughly investigated and the results embodied in standard textbooks. However, this is far from being the case. One possible explanation is that it has not been regarded as something worthy of study in its own right. Thus, while housing, transport, public goods and utilities, all in their different ways constituent parts of the urban infrastructure, have each generated a substantial literature, it has been somewhat compartmentalised and a unified treatment of the role of infrastructure in urban development has been lacking. In the case of utilities, in particular, the focus of research has been industrial organisation and (more recently) regulation, rather than the contribution of utilities to urban development or agglomeration economies.

However, awareness of the interaction between different elements of the urban infrastructure has increased in recent years. As evidence, one can cite the contents of successive volumes of the Handbooks of Regional and Urban Economics[12]: the juxtaposition of chapters on Housing, Urban Transportation and Public Facility Location in Vol 2, for example, must, one supposes, have prompted at least some readers to speculate about the relationship between these topics.

Nevertheless, in all four volumes, there is only one chapter which puts infrastructure centre stage. That is the contribution of Eberts & McMillen (1999) on “Agglomeration Economies and Urban Public Infrastructure” who summarise the position thus (p.1456):

> “Theory links [agglomeration economies and urban public infrastructure] by positing that agglomeration economies exist when firms in an urban area share a public good as an input to production. One type of shareable input is the close proximity of businesses and labor that generates positive externalities … Another perhaps more tangible type of shareable input is urban public infrastructure. Public capital stock, such as highways, water treatment facilities and communications systems, directly affect the efficient operation of cities by facilitating business activities and improving worker productivity. The literature has devoted considerable attention to both topics, but not together …Only a handful of studies have focused on the metropolitan level, and even fewer have estimated agglomeration economies and infrastructure effects simultaneously. Results from studies that include both types of shared inputs suggest that both spatial proximity and physical infrastructure contribute positively to the productivity of firms in urban areas.”

They conclude:

[12] Nijkamp (Ed) (1986) Vol 1; Mills (Ed) (1987) Vol 2; Mills & Cheshire (Eds) (1999) Vol 3; Henderson & Thisse (Eds) (2004) Vol 4.

> "More research is needed to explore the inter-relationships between urban size and urban public infrastructure and to open the 'black box' of agglomeration economies and estimate how the various other factors associated with urban size affect productivity."

One reason identified in the urban economics literature why larger size may bring costs as well as benefits is commuting costs. As Fujita & Thisse (2002, pp.108-9) express it:

> "Intuitively, the reason is that, because of an increase in travel distance, the total commuting costs within the city increase more than proportionately with the population size. In other words, given the monocentric structure, *there are diseconomies in urban transportation when the population rises*. This result coincides with another well-documented fact in economic history that high commuting costs placed an upper limit on the growth of cities for fairly long periods (see Bairoch (1985), chap. 12)."

An early statement of this result is provided by Arnott (1979). If it is assumed that all employment is concentrated in a central business district, population density is uniform over the city area and commuting cost is increasing in distance from the CBD, then it is not difficult to show that total commuting cost for the city is increasing in city size as measured by population[13]. (Note, however, that this result depends on each commuter traveling radially and individually to the CBD; it does not allow for the possibility that large dense populations will permit the development of collective means of transport, thereby greatly reducing the average cost of commuting; nor is constant density consistent with the standard monocentric urban model which implies that density declines away from the centre.)

It follows, as (Fujita (1989, p.134) puts it: "In order to have cities, therefore, we must have technological advantages in production or consumption that exceed the transport cost increase". Developing the argument, Fujita continues:

> "Perhaps the most fundamental reason for the existence of cities stems from *economies of scale* in production and consumption, which are, in turn due largely to the *indivisibility* of some commodities (such as persons, residences, plants, equipment, and public facilities). The indivisibility of persons leads to the specialization of labour, and some equipment can be effectively used only on a larger scale. Moreover, the efficient coordination of many specialized persons, equipment and production processes requires them to locate nearby – due partly

[13] For a circular city of uniform population density, where all commuting is to a central business district and transport cost is proportional to distance, total commuting costs will be given by:

$$TCC = \int_0^R r.2\pi r.dr = \frac{2}{3}\pi.R^3 = \frac{2}{3\sqrt{\pi}} N^{3/2}$$

Where *R* is the radius of the city and *N* is its population, i.e. total commuting costs increase more than proportionately with population, and average commuting cost is an increasing function of *N*.

> to the facility of communication and partly to transport cost savings in various production processes. Therefore, the average total cost of the production of a good will be smaller (to a certain extent) if it is performed at a *larger scale* and at a *contiguous location*. In addition, if the production of one firm uses an output of another firm, the two firms may find it economical to locate near each other. Hence, through *input-output linkages*, many large firms may find it economical to locate closely, and these firms will provide the basic sectors of a large city. Moreover, the provision of many *public services and facilities* (such as schools, hospitals, utilities, and highways) typically exhibits the characteristic of economies of scale."

This last sentence encapsulates the dominant view of the role of infrastructure in the urban economics literature. Its contribution comes from economies of scale. Indeed, in the model developed later in Fujita (1989), p.151-2, this becomes very explicit:

> "We assume that the formation of a city requires a certain amount of fixed costs *K*. For example, *K* may include construction costs of basic public facilities such as transport and water systems. Since the per capita fixed costs become smaller as the population increases, the existence of fixed costs provides an incentive for city formation."

In effect, infrastructure is treated here as a local public good, so that distribution or access costs are not considered, although some attention is later given to congestion costs and other externalities (positive as well as negative). It is one of the objectives of this thesis to bring out more clearly how distribution (or access) costs interact with economies of scale and urban density to complicate this picture of the role of infrastructure.

Duranton & Puga (2004), discussing the micro-foundations of urban agglomeration economies, suggest (p.2066) three types of micro-foundation, based on "sharing, matching and learning mechanisms". Within the sharing type, they include "sharing indivisible facilities, sharing the gains from the wider variety of input suppliers that can be sustained by a larger final-goods industry, sharing the gains from the narrower specialization that can be sustained with larger production, and sharing risks". It seems clear that sharing urban infrastructure must be included among "indivisible facilities", to the extent that urban infrastructure is to be considered a source of urban agglomeration economies.

Duranton & Puga continue (p.2068):

> "Here we just describe briefly how one large indivisibility could provide a very simple formal motive for the existence of cities. Consider then a shared indivisible facility. Once the large fixed cost associated with this facility has been incurred, it provides an essential good to consumers at a constant marginal cost. However, to

> enjoy this good consumers must commute between their residence and the facility. We can immediately see that there is a trade-off between the gains from sharing the fixed cost of the facility among a larger number of consumers and the costs of increasingly crowding the land around the facility (e.g. because of road congestion, small lot sizes, etc.). We may think of a city as the equilibrium outcome of such trade-off. In this context, cities would be no more than spatial clubs organized to share a common local public good or facility."

However, the authors then make clear that they do not regard this line of argument as particularly compelling (p. 2069):

> " … the easiest route to take in justifying the existence of cities is to assume increasing returns at the city level by means of a large indivisibility. While large indivisibilities are useful modeling devices when the main object of interest is not the foundations of urban agglomeration economies, they side-step the issue of what gives rise to increasing returns at the level of cities. Cities facilitate sharing of many indivisible public goods, production facilities, and marketplaces. However, it would be unrealistic to justify cities on the basis of a single activity subject to extremely large indivisibilities. The challenge in urban modeling is to propose mechanisms whereby different activities subject to small non-convexities gather in the same location to form a city."

These quotations can be read as casting doubt on the realism of the analysis developed in the Fujita passages just cited. However, they do not attempt to quantify the postulated trade-off, nor consider the point that higher density (here dismissed as "crowding") might contribute positively to the outcome, and their comments would seem to apply only to what we have called Point Type infrastructure. In consequence, Duranton & Puga go on to give the bulk of their attention to other forms of sharing, and to matching and learning mechanisms, with the implication that infrastructure is of little relevance to agglomeration economies.

It seems intuitively obvious that density must play an important part in urban economics, yet this aspect of the urban scene is not often directly addressed[14]. As Ciccone & Hall (1996) remark (p.96):

> "Although the idea that denser economic activity had advantages from agglomeration was implicit in a large earlier literature, there does not appear to be any earlier work in which density was an explicit element of the theory, nor has there been empirical work based on density."

The focus of their article is the benefit of density to productivity, which they find to explain more than half of the variance of output per worker across the states of the USA. They invoke three mechanisms to account for this (p.54):

[14] Indirectly, of course, terms such as "concentration" and "agglomeration" obviously imply density but a wide range of densities can be covered by such terms.

> "If technologies have constant returns themselves, but the transportation of products from one stage of production to the next involves costs that rise with distance, then the technology for the production of all goods within a particular geographical area will have increasing returns – the ratio of output to input will rise with density. If there are externalities associated with the physical proximity of production, then density will contribute to productivity for this reason as well. A third source of density effects is the higher degree of beneficial specialization possible in areas of dense activity."

Ciccone & Hall thus do not include the favourable effect of density on the unit costs of infrastructure services, as found in this research, as a further possible influence.

Our results for water supply suggest that Duranton & Puga (2004) may be too dismissive. In fact, there appear to be two kinds of sharing benefits at work in this case. An increase in city population will enable economies of scale in water production to be exploited; and if this increase takes the form of higher density settlement, there will be reductions in per capita distribution costs as well. However, if the increase takes the form of more dispersed settlement (expansion at lower density), the latter benefit may be reversed, perhaps even to the extent of outweighing the economies of scale in production. If similar conclusions hold for other infrastructure, such as sewerage, electricity supply and transport services, then, on the one hand, the cumulative advantage of high density settlement may be considerable (although, as these and other authors point out, account will also need to be taken of congestion effects) and, on the other hand, low density expansion, through higher infrastructure costs, will act as a brake on city growth. So, while invoking a single large indivisibility to explain urban agglomeration economies may indeed be unconvincing, the cumulative benefit of lower production and distribution costs across the whole range of infrastructure services when density is high cannot so lightly be dismissed. More generally, it suggests that urban theory and urban modeling should recognise that agglomeration benefits may depend as much on density as size (something that may equally well be true for the other sharing, matching and learning mechanisms to which Duranton & Puga give attention).

3. Sprawl and the cost of public services

A related literature considers how the cost of providing public services is affected by the spatial distribution of population. This has tended to focus on the question whether low density adds to costs: "Does sprawl cost us all?" as the title of one contribution puts it – Speir & Stephenson (2002). The question has quite a long history, particularly among urban planners in the US (e.g. Downing & Gusteley (1977), Frank (1989)).

In Britain, Elis-Williams (1987) used cost functions estimated from expenditure data, together with population distribution from the 1981 Census to determine the optimal location of secondary schools in the County of Gwynedd in Wales, taking into account both school costs and pupil transport costs. He observes (p.153):

> "Most local services are delivered from a number of identifiable service centres, each serving the population resident in the surrounding area. The population may travel to the service centre to receive the service e.g. hospitals, or the service centres may be bases from which the service is delivered to the population at home e.g. fire stations … In either case, there are two influences which a sparse population may have on unit costs –
> (a) it can force the operation of smaller and therefore less economic service centres;
> (b) it can cause higher transport costs because of the larger distances between the population and the service centres …
> … In the general case, it seems likely that there is a trade-off between economic operation of service centres and reducing transport costs, with authorities seeking to locate service centres at some optimum which minimizes total costs given the spatial population distribution."

Elis-Williams' study found that the actual location of secondary schools in the 5 districts of Gwynedd was reasonably consistent with the computed optimum. He also found greater sparcity to be associated with higher unit costs.

In a more wide-ranging study of 247 large counties in the US, Ladd (1992) estimated the impact on local government spending of two dimensions of residential development, growth rate and density, controlling for other determinants of per capita spending. She concludes, in contrast to the engineering and planning view that greater population density lowers the cost of providing public services, that there is a *U*-shaped relationship between spending and density: "Except in sparsely populated areas, higher density typically increases public sector spending". This study is note-worthy, *inter alia*, for its attempt to distinguish between costs and outputs. Its findings challenge the conventional wisdom but it may be noted that only the costs borne by government are considered.

More recently, Ladd's findings have been challenged by Carruthers & Ulfarsson (2003). They question the use of a simple density measure, particularly over areas as large as counties. Instead, they measure density as number of jobs and people per acre of *urbanized* land, with the spatial extent of urbanized land in a county given by the total number of developed acres. Using a cross-section of 283 US metropolitan counties, they found in an earlier report (p.506) that “per capita spending on infrastructure declines at greater densities but increases with the spatial extent of urbanized land area and property values”. Developing this approach to a wider range of government expenditures reinforced this conclusion (p. 513, p.517):

> “First, the parameter estimates for density are negative and significant in several of the models, suggesting that it creates economies of scale for: public spending on the whole (total direct expenditure), capital facilities, roadways, police protection, and education. For each of these services, the per capita cost decreases as densities increase, with the greatest savings realized in areas with very high densities. An individual police officer patrolling a square mile in a dense urban area may provide protection to many more people than his or her counterpart in a suburban area. Likewise, fewer roads are needed in high density areas, and school systems may be operated more efficiently – fewer (though larger) schools and less bussing of pupils are needed, for example … Overall, the models provide good evidence that density works to increase the cost-effectiveness of public service expenditure.
>
> Second, the spatial extent of *urbanized land* is positive and significant in most of the models, indicating that the spread of a metropolitan area plays an important role in determining public service expenditure. As explained in the background discussion, urban sprawl requires roadways and sewer systems to be extended over long distances to reach relatively fewer people. Trash collection and street cleaning activities must cover larger areas and, similarly, police and fire protection are spread thin, requiring more patrols and, potentially, more station houses to achieve a given level of service. In the case of parks and libraries, a greater number of facilities must be built in order for people throughout the metropolitan area to enjoy equal access.”

Overall, Carruthers & Ulfarsson conclude (p.518):

> “By far the most salient finding of the analysis is that the per capita cost of most services declines with density (after controlling for property value) and rises with the spatial extent of urbanized land area. This reinforces planners’ claim that urban sprawl undermines cost-effective service provision, and lends support to growth management and ‘smart growth’ programs aimed at increasing the density and contiguity of metropolitan areas – at least from the standpoint of public finance.”

Evidence of a similar, if more limited, kind is provided by Speir & Stephenson (2002). They aim to throw light on the effect of different patterns of housing development on water and sewerage costs by computing the infrastructure (pipes, valves, etc) and energy (pumping) costs of developments which vary in lot size (separation between houses), tract dispersion (separation between development tracts), and distance (separation from existing water and sewer centres). The results show that “smaller lots, shorter distances between existing centres, and lower tract dispersions reduce water and sewer costs” (p.64) but that lot size is the most significant factor, because “infrastructure within the development tract – water distribution mains and sewer collector mains – are the two largest components of total cost" (p.60).

Also of interest is a study by Sole-Olle & Rico (2008) of the effect of sprawl on the cost of municipal services in Spain. Using data from some 2,500 municipalities[15] for 2003, they estimate per capita expenditure equations for types of spending thought most likely to be affected by sprawl (community facilities, basic infrastructures and transportation, housing and community development, local police, culture and sports, and general administration) accounting for about 70% of local spending. The main measure of sprawl is urbanized land per capita in each municipality (in 4 bands: < 75 sq.m/pop (very compact); 75-160 sq.m/pop; 160-700 sq.m/person; >700 sq.m/pop (very dispersed)). Additional indicators are numbers of residential houses, % of scattered population, and number of population centres. Also included in the estimation are variables intended to distinguish the effects of urban sprawl on local costs from those of other cost and demand factors. They find (p.28):

> "In general, our estimation results indicate that low-density developments led to greater provision costs in all the spending categories considered, with the exception of housing. By adopting the piecewise linear function assumption we were able to disaggregate this total effect, revealing that the impact on total costs accelerated at very low and very high levels of sprawl … Further, the impact of urban sprawl on the provision costs of the public services considered here was particularly marked at high levels of sprawl … These results suggest that in municipalities with a spatially expansive urban development pattern, the provision costs of public services [per capita] increase initially as a result of increasing road construction costs and rising general administration costs, and then, if the urban sprawl advances further, costs continue to rise as a result of higher costs of providing community facilities, housing, local police and culture. In those municipalities with very low levels of urban sprawl (< 75 sq.m/pop), the increase in local costs was due to public services other than those analysed here."

This brief, and necessarily selective, review of some of the literature on the economic effects of sprawl generally lends support to the proposition that less dense development is associated with higher unit costs for the provision of public services, as intuition would tend to suggest. However, one problem dogging all those working in this field has been how best to characterize sprawl. Simple population density is a very crude measure, particularly if the area units are large. To overcome this, some researchers measure density in relation to developed land (so excluding undeveloped land). But this still gives only an average measure (at a point of time) for what may be quite a large

[15] i.e. nearly all municipalities with population > 1000.

area, whereas sprawl implies an expanding fringe of rather low density development. To capture this one ideally needs some measure of the rate of expansion over time and of the relative density at the fringe compared with the centre. It is a limitation of the studies cited in this section that they have not really been able to resolve this issue. A breakthrough in this regard is offered by Burchfield *et al* (2006), who propose and implement for the US a new index of sprawl as (p. 587) “the amount of undeveloped land surrounding an average urban dwelling”.

More precisely, Burchfield *et al*’s index is arrived at by measuring for each 30 x30 metre cell of residential development, the percentage of undeveloped land in the immediately surrounding square kilometer; this measure is then averaged across all developed cells in a metropolitan area. Using this measure, the authors compare the spatial structure of urban development in the US in 1992 with 1976. They find (p.597) that “While a substantial amount of scattered residential development was built between 1976 and 1992, overall residential development did not become any more biased towards such sprawling areas”. The explanation is as follows:

> “To reconcile these apparently conflicting tendencies, note that the distribution of the final stock of development across different degrees of sprawl is *not* the result of adding the distribution of the flow of new development to the distribution of the initial stock. The reason is that, by adding the flow of new development to the initial stock, the distribution of the initial stock becomes shifted to the left as infilling makes formerly sprawling areas more compact … It helps to consider how the environment might have changed near a hypothetical house located in a medium density suburb. The open space in the immediate neighborhood of this house will most likely have been partly infilled. Areas initially more compact, presumably closer to downtown, will have experienced less change. Undeveloped areas further out may now be scattered with low density development. To the family living in this house, the pattern of residential development around them is very different from the one they experienced in the 1970s. However, if we zoom out and look at the city from a distance, we see little change, at least in the proportions of sprawling and compact development: The new city is just like an enlarged version of the old city.”

This last observation is appealing but somewhat imprecise. A monocentric city can be approximately characterized by four parameters: d_0, its density at the centre; λ, the rate at which density declines away from the centre; N, its population; and R, its radius. These parameters are not entirely independent but are related to each other by:

$$N = \int_0^R 2\pi.r.d_0 e^{-\lambda r} dr \qquad \ldots\ldots\ldots\ldots.. \qquad (2.1)$$

The evolution of the typical American city, as portrayed by Burchfield *et al*, can now be represented through this relationship. Between 1976 and 1992, *N* has increased substantially, and been absorbed through a small increase in d_0, its density at the centre, a fall in λ due to infilling, and an increase in *R*, the city radius[16]. This is sketched in **Figure 2.1** below:

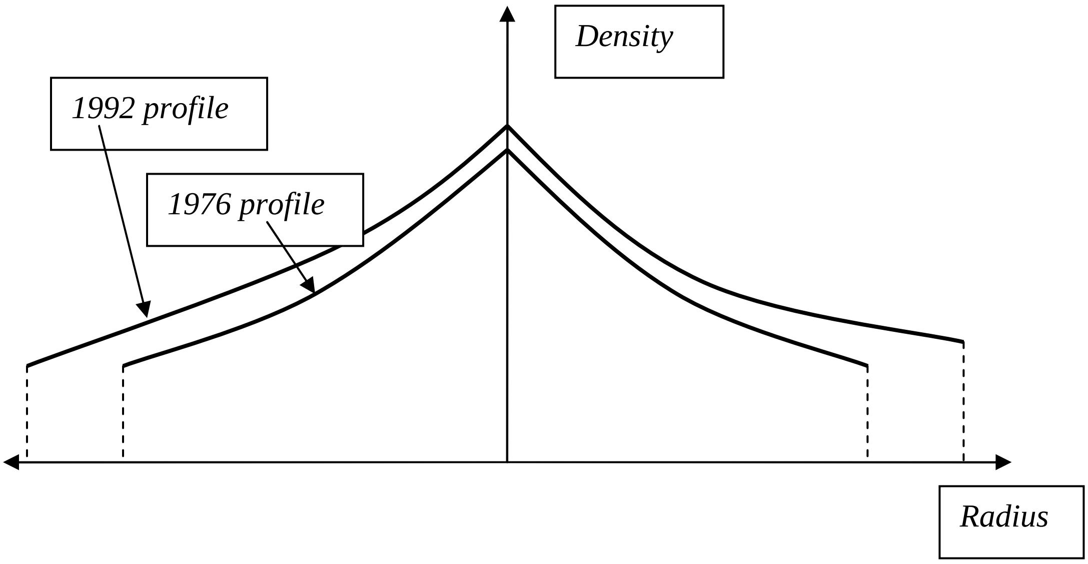

Figure 2.1: Changing profile of a typical US city (not to scale)

This representation offers scope to give a more nuanced account of whether or not sprawl has increased, by considering the relative contributions of the different parameters to the change in profile. In **Chapter V** this approach will be further developed to illustrate the impact on water distribution costs of four different development scenarios: (a) *densification* (*R* constant as *N* increases); (b) *dispersion* (*N* constant as *R* increases); (c) *suburbanization* (λ constant as *N* increases); and (d) *constant density* (density constant as *N* increases). A further advantage of this approach is that it is relatively undemanding in terms of data requirements. Given any 3 parameters, the 4th can be calculated: for example, if d_0, *N* and *R* for a city are known (as will often be the case), λ can be calculated. For given d_0 and *N*, λ or *R* provides a measure of dispersion (= sprawl). In addition, (2.1) can be adjusted to accommodate some variants on the monocentric shape e.g. for a semi-circular (coastal) city, (2.1) would become:

$$N = \int_0^R \pi . r . d_0 e^{-\lambda r} dr \qquad \ldots\ldots\ldots\ldots\ldots \qquad (2.2)$$

[16] It is an empirical matter how sharp the drop in density at the city boundary may be. Observation suggests that in some cases, it can be quite abrupt; in other cases the decline in density continues for some distance before development peters out or another settlement (with higher density) is encountered. There is also the question whether to use the administrative or some other boundary as the cut-off.

4. Types of urban infrastructure

Picking up the definition adopted in **Section 1** above: “structures and facilities that are the result of human intervention, creating something physical that was not there before”, **Table 2.1** attempts a comprehensive listing of all the different elements of the urban environment that might be taken to constitute “infrastructure”. These are grouped under five broad headings:

- **Buildings** (further sub-divided into residential, commercial, public service and leisure/entertainment);
- **Roads, streets and related items**;
- **Other transport systems**;
- **Utilities**; and
- **Other**.

The resulting categories seem reasonably well-differentiated[17], although the last is inevitably something of a rag-bag, but are they also analytically interesting? In other words, what are the distinctive economic characteristics of infrastructure and do these vary systematically according to the type of infrastructure under consideration?

[17] They may be compared with the categories developed in a regional rather than an urban context by Biehl (1986, p.109): A. Transportation; B. Communication infrastructure; C. Energy supply infrastructure; D. Water supply infrastructure; E. Environmental infrastructure; F. Education infrastructure; G. Health infrastructure; H. Special urban infrastructure (incl. Fire stations, urban parks, etc); I. Sportive, Touristic facilities; J. Social infrastructure; K. Cultural facilities (incl. Museums, theatres, etc); L. Natural endowment.

Buildings

a. residential
Houses
Apartments
b. commercial
Factories, warehouses
Offices
Shops
Hotels
Restaurants, bars
c. public service
Parliaments, town halls
Government offices
Schools, universities
Hospitals, health centers, surgeries
Post offices
Police/fire stations
d. leisure/entertainment
[Palaces]
Theatres, cinemas, concert halls
Art galleries, museums
Sports centers, gyms, swimming pools
Churches, mosques, etc
Zoos

Utilities

Water supply
Sewerage
Storm drains
Electricity supply
Gas supply
Telecommunication systems
Wireless communication facilities

Roads, streets and related items

Paved streets/roads
Pavements
Street lighting
Traffic control equipment
Bridges
Underpasses
Road drainage
Bus stations
Bus shelters/stops
Car parks, parking facilities
[Trees, flower beds]

Other transport systems

Railway stations
Railway lines
Underground stations
Underground lines
Tram lines
Light rail systems
Ferries
Canals
Docks, quays, jetties
Airports, heliports

Other

Monuments, landmarks
Parks/Open spaces
Playing fields
Children's playgrounds
Town wall

Table 2.1: types of urban infrastructure

A firm of economic consultants, OXERA (1996), have suggested that:

> "There are a number of features which distinguish infrastructure projects from other investment programmes:
>
> - *Fixity or the stranded nature of assets*. Infrastructure projects tend to be fixed – only at the place where the project exists can it offer a service to customers. It cannot service a general market in the way that a factory can. If the market has been assessed incorrectly the product cannot be sold to others elsewhere.
> - *Large units of investment*. Even small projects require relatively large sums of capital investment. It is often difficult to assess the public's response to the projects prior to their opening since the market testing techniques that exist for conventional products cannot easily be applied to new infrastructure projects in advance.
> - *Initial overcapacity*. Infrastructure is often built with long term predictions of the growth of demand in mind, such as increases in the volume of traffic,

> which lead to an initial period of overcapacity. As a consequence, the revenues that accrue from such projects tend to be relatively small compared to the level of investment in the first few years of the project's life.
>
> - *High up-front costs*. The costs incurred at the start of the project, relative to the ongoing operating costs, are high for infrastructure projects.
> - *Benefit appropriation and the existence of public good attributes*. It is generally the case that not all the benefits that result from investments in infrastructure can be captured by those responsible for the investment."

However, not all infrastructure exhibits all these characteristics and even with our relatively restricted definition of infrastructure, it is still open to question whether infrastructure so defined can be treated for analytical purposes as a single category. Gramlich (1994, p.1177) begins to wrestle with the issue, saying "The definition that makes the most sense from the economics standpoint consists of large capital intensive natural monopolies such as highways, other transportation facilities, water and sewer lines and communications systems." This opens up questions about economies of scale and market structure, which are indeed important in the analysis of infrastructure. However, he goes on to consider only public sector owned tangible capital stock because of the difficulty of obtaining data on private infrastructure capital; and he does not address the implications of differences in the character of the goods and services produced by infrastructure. In contrast Biehl (1991, p.10) suggests "If the term 'infrastructure' is used in order to designate that part of the overall capital stock of an economy that possesses high publicness, infrastructure becomes a determining or limiting factor to growth as it will not be provided by private transactions during economic growth." This puts the focus on the demand for infrastructure and how it is used rather than conditions of supply. This distinction between supply characteristics and demand characteristics seems worth developing further.

a. Supply characteristics

A first (and rather obvious) generalization as regards the **supply side** is that infrastructure involves investment; infrastructure items are capital goods; and they use land. Rather confirming the views expressed by OXERA quoted above, a pervasive feature evident in scanning the items listed in **Table 2.1** is that infrastructure provision typically involves high fixed costs and low operating costs. It generally requires building or construction work, and that means significant initial costs and a degree of indivisibility. It may indeed require a substantial minimum scale to be worth providing at all. Operating costs on the other hand are often low (although costly periodic maintenance may be required). Together these features imply strong economies of scale, with average long term supply costs falling as the amount provided increases (although in the short term, which may be quite a long time itself given the scale and fixity of much infrastructure, supply costs may rise sharply as capacity constraints bite). There may also be discontinuities in the supply function as larger scale provision allows better technical solutions to be adopted or technical progress makes older infrastructure redundant. Moreover, once installed, the resulting facility has a degree of permanency and cannot easily be relocated elsewhere – it becomes a sunk cost.

Focusing specifically on **Utilities**, they are certainly characterised by investment in large scale systems, with high fixed costs and low operating costs, the classical case of scale economies. Additionally however there are typically two parts to utility systems: production facilities (power stations, water treatment plants, etc) and distribution facilities (water mains, telephone lines, etc). With production facilities, the scale economies are no different from those encountered in manufacturing industry and process plant. There is therefore an existing body of analysis and results that can be brought to bear when considering this part of utilities' infrastructure. Utilities' distribution systems bring different considerations into play. It is much less clear whether economies of scale apply to distribution systems. As one early expert on electricity distribution, Sayers (1938, p.2), commented:

> " … if an area of supply lying wholly within a radius of, say, three miles from a generating station is extended to double that radius, two consequences follow. As the whole of the supply to the new area has to be transmitted to the boundary of the old area and then spread across the radial breadth of the new area, the average distance of transmission to the consumers in the new area will be about doubled. Whilst the new area is three times as extensive as the old one, it is generally less densely populated so that there are fewer consumers per mile of mains with, consequently, a smaller annual consumption in proportion to the capital employed …"

It thus appears distinctly possible that distribution may be subject to *diseconomies of scale* with respect to the geographical size of the supply area. On the other hand, the distribution cost per consumer will be affected by the density of the population. Denser populations will generally require less reticulation per head and may offer other savings, such as reduced pumping costs or lower transmission losses. *Density economies* of this kind are likely to be important in considering urban infrastructure. Where a utility produces more than one service (e.g. water companies that offer both water supply and sewerage) *economies of scope* may also need to be considered.

b. Demand characteristics

Turning to the **demand side**, the emphasis shifts from the character of the facility to the nature of the services being provided. There is generally a communal element in the consumption of infrastructure services and there may be difficulty about excluding people from participation, at least to some extent. As has been recognised since Samuelson (1954), these features make charging difficult or inefficient. Moreover, the services provided by the infrastructure are accessible only within a defined area and the enjoyment of infrastructure services is likely to be impaired because of congestion as the numbers seeking to take advantage of them increase - and this impairment may start to occur well within any physical capacity limit. In some cases, the service is only accessible at the point where a facility is located. In such cases, access costs as well as the potential for congestion will affect demand.

These features tend to differentiate infrastructure services from other goods and services. A classification scheme which incorporates these distinctions is set out in **Table 2.2** below. As can be seen infrastructure services are mainly found in categories B and C.

	Tangible (“goods”)	**Intangible (“services”)**		
	Individually consumed	**Individually consumed**	**Collectively consumed**	
			Excludable	**Non-excludable**
A. Can be made available anywhere	A1. Standard goods (eg. Can of beans)	A2. (eg. Insurance policy)	A3. (eg. GPS, radio/TV programmes)	A4. Pure public goods (eg. National defence)
B. Available only within a defined area	B1. (eg. Water supply, electricity)	B2. (eg. Mobile phone service, sewerage)	B3. (eg. Transport services)	B4. Local public goods (eg. Police service, untolled road)
C. Available only at point of supply	C1. (eg. Restaurant meal)	C2. (eg. Haircut)	C3. (eg. Public park, art gallery, concert/theatrical production, education services, public convenience)	C4. (Items such as the view of a landmark or monument might fit here)

Table 2.2: A classification scheme for goods and services, based on degree of “publicness” and accessibility

We may compare this classification scheme with the typology of collective goods proposed by Starrett (1988, p.42-3). He says "In characterising collective goods, it suffices to concentrate on features that make them 'unmarketable'". He identifies two features as key: *Non-excludability* and *non-rivalrousness*. In discussing the former, he observes: "To establish a workable property right, we must be able to assign ownership of the good to a single individual in a meaningful way and monitor the transfer of it among individuals in a (relatively) costless way. A meaningful assignation of ownership requires that the holder be able to withhold the benefits (or costs) associated with the commodity from others – thus, the idea of excludability." He goes on to note that in some cases exclusion is possible but only at a cost (which may be prohibitive). He suggests that therefore "we can rank commodities according to the cost of setting up and enforcing a private property right" while "for some intermediate cases, the answer may depend on context". In **Table 2.2** we recognise non-excludability in column 4; the point about exclusion possibly carrying a cost applies to items in column 3, and where that cost is prohibitive the items can be regarded for practical purposes as if they were in column 4.

Non-rivalrousness is a subtly different feature. Even if it is possible to exclude, it may not be desirable to do so. Starrett cites the example of a radio broadcast "where my access to it does not in any way diminish your capacity to benefit from it". He offers the definition "A good is non-rivalrous when the opportunity cost of the marginal user is zero". As with excludability, rivalrousness may be a matter of degree, depending on the size of the marginal opportunity cost. It may also depend on the level of provision. Starrett suggests that the opportunity cost may be measured either in terms of resources required (e.g. the costs of maintaining a highway or bridge if these vary with use) or utility foregone (e.g. the congestion costs associated with extra use of a park or museum). In **Table 2.2** congestibility increases in moving down the table, and is particularly strong in row C.

A point particularly worth noting from **Table 2.2** is the different implications for transport costs. In general, *goods* can be transported wherever they are required, at some cost, whereas *services* cannot. However, some services can be provided within a defined area but only if an appropriate distribution system is also provided, implying additional costs. For goods or services that cannot be transported or distributed, consumers must travel to access them, again at some cost. This distinction has been noted previously. Thus Thisse & Zoller (1983, p.2) observe:

> "Tiebout's second paper (1961) offers the basis of another perspective by reversing the problem under consideration. No longer concerned with consumers' choice among given service packages, he assumes a fixed spatial distribution of users. As consumers' benefits are now dependent on the distance to supply points (they are decreasing functions of distance to the closest facility) the notion of a pure public good becomes of little relevance. Investigating such space-generated impurities more deeply, Lea (1979) suggests making an essential distinction between traveled-for goods and delivered goods. To make this distinction clear, we hypothesize a system of established facilities. In the former case, users must travel to a facility in order to consume the public output. Examples are parks, libraries, hospitals, and so on. A demand function for services can then be derived (see Shepherd (1980)), which relates consumption to distance or transportation costs between residence and facilities, provided that people are actually able to adjust the level of services consumed … By contrast, consumers of *delivered goods*, such as emergency services or mail delivery, are not allowed to determine what facility will provide the service and, in general, do not have to bear the cost or inconvenience of travel. Distance *may* affect consumption, however, through service quality (fire protection is an example); this is reminiscent of Buchanan's (1965) theory of public goods subject to congestion, where consumers' utility decreases as the number of consumers increases. Two different mechanisms therefore lead to the same result, namely that *users' benefits are in most cases distance-dependent*."

However, the point does not appear to have been followed up to any great extent. In fact, Fujita & Thisse (2002, p.165, note 3) later remark "The distinction between *traveled-for goods* and *delivered goods* made by Lea (1979) is not essential for our purpose." This seems to be because they do not concern themselves with delivery costs, only with the perceived or actual deterioration in the quality of a delivered good as distance increases.

For infrastructure services such as water supply, however, delivery costs would seem to matter a good deal. We therefore see merit in this context in bringing back a distinction based on whether the service is delivered to the customer (at some cost) or has to be accessed by the customer (again at some cost). For this purpose, the terms Area-type infrastructure and Point-type infrastructure are proposed.

c. Network economics

Some infrastructure has network industry characteristics. Apart from economies of scale, Shy (2001, p.1) lists these as: complementarity, compatibility and standards; consumption externalities; and switching costs and lock-in. On the first point, Shy observes (p.2) that "Complementarity means that consumers in these markets are shopping for systems … rather than individual products … In order to produce complementary products, they must be compatible … This means that complementary products must operate on the same standard.” Network consumption externalities, the second characteristic, are illustrated by the example of a telephone service: Would anyone subscribe to a telephone service if nobody else had subscribed? In cases like this, the utility derived from the consumption of these goods is affected by the number of other people connected to the system. On the third characteristic, Shy says (p. 4): “The degree of lock-in is found by calculating the cost of switching to a different service or adopting a new technology, since these costs determine the degree to which users are locked into a given technology.”

d. Summary

It is evident that there are a variety of established economic approaches available to apply to the analysis of infrastructure. Useful generalizations to emerge from this discussion include:

- **Access type**: Most infrastructure services are distinctly local, being supplied either over a defined area (Area-type) or at a particular location (Point-type).
- **Collectiveness**: Many types of infrastructure involve collective use. However, this is not the case with residential buildings nor with utility services supplied to businesses and households.
- **Network effects**: This particularly applies to telecommunications, where much of the value of a system depends on the number of other subscribers; however, analogous effects are present in transport networks, as the number of places connected (by a metro, for example) increases.
- **Excludability**: While exclusion is possible in principle for most kinds of infrastructure, the cost and practicality of doing so varies widely. This aspect of infrastructure is perhaps best viewed as a continuum, ranging from the straightforward (cinemas) to the very difficult/expensive (urban roads). In some cases, such as enjoying the view of a landmark, exclusion may be impractical.
- **Congestibility**: Very few types of infrastructure are not congestible. This is a consequence of their local character (see “Access type” above). And, while in many cases, it may be possible to relieve congestion by increasing the scale of supply (widening a road, enlarging a museum), this will usually involve taking more land or otherwise impinging on existing activities. The main exception is wireless services (TV, radio, GPS, etc).

It is perhaps worth adding here that infrastructure may be the vehicle for some wider public goods effects, such as the public health benefits of clean water supply and good sanitation; it might also be argued that communications (whether of the physical or electronic kind) contribute a wider benefit in the form of “social glue”. Such considerations have implications for the optimal amounts of such services to provide but do not affect the costs of provision, which will be the main focus of this research.

While the above discussion applies particularly to individuals or households, the considerations are not very different for firms. It is commonly thought that firms may choose to locate in urban areas so as to be able to take advantage, at little or no direct cost, of shared infrastructure. However, this is quite compatible with the characteristics identified above and special consideration of firms’ use of infrastructure is not necessary here.

5. Water supply as a case study of urban infrastructure

Historically, the enormous implicit value placed on urban water supply is evidenced by the size of the investments towns and cities have made over the ages. In Europe, the Romans provided many spectacular examples. Frontinius, curator of Rome’s aqueducts in the AD 90s claimed that the maintenance of aqueducts was ‘the best testimony to the greatness of the Roman Empire’ (Bromwich (1996, p.110). Rome itself was an outstanding example: Eventually there were 11 aqueducts that supplied water to Rome, according to Mays (2002). Although, Mays adds (p. 1.28), “ … throughout the history of Rome, aqueduct construction was generally not planned in an orderly manner. During Republican Rome the city fathers tended to allow needs to become critical before aqueducts were built, similar to modern day practice.”

Other examples can be found in France. The supply to Nimes from Uzes is particularly well known. This remarkable aqueduct built around 20 BC, which includes the famous Pont du Gard (itself a massive three tier construction, some 250 metres long, carrying the water 50 m over the river bed), runs nearly 50 km, with a drop of only 17 metres. It also includes three smaller bridges and 35 km of underground channels. Less well known is that Paris (then known as Lutetia) also benefited from a Roman aqueduct some 26km long, running from Wissous (now part of Orly airport) to the Thermes de

Cluny[18]. It is not clear how long this remained in use but today only isolated fragments can be seen.

For an example nearer home, consider Birmingham. During the 19th century, as a consequence of the industrial revolution, Birmingham's population grew rapidly. Clean water was in short supply and there were major epidemics of water-borne diseases, including typhoid and cholera. Birmingham City Council, led by Joseph Chamberlain, set about finding a clean water supply for the city. A potential source was identified in the Elan and Claerwen valleys in North Wales, where there was high rainfall and geological conditions suitable for dam-building. Work started in 1893 and in 1904 the Elan dams were opened and water started flowing along 118km of pipeline to Birmingham. As the offtake is 52m above the Frankley Reservoir, the water flows by gravity alone. Now 300 million litres of water a day can be extracted from the Elan Valley to supply Birmingham[19]. In similar vein, much of modern Manchester's drinking water comes from Lake Vrynwy, also in Wales.

The enormous value of the accumulated investment in urban water supply in England & Wales is indicated by the following quotation from a pamphlet accompanying a Water UK Press Release dated 17 June 2004 "Water infrastructure: Building on our inheritance":

> "The total cost of replacing all the [water] industry's physical assets in England and Wales would be over £200bn. Three quarters of this is below ground. Collecting and dealing with wastewater costs more than supplying drinking water, mainly as larger pipes are needed. It would cost twice as much to replace the sewerage system than the water supply.
>
> At the moment, annual expenditure on maintaining these assets is £1.6bn. This is a significant sum but it is less than 1% of their replacement value. … There are 325,000km of water mains serving 23.6 million connections. On average each km of main serves 73 households. Mains vary significantly. Trunk mains, which transport water in bulk, can be 300mm to 1,800mm in diameter. Local distribution mains are usually smaller, with 125mm being a common size. Households are connected to the mains via service pipes. They are usually quite small, 25mm. …

[18] More details can be found at W D Schram's website www.cs.uu.nl/people/wilke/aquasite/paris/
[19] The information in this para has been extracted from http://www.bbc.co.uk/wales/mid/sites/history//pages/facts.shtml.

Water and sewerage systems at a glance

	Water	**Sewerage**
Length	325,000km of mains	302,000km of sewers
Connected properties	23.6m	21.8m
Treatment works	2,500	9,000
Replacement value	£70bn	£140bn
Layout	Mainly inter-connected networks (except in rural areas). Booster pumping stations to maintain pressure.	Small sewers joining large ones to single destination (the treatment works). Underground chambers to prevent flooding.

> … More than half of the mains below London are reckoned to be over 100 years old. One third are over 150 years."

Evidently, if all the different forms of urban infrastructure (buildings, roads, transport systems, other utilities such as electricity, gas and telecoms, etc) could be similarly valued, the resulting total would be very large indeed – probably dwarfing the investment in manufacturing industry, for example, which features so much more prominently in the urban economics literature. It does seem surprising therefore that urban infrastructure does not attract more attention.

Its evident importance provides one good reason for focusing in this research on urban water supply. It is also a good example of Area Type infrastructure. Furthermore, it has the advantages of a relatively straightforward technology, which does not vary much from place to place and evolves only slowly; there is only one (free) raw material and the costs of distribution are significant[20] – all of which should help to bring to light the effects of interest here. A further advantage is the public availability of most of the data submitted annually to the Office of the Water Regulator (Ofwat), known as the June Returns (Ofwat (2003a)). 43 tables in all, covering both financial and non-financial information, it is all compiled using the same guide-lines and so should be consistent across companies. However, as will later become apparent, because the water companies in England & Wales serve large areas with many settlements, it was necessary to seek more disaggregated information from other sources.

At the same time, there are some limitations to the use of water supply as a model for other types of urban infrastructure. When each town had its own gas works, and electricity generation was more local, the similarities were substantial. However, since

[20] "In gas and electricity, the indicative additional costs of transportation are approximately 2.5 – 5% per 100 km, while in water they are approximately 50%" Byatt *et al* (2006, p.390)

the 1960s, town gas works have been replaced by bulk supplies of natural gas from the North Sea and elsewhere, changing fundamentally the economics of gas production and distribution; similarly, electricity production has increasingly been concentrated in very large power stations, although in this case some residual trade-off between economies of scale in production and diseconomies in distribution may still be at work. In consequence, long distance bulk transmission plays an important role in electricity and gas distribution. This is less a feature of water distribution where treatment works tend to be located near the settlements they serve – although bulk supplies *to* treatment works are of some importance.

Application to the transport sector may also not be immediately evident but consider the functional analogy between water distribution systems and roads or railway lines (whether over or under ground); and between treatment works and stations or bus termini. Transport does however raise additional complications, such as that transport itself is part of distribution costs; and that traffic flow consists of units that can exercise some choice about routeing.

Perhaps more encouragingly, the conclusions should be applicable, if distribution costs are replaced by access costs, to Point Type infrastructure (such as hospitals) without undue difficulty. The trade-offs will of course be different, and explicit consideration may need to be given to how transport costs are affected by different scales of operation, but the consequences for access costs if a larger facility requires a larger service area are amenable to analysis using a similar framework to that developed here for water supply.

Taking then water supply, there are two main elements in any urban water supply system: water production (which can be sub-divided into water acquisition and water treatment) and water distribution, each with its own distinctive economic characteristics. These characteristics can be summarised as:

a. Water production

i. Water acquisition

This is highly dependent on the geography and geology of local water resources but typically involves some or all of:

- Impounding dams and reservoirs;
- River abstractions; and
- Boreholes to tap underground water.

The economics of water acquisition reflect these technologies. Dams are clearly large, indivisible items; and an increase in the height of a dam will generally result in a more than proportionate increase in water stored. River abstractions may also enjoy some scale economies due to pumping technology and the volume benefits of larger pipes (the volume of a pipe varies with radius squared, surface area with radius). With boreholes, however, abstraction tends to be optimised with several small ones rather than a few large ones. Nevertheless, overall water acquisition is likely to be characterized by significant scale economies. But there is an important qualification: water has a high weight to value ratio so it quickly becomes uneconomic if pumping is required, either to bring it up from great depths, or to deliver it over long distances where there is insufficient difference in levels to allow gravity feed. There is thus a trade-off between scale economies in water acquisition and transmission costs. Distance introduces diseconomies, a point that will re-appear more strongly when water distribution is considered.

ii. Water treatment

Water taken from boreholes is generally of high quality, needing little further treatment (although there are exceptions to this generalization). Because of this, such treatment as is required can often be provided at or near the wellhead and a separate treatment works may not be required. Where a full treatment works is needed, a near universal requirement for surface water, this is a generally a relatively straightforward semi-industrial facility involving processes such as filtration and chemical treatment. As such, treatment works show the kind of scale economies typical of industrial processes. However, Nick Curtis of Strategic Management Consultants (2002, p. 61) reports that the Minimum Efficient Scale (MES) of water treatment plant is relatively low at about the size required to serve some 50,000 properties (about 30 Ml/day). Unit cost curves estimated by both Curtis and Deloitte, Haskins & Sells (1990) indicate that a doubling of output secures a 20% reduction in costs although it is not clear whether such savings continue much beyond 100,000 properties served. Curtis further reports (p. 30) that “the average size of surface water treatment plant of the five largest water industry companies [in UK] in 1993 … was 44,500 properties.” This may be because in practice the size of treatment works is determined less by the cost-minimising scale of plant than by distribution costs, which we consider next.

b. Water distribution

The water distribution system of any settlement tends to be a reflection of history and local geography rather than technical or economic optimisation, making generalisation difficult. However, modelling – see **Chapters III** and **V**, and **Appendix F** – indicates that unit water distribution costs are likely to increase with size of service area. This is essentially because as the size of the service area increases, the average distance over which water must be delivered increases. However, the modelling also indicates that higher population densities should be associated with lower unit distribution costs, *ceteris paribus*. As a result, the higher costs of distributing to a larger area may be offset to the extent that larger populations are more densely settled[21].

6. The focus of this research

One of the conclusions from the analysis in **Section 4** above is that much of the man-made urban infrastructure can be seen as belonging to one of two broad types:

- **Area-type**: Provides services within a defined area (e.g. utilities, transport systems). In such cases, getting the service to users involves distribution costs;
- **Point-type**: Provides services at a specific point (e.g. hospitals, schools, offices, shops, museums, theatres, etc). In such cases, the equivalent consideration is the cost to users of accessing the facility.

For the former, the cost of supply seems likely to be driven by:

1. Possible scale economies in production (e.g. water treatment works);
2. Possible diseconomies in distribution costs, which may increase more than in proportion to the size of the area served;

[21] As Glaister (1996) has commented: “The [water] industry is likely to exhibit non-constant returns to scale for a variety of reasons. It has long been recognised that the network effects make this the most natural of monopolies. Yet there are likely to be increasing returns to density of supply wherever one has capacity of storage and delivery which depend upon the square of the linear dimensions.”

3. Possible savings in distribution costs related to higher population densities.

For the latter, the equivalent influences are:

1. Any scale economies in the basic facility (e.g. hospital, school, museum);
2. Possible diseconomies in access (e.g. transport) costs, which may increase more than in proportion to the size of the catchment area (cf. the analysis of commuting costs by Arnott (1979));
3. Possible savings in access costs related to higher population densities; and, in addition
4. Possible congestion costs, which are likely to increase with size of catchment area and population density.

It is indeed precisely the interaction between these effects, i.e. economies of scale, distribution costs and density effects, that this research aims to elucidate, using water supply to provide illustration and quantification.

The results of the empirical investigations carried out are reported in **Chapter IV** (water production), **Chapter V** (water distribution) and **Chapter VI** (the interaction between production and distribution). But first, **Chapter III** draws attention to the special issues that arise in considering the distribution stage of water supply, developing simple models which help to throw light on how distribution costs at settlement level can be expected to vary with size of population and service area characteristics.

III. METHODOLOGICAL IMPLICATIONS OF THE DIFFERENT CHARACTERISTICS OF WATER PRODUCTION AND DISTRIBUTION

1. Introduction

Analysis of water supply costs, when the distribution stage is included, raises a number of methodological issues that do not arise when water production is considered on its own. The semi-permanent character of the main assets employed in water supply, particularly on the distribution side, has led some authors to treat capital in this industry as "quasi-fixed". This question is considered in **Section 2** below. Then the potential for interaction between economies of scale in production and diseconomies in distribution (non-separability) gives rise to problems in the specification of production or cost functions to test for scale effects, which are discussed in **Section 3**. Moreover, distribution output has a spatial dimension, raising questions about how it should be measured, and this is tackled in **Section 4**. Here, by modelling distribution areas as monocentric settlements, a measure of distribution output (*DO*) which is the product of water consumption (*QC*) and average distance to properties (φ) is derived. Both this method and alternative models of water distribution costs developed in **Appendix F** lead to the conclusion that distribution output can be viewed as a function of three key variables: consumption per property (*w*), numbers of properties served (*N*) (together making up water consumed, $QC = w.N$) and some measure of the distance or area over which water has to be distributed (in this research, the main emphasis is put on φ but there could be simpler measures related to length of mains or size of service area). Consequently, there will be more than one scale effect to consider. The various possibilities are examined in **Section 5**. A further issue is the treatment of water lost in distribution (leakage); this is taken up in **Section 6**. Finally, in **Section 7** conclusions on how best to proceed are drawn.

In considering the arguments and methods developed in this Chapter**,** and the results of the empirical work carried out on this basis in later chapters, it is important to keep in mind that the purpose of this part of the research is to arrive at a reasonable general representation of scale effects in urban water supply, not to make a precise estimate for a particular company or town.

2. The quasi-fixity of capital

In standard production theory, capital is taken to be fixed in the short term but variable in the longer term. Accordingly, a distinction is made between the short run cost function (in which capital is fixed) and the long run cost function (when it is not). Garcia & Thomas (2001) seem to have been the first to propose that in the cost function for water supply, the capital stock should be treated as "quasi-fixed" because "its modification in the short run is either not feasible or is prohibitively costly" (p.11). In fact, the implication is that the capital stock cannot be changed much even in the longer run, so that it is best to concentrate on results obtained using a short run cost function, conditional on *K*, a vector of quasi-fixed inputs.

Torres & Morrison Paul (2006) concur, arguing that:

> "The choice between long and short run models to represent water utilities' production structure depends on, among other things, the presence of quasi-fixed inputs in the water production and distribution processes. The water utility industry is highly capital intensive, with most of its capital assets underground, which may severely restrict the capital adjustment process. We thus use a short-run cost function framework to represent water utilities' production technology and decisions."

This short run cost function can be expressed as:

$$VC = VC(Q, p, \overline{K}, Z) \qquad \dots\dots\dots\dots\dots \qquad (3.1)$$

Where *VC* is variable costs, *Q* is a vector of outputs, *p* is a vector of variable input prices, $\overline{K}$ is a vector of quasi-fixed inputs and *Z* is a vector of technical/environmental characteristics.

Although Stone & Webster Consultants (2004) estimate both short and long run cost functions for water companies in England & Wales, they also argue (p.14) that:

> "In the water company context, this formulation [assumption of long run adjustment] may be less than helpful … First, the technology used in water services can be indivisible and associated with very long service lives … Secondly, companies do not have total influence over fixed factors such as capital. Legal obligations to meet quality standards or connect customers to network systems means that it can be more appropriate to treat capital in particular as a quasi-fixed input."

Their main results therefore come from a specification based on (3.1). They comment:

> "This variable cost function satisfies the same properties as the long run function, without imposing the assumption that quasi-fixed inputs such as capital have been optimally chosen by the firm. Hence, from an empirical viewpoint, estimation of the variable cost function will yield the same economically

> relevant information contained in the underlying production technology, but without the risk of mis-specification because the level of observed capital inputs have not been optimally determined."

They go on to show that modeling variable costs provides a way of distinguishing between short-run and long-run economies of scale. For example, with output elasticities derived from (3.1), returns to scale (*RTS*) are given by:

$$\textit{Short run}:\ RTS_{SR} = \frac{1}{\varepsilon_S} \text{ where } \varepsilon_S = \frac{\partial(\ln VC)}{\partial(\ln Q)} \qquad \ldots\ldots\ldots\ldots \qquad (3.2)$$

$$\textit{Long run}:\ RTS_{LR} = \frac{1-\varepsilon_K}{\varepsilon_S} \text{ where } \varepsilon_K = \frac{\partial(\ln VC)}{\partial(\ln \overline{K})} \qquad \ldots\ldots\ldots\ldots \qquad (3.3)$$

The arguments for taking this approach in the water industry are strong. On the distribution side, water mains which constitute the vast majority of the assets tend to have very long lives; on the production side, impounding reservoirs are also long-lived while water resources, such as boreholes and river abstractions, cannot be quickly changed. However, not all capital assets are so impervious to change: water treatment works can be expanded or upgraded, pumping stations and monitoring systems can be improved and the formation of new settlements provides opportunities for new technologies to be adopted. Nevertheless, the assumption of "quasi-fixity" is clearly more realistic than assuming complete flexibility. Indeed, in the case of water distribution, we will take this argument a bit further, proposing a Leontief-type production function.

3. The (non-)separability of water production and water distribution

a. The trade-off between the costs of production and the costs of distribution

The tension between economies of scale pulling production to a single point and transport costs pulling production towards the places where customers are located can be seen as lying at the heart of spatial economics[22]. For utilities, the key issue on the distribution side is coming to grips with the implications of Schmalensee's (1978, p.271) observation that: "When services are delivered to customers located at many points, cost must in general depend on the entire distribution of demands over space." The question then is how important in practice is the trade off between economies of scale in production and the costs of distributing the larger volume of product over a larger service area.

[22] See Fujita & Thisse (2002, Ch.2) for a general discussion of location and pricing in a spatial economy.

On this key question, several of the references reviewed in **Appendix B** refer to the possibility of such a trade-off – e.g. Nerlove (1963), Clark & Stevie (1981), Kim & Clark (1988), and Torres & Morrison Paul (2006). However, only Clark & Stevie attempt to investigate this trade-off in a systematic way and their approach is open to criticism as too *ad hoc*. It seems likely that in general there is a trade-off, and that it may be particularly important in the case of water supply because of high distribution costs, but there appears to be plenty of scope for it to be further explored.

It is assumed by Roberts (1986) and Thompson (1997) that electricity production is separable (in the formal economic sense)[23] 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[24]. However, if there are scale economies in the production stage but diseconomies of scale in distribution, this assumption is inappropriate. Transferring attention from electricity to water supply, the point can be simply illustrated by reference to the diagrams in **Figure 3.1** below:

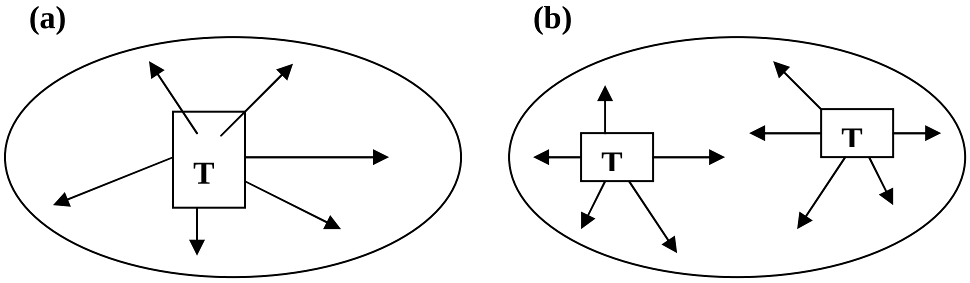

Figure 3.1: Water supply: Should this area be served by (a) one treatment works or (b) two (or more) treatment works?

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

[23] See Chambers (1988) pp.41-48 on separability in production functions and pp.110-119 on separability in cost functions.

[24] A similar assumption is made by Duncombe & Yinger (1993) in their two stage specification of a cost function for fire protection.

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 as it is central to the questions being investigated in this research, this potentially important element of the situation will be missed if one proceeds to try to estimate scale economies in water supply with a cost function specification which assumes separability.

b. Separating distribution from production using production/cost functions

How best then to bring out the distinctive features of water distribution when analyzing water supply costs? Among those using production and/or cost functions, two broad approaches can be identified in the literature:

(a) Model water supply as a single activity but seek to identify distribution effects by treating distribution as an additional output in a multi-output framework or by adding suitable explanatory variables. Thus (Stone & Webster (2004) use number of connections as a measure of distribution output while (Kim & Clark (1988) introduce miles of pipes as an explanatory variable and (Torres & Morrison Paul (2006) introduce service area. It would also be possible to use some composite of these, such as connections/mile of pipe or connections/service area, i.e. measures of density, although this is not done directly in the studies mentioned. The main problem with this approach is that it may fail to expose fully the distinctive economics of the distribution stage.

(b) Develop a two stage model of production and supply, either based on network costs (Clark & Stevie (1981)) or on a two stage production function – e.g. Roberts (1986) and Thompson (1997) for electricity supply, Duncombe & Yinger (1993) for fire protection, with distribution effects being directly identified in the second stage. The main problem here is that multi-collinearity between production and distribution variables will arise unless cost minimization at the first (production) stage is assumed, but that is inappropriate if the two stages are not separable (in the formal economic sense) – see **Section 3(a)** above.

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

The strengths and weaknesses of the first approach can be seen in Torres & Morrison

Paul (2006)[25]. Although their cost function does not distinguish between water treatment and water distribution, volume economies (ε_{CY}) in their analysis can be seen as likely to arise mainly at the treatment stage while economies (or diseconomies) linked to customer numbers (ε_{CN}) or service area (ε_{CS}) are likely to 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, as they assert (p. 105), "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.

There is also a problem regarding Torres & Morrison Paul's measurement of the effect of size of service area. Although they considered including length of pipes in the vector of quasi-fixed inputs, they 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.111, 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 their short run specification of the production/cost function has adequately represented differences in the capital invested in systems of different sizes and densities.

On the face of it, some of the problems identified above might be avoided, if suitable data is available, by estimating a production function which includes all the separate inputs to production and distribution in a single function, such as:

$$Q = f(K_P, L_P, Z_P, K_D, L_D, Z_D) \qquad \ldots\ldots\ldots\ldots \qquad (3.4)$$

Where *Q* is final output, and *K*, *L* and *Z* are (vectors of) capital, labour (and other operating) inputs and environmental factors, relating to production (*P*) and distribution (*D*) respectively.

[25] See **Appendix B, section 4(e)** for a fuller account.

This is the approach taken initially by Roberts (1986) and Thompson (1997) for electricity supply[26]. Apart from data issues – e.g. implementation will require suitably disaggregated data, with the added complication that production units may not align with distribution areas – there is also the likelihood of unacceptably high collinearity between variables. In fact, if a cost function is derived from the composite production function (3.4), 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. It is probably for this reason that Roberts and Thompson, in their cost functions, replace the production variables by a single price for electricity input into transmission and distribution, arguing that there are in effect constant returns to scale in electricity production, contrary to at least some of the evidence, e.g. Nerlove (1963).

An alternative to either of the above procedures would be to work with a separate production or cost function for each stage of water supply (although no studies which do this have come to light in our literature survey). This route, while feasible, is also not without problems, as explained below.

- *Production stage*

Following standard procedure, one would start by postulating a production function for water production[27] of the general form:

$$QP = f(K_P, L_P, Z_P) \qquad \text{……….} \qquad (3.5)$$

where QP is quantity of water produced, K_P is (a vector of) capital employed in water production, L_P is (a vector of) production operating costs and Z_P 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:

$$CP = C(QP, p_{KP}, p_{LP}, Z_P) \qquad \text{…………..} \qquad (3.6)$$

Where CP is the full cost of water production and the ps are prices related to K_P and L_P. Or, if capital is taken to be quasi-fixed (see **Section 2** above):

$$VCP = C(QP, \overline{K_P}, p_{LP}, Z_P) \qquad \text{……….} \qquad (3.7)$$

[26] See **Appendix B, sections 3(b)** and **(c)**.

[27] The term “water production” here and elsewhere means water acquisition and treatment.

Where *VCP* is variable costs of water production and $\overline{K_P}$ is a measure of the quasi-fixed capital.

There does not appear to be any reason why this method should not be successfully applied to water production, as it was to electricity production by Nerlove (1963), although some practical problems will need to be addressed. For example, the specification strictly relates to individual plants so ideally implementation requires plant level data. 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, although more data is available at company level, most water companies are rather large, and operate large numbers of plants, with very limited plant level data publicly available (not including information on costs at plant level).

- *Distribution stage*

Following the same approach as for water production, one might postulate a production function for distribution having the general form:

$$DO = f(K_D, L_D, Z_D) \qquad \ldots\ldots\ldots\ldots\ldots.. \qquad (3.8)$$

Where *DO* is a measure of distribution output and K_D, etc are the distribution equivalents of the treatment variables – see (3.5) above. It would then in principle be possible to proceed to derive a distribution cost function of the general form:

$$CD = C(DO, p_{KD}, p_{LD}, Z_D) \qquad \ldots\ldots\ldots \qquad (3.9)$$

Where *CD* is the full cost of distribution and the *p*s are prices related to K_D and L_D. Or, if capital is taken to be quasi-fixed

$$VCD = C(DO, \overline{K_D}, p_{LD}, Z_D) \qquad \ldots\ldots\ldots \qquad (3.10)$$

Where *VCD* is the variable costs of distribution.

However, the processes involved in distribution are rather different in character from those involved in production. By far and away the largest capital input to water distribution is the network of pipes through which the water is delivered to customers. This basic system may be augmented by service reservoirs (to help manage fluctuations in demand), pumping stations (to boost pressures) and, in some countries, water towers (which serve both purposes); and the system may be subject to a greater or lesser degree of monitoring, which may be more or less automated. Operating costs include teams to carry out inspections and repairs, pumping costs and leakage control. Thus although there is some scope to vary the proportion of capital to other inputs, so that a production function for distribution can be said to exist, in practice the network of pipes is more or less fixed and there is very little choice of technology so that significant change in input factor intensity is unlikely to be observed even in the longer term.

Moreover, within any one company, there will be little variation in factor prices from one area to another, so that (whether or not there is much choice of technology) economic considerations would lead one to expect more or less the same technology to be adopted throughout the company area. The only real variable is the scale of output and that will be determined by the size and location of customer demands in relation to the water production facilities. In this case therefore it would appear an acceptable simplification to consider the ‘production function’ to have become reduced to a single point for each level of output, with factor proportions fixed by the technology that has been chosen (or, more often, inherited from the past). This is the Leontief form of production function[28] but without constant returns to scale and is portrayed in **Figure 3.2**

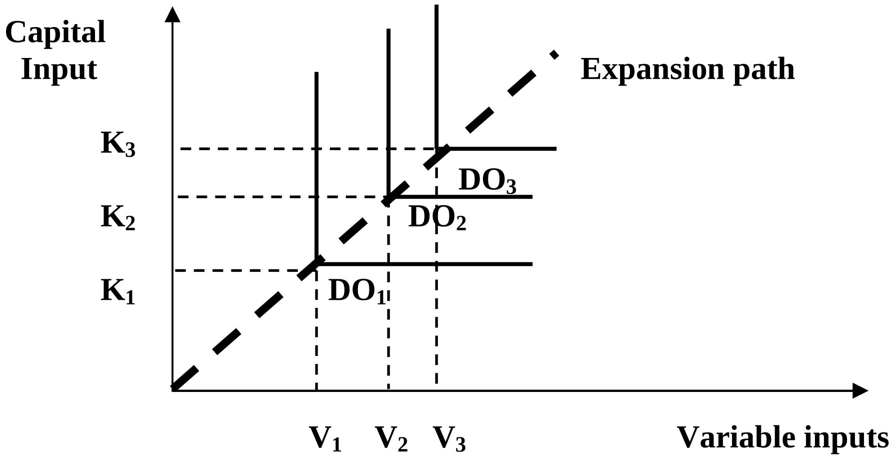

Figure 3.2: Production functions for water distribution

The implication of **Figure 3.2** is that there is a particular amount of variable input associated with any particular level of output, i.e:

$$VC = V(DO) \qquad \ldots\ldots\ldots \qquad (3.11)$$

If *VC* is measured as the variable costs of distribution (*VCD*) this becomes (3.10) shorn of the additional variables on the RHS, although it would still be right to include any technical/environmental variables from Z_D which might affect this relationship. And for

[28] The Leontief production function can be regarded as a special case of the CES production function, when the coefficient of substitution $\gamma = -\infty$, although this not particularly helpful.

any particular level of output and variable input, there will be an associated amount of capital input, which is why a capital variable is not needed in (3.11)[29].

A note on Leontief production functions

If technology is such that *Q* units of output require $u.Q^{\alpha}$ units of fixed capital input and $v.Q^{\beta}$ units of variable inputs, three distinct cases arise:

1. α = β = 1: This is the textbook Leontief production function, which has the two properties: (a) K/V = u/v (i.e. a constant); and (b) $\frac{\partial(\ln K)}{\partial(\ln Q)} = \frac{\partial(\ln V)}{\partial(\ln Q)} = 1$; i.e. constant returns to scale.
2. α = β = γ (γ≠1): This can be called a Leontief-type production function. It has the two properties: (a) K/V = u/v (i.e. a constant); and (b) $\frac{\partial(\ln K)}{\partial(\ln Q)} = \frac{\partial(\ln V)}{\partial(\ln Q)} = \gamma$, i.e. increasing or decreasing returns to scale depending whether γ is <1 or >1.
3. α ≠ β: This is a new case, which does not seem to be discussed in the literature. It has the properties: (a) $\frac{K}{V} = \frac{u}{v} Q^{\alpha-\beta}$ (i.e. varies with the level of output); and (b) returns to scale also varies with output, being a function of $\frac{\partial(\ln K)}{\partial(\ln Q)}$ and $\frac{\partial(\ln V)}{\partial(\ln Q)}$.

As with water production, there will be a number of practical problems to address:

- Just as the production function for water production 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 distribution system is not easily accessible.
- Secondly, there is a question about how distribution output (*DO*) should be measured. Volume of water is 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”. In **Section 3** below, a new composite measure is proposed, which incorporates both volume and distance.

[29] It is in this respect that the approach here differs from the ‘quasi-fixed’ capital approach of Garcia & Thomas (2001).

- Thirdly, there is the question of how to deal with leakage which, in UK at least, is significant, varying between about 10% and 30% across companies. This is discussed in **Section 6** below.

c. Proposed way forward

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) using production or cost functions`:

i. A possible starting point is a *composite production function* like (3.4) above, provided appropriate data is available, and there is not excessive collinearity between variables. However, it would not be possible to estimate a cost function based on this production function because of collinearity in the prices.

ii. A better prospect would be to start from the *separate production functions* for water production (3.5) and water distribution (3.8) which, assuming capital to be quasi-fixed, then lead, as discussed above, to the variable cost functions (3.7) for water production and (3.11) for water distribution. There would still be a number of practical issues to resolve, as noted above; and some way of bringing the separate results together will be needed.

iii. If neither of the above approaches can be successfully implemented, the *aggregate cost function* used by Torres & Morrison Paul (2006), perhaps with different distribution variables, remains a possibility although it may not fully expose the different economics of production and distribution.

Method (i) above seems to be ruled out because direct estimation of the production function is unlikely to work well while the cost function cannot be estimated because of collinearity in prices. Although method (ii) might appear to ignore non-separability, this is not in fact the case. There is clearly no objection to estimating economies of scale in production at plant level, if suitable data is available, as done by Nerlove (1963) and those following in his footsteps. Similarly, scale effects in distribution can be investigated independently of production. However, to determine the cost-minimising arrangement taking production and distribution together will then require a sort of “trial and error” assessment of different combinations of treatment plants and service areas – rather in the manner of Clark & Stevie (1981). So this is a viable if somewhat clumsy approach. Finally, method (iii) is also feasible, given suitable data, and results obtained in this way can then be compared with those obtained using method (ii) to see whether different conclusions emerge.

4. Defining distribution output

At first sight, it might seem that the output of the distribution system is simply the volume of water delivered. If it were all delivered to one place, this might be acceptable. But the essence of the distribution function is to deliver water to many different places, in the amounts and at the times when it is required[30]. These wider functions need somehow to be reflected in the way output is measured.

As a starting point, distribution output might be measured, by analogy with measures such as tonne-kms and passenger-miles used in transport studies, as:

$$DO = \sum_{i=1}^{N} w_i . r_i \qquad \ldots\ldots\ldots\ldots \qquad (3.12)$$

where: N = number of properties being supplied;

w_i = water consumption by property i;

r_i = distance of property i from water treatment plant.

In this formulation, the quantity of water used at each property is weighted by the distance it has had to be transported to reach the property. It therefore leaves out some other features of distribution such as height (pumping head) and variations in diurnal and seasonal demand[31]. Nor does it say anything about the technology of distribution although it leaves scope for the efficiency of distribution to vary depending on the technology used (e.g. size of pipes, use of booster pumps, number of service reservoirs, etc.).

In practice, information about the consumption and location of every individual property is unlikely to be available so it will be necessary to work with average consumption per property, or averages for groups of consumers such as households and non-households (or large industrial consumers and others) and to find ways of

[30] This point is well-recognised in logistics: “Logistics … is the positioning of resource at the right time, in the right place, at the right cost, at the right quality.” (Rushton *et al* (2000)). More generally, the functions of the distribution system can be summarised as making product available *where* and *when* it is required, as well as in the *quantity* demanded, i.e. it involves changing location (transport) and timing (storage) as well delivery to individual consumers (breaking bulk). With high value products, the value of the product in the pipeline can also be an important consideration.

[31] Arguably, if water supply was priced in a competitive market, no such adjustments are needed as the price paid by consumers should reflect all these factors. However, in the case of water supply, prices are often not market-determined and it is necessary to work with quantities supplied rather than value.

approximating distances. The simplest assumption would be that each property has the same water consumption, *w*, which is equal to total consumption averaged over all properties. Adopting this simple assumption, (3.12) can then be expressed as:

$$DO = N.w.\varphi \qquad \ldots\ldots\ldots\ldots\ldots \qquad (3.13)$$

where *φ* is the average distance between properties and the treatment plant. *N* and *w* (or *N.w* = total consumption, *QC*) are usually readily available, but how to estimate *φ*?

Although in practice a water treatment works may serve more than one settlement, or a large settlement may be served by more than one works, it is convenient to start by supposing that each treatment works serves a settlement proportional in size to the capacity of the works. Then, treating each settlement as circular and monocentric, with its treatment works centrally located[32], the following results can be used:

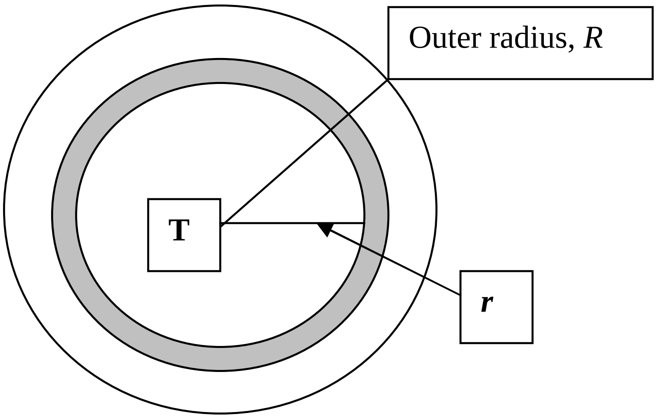

Figure 3.3: Circular settlement

In **Figure 3.3** if property density at radius *r* = *d(r)*, and the width of the shaded area is *δr*, then

Number of properties in the shaded ring, $n_r = d(r).2\pi r.\delta r$ ………… (3.14)

Distance to properties in the shaded ring, $\psi_r = r.n_r = r.d(r).2\pi r.\delta r$ ………… (3.15)

and

Total number of properties in the settlement, $N = 2\pi\int_0^R d(r).r.dr$ ……….. (3.16)

Total distance to properties in the settlement, $\psi = 2\pi\int_0^R d(r).r^2.dr$ ………….. (3.17)

So, average distance to properties in the settlement, $\varphi = \dfrac{\psi}{N}$ ………….. (3.18)

[32] More commonly, the treatment works will be towards the edge of a settlement but the exact location is relatively unimportant if water is delivered in bulk to the distribution system.

If it is further supposed that property density is uniform across the settlement[33], so that *d(r) = d*, (3.18) then yields:

$$\varphi = \frac{2}{3}R \qquad \text{where } R \text{ is the settlement radius} \qquad \ldots\ldots\ldots\ldots\ldots \quad (3.19)$$

In this case therefore *DO* is linear in *R*.

If, more realistically, and consistently with the monocentric urban model (see, for example, DiPasquale & Wheaton (1996, pp.61-64)), a declining density is assumed, so that $d(r) = d_0.e^{-\lambda r}$ (i.e. density declining exponentially at the rate λ away from the centre, where density is d_0), then (3.16) gives:

$$N = \frac{2\pi d_0}{\lambda^2}\left[1 - e^{-\lambda R}(1+\lambda R)\right] \qquad \ldots\ldots\ldots\ldots\ldots \quad (3.20)$$

And (3.17) gives:

$$\psi = \frac{4\pi d_0}{\lambda^3}\left[1 - e^{-\lambda R}\left(1+\lambda R + \frac{\lambda^2 R^2}{2}\right)\right] \qquad \ldots\ldots\ldots\ldots\ldots \quad (3.21)$$

And (3.18) then becomes:

$$\varphi = \frac{\psi}{N} = \frac{2}{\lambda}\frac{\left[1 - e^{-\lambda R}\left(1+\lambda R + \frac{\lambda^2 R^2}{2}\right)\right]}{\left[1 - e^{-\lambda R}\left(1+\lambda R\right)\right]} \qquad \ldots\ldots\ldots\ldots\ldots\ldots \quad (3.22)$$

Here, *φ* is increasing in *R* (in a non-linear way) and so therefore is distribution output. Although in both cases average distance is a function of *R*, this does not mean that density has no effect on costs. For a settlement of a given population size, *R* will vary inversely with density. This can also be seen by noting that the expression (3.13) for *DO* includes *N*, the number of properties, which may be larger or smaller independently of *φ*.

Now, a measure of the distribution output of a settlement can be obtained as the product of *N*, *w* and *φ*. In the constant density case, using (3.19) and (3.13), this gives:

$$DO = N.w.\frac{2}{3}R \qquad \ldots\ldots\ldots\ldots \quad (3.23)$$

Similarly, in the more realistic declining density case, using (3.22) and (3.13), it leads to:

[33] An assumption often adopted for simplicity although inconsistent with standard urban theory which suggests that density will decline away from the centre.

$$DO = \frac{2}{\lambda} w.N \frac{\left[1 - e^{-\lambda R}\left(1 + \lambda R + \frac{\lambda^2 R^2}{2}\right)\right]}{\left[1 - e^{-\lambda R}(1 + \lambda R)\right]} \qquad \ldots\ldots\ldots\ldots\ldots \qquad (3.24)$$

This is a rather more complicated expression than (3.23) and its evaluation requires an estimate of (or a plausible assumption for) λ. If desired (3.23) and (3.24) could be expressed as functions of service area *A* rather than *R*, using $A = \pi R^2$. However, because reported service areas often include areas which are unoccupied or unserviced, it is likely to be desirable to make a further refinement to exclude areas not reached by water mains, when measuring *A* or *R*.

Hitherto, studies of water supply have always measured output as the quantity of water supplied or consumed, so failing to take into account the distance aspect of water distribution. These new measures, although approximations, are clearly superior in this respect: as can be seen, in both cases *DO* is the product of quantity consumed (*w.N*) and a measure of average distance to properties (φ). In **Chapter V**, methods to implement (3.24) are developed and the results of using this measure to estimate scale effects in distribution are reported.

Using a different kind of model also leads to the conclusion that water distribution costs are driven by three key variables: consumption per property (*w*), number of properties (*N*) and length of mains (*M*). This alternative approach is set out in **Appendix F**.

5. Assessing scale effects in water production and distribution

Now, if from (3.11), the cost function for water distribution is of the form:

$$VCD = f(DO, Z_D) \qquad \ldots\ldots\ldots \qquad (3.25)$$

And from (3.23) or (3.24):

$$DO = f(w, N, \varphi) \qquad \ldots\ldots\ldots \qquad (3.26)$$

It can be seen that there is more than one cost elasticity to consider when assessing scale effects. Three are of particular interest:

a. ε_W – the elasticity of distribution cost with respect to consumption per property – the pure quantity effect (numbers of properties and size of distribution area held constant);

b. ε_N – the elasticity of distribution cost with respect to numbers of properties – the density effect (consumption per property and size of distribution area held constant);

c. ε_A – the elasticity of distribution cost with respect to size of service area – the size of distribution area effect (which is also a kind of density effect).

In deriving values for these elasticities from estimating equations based on (3.26), it needs to be noted that both N and φ are functions of λ and R, and are therefore not independent of each other. The elasticities ε_N and ε_A cannot be read off from the estimated coefficients. Their evaluation is taken up for further discussion in **Chapter V**.

Other elasticities potentially of interest include:

d. $\varepsilon_S = \varepsilon_N + \varepsilon_A$ – the elasticity of distribution cost with respect to size of settlement (density constant);

e. ε_D – the elasticity of distribution cost with respect to density of settlement.

Returns to scale are then measured by the reciprocal of these elasticities, with values greater than 1 indicating economies of scale (greater than 0 if quantity does not change, as with ε_A and ε_D). And, following Stone & Webster (2004), the relationships (3.2) and (3.3) provide a means to distinguish between short term and long term returns to scale, in cases where the cost function includes a term for capital.

6. Treatment of leakage

Thus far, the treatment of leakage (distribution losses) has not been considered. In fact, leakage rates are typically in the range 10%-30%[34]. Leakage thus represents a significant cost in the water supply process. Not only does water that has been acquired and treated at some cost get lost but part of the costs of distribution are incurred in the transport of water that never reaches consumers.

One approach, pioneered by Garcia & Thomas (2001), is to treat leakage as an additional output (albeit an undesirable one) in a multi-product analysis. Hence in their cost function (see **Appendix B, section 4(b)**) the output vector has two components – water delivered to customers and water lost in distribution. This approach, also used by Stone & Webster (2004), is attractive when the focus is on industry efficiency because it enables the trade off between higher expenditure on leakage control and expenditure on other ways of increasing supplies to be exposed. However, the focus in this research is

[34] See **Appendix D** for a full breakdown of distribution losses (incl. leakage) for one water company.

different. We want to be sure that if there are any systematic differences between leakage costs related to settlement size or density, it will be reflected in our results. For this purpose, it is sufficient that water production costs include the cost of producing amounts lost to leakage, while for water distribution, the recorded costs should include the cost of leakage control activities as well as the cost of transporting lost water (as they do). If a measure of the cost of leakage is required, the difference between unit costs with quantity consumed (*QC*) as the divisor and unit costs with quantity put into distribution (*QDI*) as the divisor will provide it.

7. Data sources

a. Ofwat data for water companies in England & Wales

Water supply in England and Wales is currently the responsibility of 10 combined water and sewerage companies (WaSCs) and 12 water supply only companies (WoCs)[35]. In the areas where the latter supply water, sewerage is the responsibility of one of the combined water and sewerage companies. Whereas the WaSCs cover very large areas, based in principle (following a reorganization of the industry in 1973) on river basins, the WoCs generally cover rather smaller areas, reflecting their origins as municipal water suppliers (although with the passage of time, some have come to serve more than one urban area).

As the ultimate purpose of this research is to throw light on how infrastructure affects the economics of urban settlements, the ideal would be to test the relationships developed in **Chapters III**, **IV** and **V** using data from individual urban areas. Data disaggregated to urban area level on the water supply activities of the WaSCs is not publicly available. For the WoCs there is, at least in some cases, a closer match between responsibilities and particular urban areas (e.g. Bristol, Cambridge, Portsmouth). However even in these cases the correspondence with urban areas, as defined for other purposes, e.g. Census key statistics for urban areas (ONS (2004)), local authority administrative boundaries or the Functional Urban Regions favoured by some researchers, is not very good; and in other cases (e.g. Three Valleys, South East Water), the correspondence appeared to be quite remote.

[35] Omitting the Cholderton & District Water Co, for which Ofwat does not publish data because it is too small.

Tables 3.1A and **3.1B** below show the key water supply figures for 2003 for the WoCs and the WaSCs respectively, as recorded in Ofwat (2003a):

Company	**Acronym**	**Area[36] (sq km)**	**Properties served ('000)**	**Treatment plants (No)**	**Water supplied (Ml/day)[37]**
Bournemouth & West Hampshire Water plc	BWH	1,041	188	7	160
Bristol Water plc	BRL	2,391	483	23	292
Cambridge Water plc	CAM	1,175	120	14	73
Dee Valley Water plc	DVW	831	117	9	70
Folkestone & Dover Water Services Ltd	FLK	420	72	18	50
Mid-Kent Water plc	MKT	2,050	242	29	157
Portsmouth Water Ltd	PRT	868	290	20	177
South East Water plc	MSE	3,607	590	65	376
South Staffordshire Water plc	SST	1,507	548	29	331
Sutton & East Surrey Water plc	SES	833	270	11	160
Tendring Hundred Water Services Ltd	THD	352	70	2	30
Three Valleys Water plc	TVW	3,727	1,224	99	864

Table 3.1A: Water only companies (England & Wales, 2003)

Company	**Acronym**	**Area[36] (sq km)**	**Properties served ('000)**	**Treatment plants (No)**	**Water supplied (Ml/day)[37]**
Anglian Water Services Ltd (incl. Hartlepool)	ANH	22,090	1,930	143	1,159
Welsh Water (Dwr Cymru)	WSH	20,400	1,317	105	883
Yorkshire Water Services Ltd (incl. York)	YKY	14,240	2,109	90	1,299
Northumbrian Water (incl Essex & Suffolk Water)	NES	11,843	1,899	67	736
South West Water Ltd	SWT	10,300	726	40	447
Severn Trent plc	SVT	19,745	3,279	173	1,958
Southern Water	SRN	4,450	1,007	102	595
Thames Water	TMS	8,200	3,474	99	2,804
United Utilities (NW Water)	NWT	14,415	3,120	137	1,952
Wessex Water Services Ltd	WSX	7,350	537	119	368

Table 3.1B: Water and sewerage companies (England & Wales, 2003)

[36] Figures for water company area (in sq. km) are from Ofwat (2003, Appendix B5, p.94) but note that these are company, not Ofwat, estimates.

[37] Water production (and works capacity) is usually quoted in Megalitres per day (Ml/d);
1 Ml = 1,000,000 litres

Each year, all the water companies submit to Ofwat in a standard format (known as the "June Return") a large amount of data, both financial and non-financial, for regulatory purposes. This process has been in operation since 1992. Most of this data (omitting only a small amount judged to be commercially confidential) is publicly available on the Ofwat website or on CD-ROMs. The data is used by Ofwat to inform its regulatory activities; and analyses using appropriate parts of the data are included in Ofwat publications, notably (in the present context) an annual report on "Water and sewerage service unit costs and relative efficiency" (e.g. Ofwat (2004)). As noted in **Appendix B, section 5(b)**, a key difference between the Ofwat analyses and those reported here is that Ofwat's focus is on differences in the relative efficiency of companies, after allowing for differences in their operating environments, whereas our emphasis is precisely on how environmental factors (such as differences in population densities and the size of areas served) affect costs, at settlement rather than company level. Hence this research looks at the data from a different perspective.

Data for each of the years 1998-2003 was extracted for all the reporting companies from the Ofwat June Returns. During this period the number of WoCs declined from 17 to 12, owing to amalgamations and absorption into WaSCs. The original intention had been to carry out analysis using this panel data. However, in addition to the problem of the changing number of companies (which can largely be overcome), it was found that the year to year variation in key quantities was rather small and random so that when working in differences (as panel methods do), the results obtained were very poor[38]. Therefore, analysis was carried out primarily using cross-section data for 2003. **Appendix A** explains in detail how the data has been compiled, giving for each item the June Return (JR) Table number and line reference.

b. AWWA data for water undertakings in the USA

The water industry in the US is highly fragmented. The USEPA in 1993 recorded nearly 60,000 water systems. However, over 60% of these were classified as "very small", serving populations of less than 500. Larger systems mostly belong to members of the American Water Works Association (AWWA) and the AWWA has carried out a

[38] In a different context, Lundberg & Squire (2003) observe that "Within cross-sectional data, all unobserved cross-country variation is relegated to an error term … Panel-data formations make it possible to control for the unobserved cross-country effects … However, inequality varies much more across countries than over time, and the characteristics of this variance cannot be examined by techniques that eliminate cross-country effects and focus exclusively on the within-country relationships …"

number of surveys of its members in recent years, which provide a rich source of data for research.

Table 3.2 below sets out a comparison between the size distribution of utilities responding to the 1996 AWWA survey and the USEPA data on systems. Overall, the AWWA figures appear to cover about 40% of the population in the USEPA analysis. It is clear that the systems included in the AWWA figures are on average larger than those recorded by the USEPA, even within size groups. In part this may be because the AWWA respondents are utilities, some of which may operate more than one system, particularly in the case of the larger utilities (population growth and amalgamations of water utilities between 1993 and 1996 could also provide part of the explanation).

USEPA designation	**Population served**	**USEPA (1993)**		**AWWA (1996)**	
		No of systems	**Population served (million)**	**No of utilities**	**Population served (million)**
Very small	25-500	36,515 (62%)	5.569 (2%)	- (0%)	- (0%)
Small	501-3300	14,516 (25%)	20.053 (8%)	3 (0%)	0.003 (0%)
Medium	3301-10,000	4,251 (7%)	24.729 (10%)	14 (0.03%)	0.135 (1.4%)
Large	10,001-100,000	3,062 (5%)	85.035 (35%)	358 (66.9%)	13.845 (14.1%)
Very large	> 100,000	326 (1%)	109.797 (45%)	161 (29.9%)	83.981 (85.7%)
Total		**58,670**	**245.183**	**538**	**97.964**

(USEPA data from Twort et al (2000), Table 2.1)

Table 3.2: Comparison of size distribution of US water utilities

For this research, the AWWA's Data Manager provided a disk containing the results of the 1996 survey (which was also the source of data for Torres & Morrison Paul (2006)). Information was extracted from three of the tables on the disk:

- **Utility general information**: This table provided the name, city and state of each water utility, together with the retail and wholesale population served, the size of the service area (sq. miles), and the volume of water produced, subdivided into ground water, surface water and purchased water (all in million US gallons/year).

- **Annual Operation & Maintenance expenses**: This table provided total operating and maintenance expenditure, subdivided into supply, water treatment, distribution, customer accounts, administration and other; and also employee numbers.
- **Plant ID table**: This was used to infer the number of water treatment plants operated by utilities which process surface water (groundwater systems are treated as a single system).

c. Data for one company ("BWC") in England & Wales

The focus of this research is on the economics of infrastructure at settlement level. A problem with taking water supply in England & Wales as a case study is that the water companies mostly cover rather large areas, serving many urban settlements. This makes it difficult using company level data to discern clearly what is happening at this lower level. Fortunately, one of the larger water companies (which does not wish to be identified) kindly agreed to provide a considerable amount of disaggregated information on a confidential basis for the purposes of this research. This has proved extremely useful in throwing more light on the questions of interest than is possible with company level data.

The company concerned, which we shall for convenience refer to as "BWC" (Britannia Water Co), is fairly typical of the larger WaSCs in terms of size of service area, numbers of customers, mix of urban and rural areas, sources of water and types of treatment plants operated. While not strictly "representative" in the statistical sense, observations based on its experience can be taken as providing a picture that is not seriously misleading. Fuller discussion of the information provided by BWC and how it was processed can be found in **Appendix H**.

8. Conclusions

A number of conclusions have emerged about the appropriate methodologies to use when the aim, as here, is to estimate scale effects in water supply at settlement level.

- First, the quasi-fixed character of much of the capital invested in the water industry justifies the use of variable cost models, with capital treated in effect as a control variable. Indeed, in the case of water distribution, the lack of much choice of technology justifies the adoption of a Leontief-type production function, when no capital term is required.

- Secondly, the non-separability of water production and water distribution means that treating water supply as a single activity risks obscuring the distinctive characteristics of water distribution. Equally, it may not be valid to assume cost minimization at the production stage if (as is likely) there is interaction with distribution costs. There is merit therefore in examining water production and water supply separately, even if this means a somewhat clumsy procedure to analyse their interaction.
- Thirdly, the measurement of distribution output needs to capture in some way the spatial aspect of distribution. In **Section 4**, a measure of distribution output (*DO*) as the product of quantity consumed ($QC = w.N$) and the average distance to properties (φ) is developed. Conceptually, this is similar to the use of tonne-kms or passenger-miles in transport studies. Implementation of this measure is left to **Chapter V** but it will be evident there that it offers useful insights. In fact, it might sometimes prove useful in studies of utilities other than water, when distribution as well as production are under consideration.
- It follows that assessment of scale effects will require more than one elasticity to be considered. Of particular interest are likely to be:
  a. ε_W – the elasticity of distribution cost with respect to consumption per property;
  b. ε_N – the elasticity of distribution cost with respect to numbers of properties;
  c. ε_A – the elasticity of distribution cost with respect to size of distribution area.

Development of specifications to implement these conclusions will be taken up in the chapters that follow. Based on **Section 3(c)** above, the basic strategy will be to separately estimate cost functions for water production (**Chapter IV**) and water distribution (**Chapter V**); then to use the results (**Chapter VI**) to assess the interaction between them, and to compare with estimates obtained by other researchers using cost functions which incorporate both production and distribution.

IV. ECONOMIES OF SCALE IN WATER PRODUCTION: EMPIRICAL INVESTIGATION

1. Framework for investigation

a. Cost functions: General considerations

The aim of the empirical work reported in this chapter and in **Chapters V** and **VI** is to use data on water supply to throw light on the interaction between economies of scale, distribution costs and density effects at settlement level. The methodologies used build on the approaches found in the literature surveyed in **Appendix B**. However, the objectives of this research are different from those in the mainstream utilities literature in that the focus is on settlement (not company) level effects; and what can be done is limited to some extent by the availability of suitable data so that some compromises have had to be made. For example, while the most complete data available is that provided by the June Returns to Ofwat, most companies in England & Wales serve a large number of settlements so that this data does not directly reveal settlement level effects. It has required some ingenuity to adapt the methodologies and manipulate the data to produce results which, it is hoped, provide a plausible assessment of the likely size of the effects of interest. Per contra, while use of the results to model urban water supply seems reasonable, their use to assess the performance of individual companies would be inappropriate.

Generally, the literature points to the use of cost functions as the way into assessing scale effects. For any production activity, it can be supposed that there exists a *production function*, which expresses the conversion of inputs into outputs:

$$Q = Q(L, K, Z) \qquad \ldots\ldots\ldots\ldots \qquad (4.1)$$

Where *Q* represents output, *L* represents variable inputs, *K* represents capital inputs and *Z* represents external factors which may affect the relationship. *Q, L, K* and *Z* may be vectors with several elements each. The "=" sign implies that production is at the efficient frontier of the production set.

The cost of producing the output *Q* can be expressed as $C_Q = p_L.L + p_K.K$, where p_L and p_K are the prices applicable to *L* and *K* respectively. Then, assuming cost minimisation subject to the production function constraint, this leads to the long run *cost function*:

$$C_Q = C_Q(Q, p_L, p_K, Z) \qquad \ldots\ldots\ldots\ldots \qquad (4.2)$$

(together with a set of cost share equations, one for each input).

This formulation assumes that firms are able to adjust their capital inputs optimally. However, in the case of water supply, many of the capital inputs are very long-lived (e.g. reservoirs, water mains) and cannot be quickly adjusted. Following Garcia & Thomas (2001) and others, it is arguably more realistic to treat such inputs as ‘quasi-fixed’. This leads to a modified, ‘short run’, cost function:

$$VC_Q = VC_Q(Q, p_L, \overline{K}, Z) \qquad \ldots\ldots\ldots\ldots \qquad (4.3)$$

Where VC_Q is the variable cost associated with the output level Q and $\overline{K}$ is a measure of the ‘quasi-fixed’ capital inputs. In this formulation, the term $\overline{K}$ becomes in effect a component of Z, one of the conditioning variables.

If this latter approach is adopted, a particular issue arising is how capital maintenance fits into this framework. In the Ofwat data, capital maintenance for infrastructure assets[39] is “the annual expenditure required to maintain the operating capability of the existing network”, while for non-infrastructure assets it is the CCA depreciation charge. Should this be treated as part of variable costs? The practical arguments for doing so appear strong: water supply is a highly capital intensive activity so that leaving out capital maintenance would omit about half of the costs charged to water companies accounts; and the distinction between current maintenance (which is included in operating costs) and capital maintenance is somewhat arbitrary. However, this would not be consistent with the theoretical reasoning which leads to (4.3). If the capital input $\overline{K}$ is fixed (by the assumption of quasi-fixity) then the amount of capital maintenance is pre-determined and not a quantity that can be optimised. Therefore capital maintenance is not included in specification (4.3). (On the other hand, the correct treatment if using specification (4.2) would be that adopted by Stone & Webster (2004, p.33-4) where the price p_K includes both the return on capital (τ in Stone & Webster’s notation) and

[39] The definitions of “infrastructure assets” and “operational assets”, indicate that the former include some assets related to water acquisition (e.g. dams and reservoirs) although the majority relate to distribution (e.g. water mains), while the latter relate almost entirely to water acquisition and treatment: “**Infrastructure assets** cover the following: underground systems of mains and sewers, impounding and pumped raw storage reservoirs, dams, sludge pipelines and sea outfalls.”
“**Operational assets** cover the following: intake works, pumping stations, treatment works, boreholes, operational land, offices, depots, workshops, etc …”

capital maintenance/depreciation (δ in Stone & Webster's notation), both expressed as a proportion of the capital stock.)

If the case for treating capital as quasi-fixed is accepted, the Stone & Webster report argues that returns to scale can be assessed working with just the variable cost model (4.3), because a measure of long run returns to scale can be obtained using the relationships:

$$\text{Short term returns to scale, } RTS_S = \frac{1}{\varepsilon_S} \text{ , where } \varepsilon_S = \frac{\partial(\ln VC)}{\partial(\ln Q)}$$

and $$\text{Long term returns to scale, } RTS_L = \frac{1-\varepsilon_K}{\varepsilon_S}\text{, where } \varepsilon_K = \frac{\partial(\ln VC)}{\partial(\ln \overline{K})}\,.$$

This has provided the starting point for our investigation of returns to scale in water production and water distribution. Where the data permits, the long run model (4.2) has also been deployed. However, as will be explained in **Chapter V**, it was found preferable to adopt a different approach for water distribution, deriving from a Leontief-type production function.

Moving on to consider implementation in more detail, there are three steps to address.

b. Cost function for water production

Based on (4.3) above, the starting specification proposed is:

$$VCP = VC_P(QP, p_{LP},.\overline{K_P}, Z_P) \qquad \ldots\ldots\ldots \qquad (4.4)$$

Where the *P* subscript signifies quantities related to water acquisition and treatment (hereafter called 'water production'). *VCP* should therefore include the variable costs of both water acquisition and treatment. In the Ofwat data acquisition and treatment are not distinguished; in the AWWA data they are separately recorded but can easily be combined; in the BWC data, some elements of operating costs had to be allocated to achieve the same coverage as the Ofwat figures. *QP* should be quantity of water actually treated (so excluding any imported or purchased water that has already been treated). The cost of imported/purchased water should therefore only be included in *VCP* if it is untreated.

In applying (4.4) to BWC zones, it is reasonable to assume that the variation between cases in p_{LP} is sufficiently small to be ignored. For simplicity, the same assumption is adopted for companies reporting to Ofwat and for US utilities, although this is more questionable. Additional arguments for this simplification are that technology is fairly standard in the water industry and the scope for capital/labour substitution does not appear to be large; also the assumption that capital is quasi-fixed implies that such substitution does not take place in the short run.

Using Ofwat data, a measure of $\overline{K}_P$ can be derived from information in the June Returns – see **Appendix C**; however, with AWWA utilities and BWC zones it does not appear that even a proxy for $\overline{K_P}$ is available. The components of Z_P will be variables such as surface water proportion, resource pumping head, etc. Where the variable may take a zero value, it will be used in (1 + *variable*) form.

One further issue requiring attention is how best to distinguish between boreholes and other sources. Initially, this is done by having a control for the proportion of surface water. However, it also turns out – see **Section 5** – that the Ofwat data on the size distribution of works can be exploited to yield some insight, even though the costs of boreholes and other supplies are not separately identified.

The detailed methodologies and results for water production are reported below in **Sections 2-7** of this chapter.

c. Cost function for water distribution

The general specification for a cost function for water distribution following the approach in **(a)** above would be:

$$VCD = VC_D(DO, p_{LD}, \overline{K_D}, Z_D) \qquad \ldots\ldots\ldots\ldots \qquad (4.5)$$

Where *VCD* is the variable cost of water distribution, *DO* is a measure of distribution output, p_{LD} is a price for variable inputs, $\overline{K_D}$ is a measure of water distribution capital and Z_D is a vector of control variables. The measurement of *DO* is not straightforward because of the spatial aspect of distribution. In **Chapter V** a new measure is developed and tested. For similar reasons as with water production, it is assumed that the variation between cases in p_{LD} is small, so that this term can be dropped. There are other practical and conceptual issues that arise in trying to implement (4.5) but, as explained in **Chapter V**, it proved better to work with a cost function derived from a simpler Leontief-type production function for water distribution.

The detailed methodologies and results for water distribution are in **Chapter V**.

d. Cost function for water production and distribution combined

Based again on (4.3) above, the starting specification would be:

$$VCPD = VC_{PD}(DO, p_{LPD}, \overline{K_{PD}}, Z_{PD}) \qquad \ldots\ldots\ldots\ldots \qquad (4.6)$$

Where *PD* subscripts signify quantities related to water production and distribution. Accordingly, in this specification, the variables will need to be measured so as to cover both production and distribution[40]. The arguments for using *DO* in (4.6) rather than the quantity of water produced or consumed as in conventional utility studies is that this is the relevant measure of output when water is being both produced and distributed. Other points discussed in **Sections b** and **c** above, such as the assumption of small variation in *p*, continue to be relevant in this context.

The detailed methodologies and results for water production and distribution combined are reported in **Chapter VI**.

e. What about using a panel data approach?

A natural question to arise at this point is whether it would be productive to use a panel data approach. At the beginning of this research, it had indeed been the intention to put together a panel of Ofwat data, as had been done by previous researchers (notably Stone & Webster Consultants (2004)). The arguments in favour of this approach are very strong when the aim is to estimate a structural relationship and there are thought to be persistent unmodelled factors present which vary between cases but not across time. The use of panel data methods then enables these 'fixed effects' to be eliminated and the relationship of interest to be more clearly exposed.

There are however some substantive arguments against using this approach in the present context, as well as significant practical difficulties:

a. Most of the companies reporting to Ofwat serve too many settlements for settlement level effects to be observed through the Ofwat data. So, although a more than sufficient number of years of Ofwat data now exist (1992-2007) to enable panel methods to be used (notwithstanding a steady diminution in the total number of companies due to mergers and take-overs), the results would be of limited value for the purposes of this research.

[40] It should be noted that overhead costs (such as billing and research) which are not allocated to water production or distribution in the Ofwat data have not been included in the analysis.

b. In the case of the BWC data, the data made available is only for one year. Even if comparable data for other years could be obtained (unlikely), the time required to process the information for a single year was very great (see **Appendix H**) and the time and resources are not available to repeat this for additional years.

c. Apart from these practical issues, it is not clear what are the fixed effects that one would be trying to remove. Most of the obvious candidates (size of service area, density of settlement, proportion of borehole water) can be measured and are of interest in their own right so that it would seem better to keep them visible, as in a cross-section analysis.

d. A further concern is that examination of 6 years' Ofwat data for the WoCs found that the year to year variation in key variables was small, so that these differences appeared comparable in size to the likely measurement errors, raising the possibility that regressions using these differences (as panel methods do) would be nearly meaningless. (It may be wondered whether the results obtained by Garcia & Thomas (2001), which are based on a 3-year panel of 55 utilities in the Bordeaux area, might also be vulnerable on this score, despite the sophistication of their methods.)

In the light of these arguments, a panel data approach has not been pursued in this research. The emphasis instead has been put on obtaining results using a single year cross-section, exploiting to the full the detailed information in the Ofwat data or provided by BWC – for example by constructing a set of data for BWC zones and urban districts covering both production and distribution with which to carry out a cross-section analysis.

2. Application to water production

a. Introduction

A first step towards understanding the economics of water supply at settlement level is to make an assessment of scale economies in water production at plant level. Although cases where a single plant serves a single settlement are not very common in England & Wales (they are more common in the USA), economies of scale at plant level provide a useful starting point even where a settlement is served by several plants. Of course, the number of plants used may not be determined solely by cost considerations. The capacity and other characteristics of local water resources will vary from place to place, and there are likely to be limits placed on the amounts that can be abstracted from rivers or pumped from boreholes[41]. Moreover, companies often try to ensure that communities are served by more than one source for water quality and security of supply reasons. Nevertheless, there is likely to be some discretion about the amounts taken from different sources, and hence some scope for relative cost to play a part in determining the pattern of supply.

There are important differences in the processes involved in the production of treated water depending whether the source is groundwater or surface water:

> *Groundwater* is obtained from boreholes[42] and is generally of relatively good quality, requiring less treatment[43]. Boreholes (BHs) usually have a relatively low capacity (up to about 15 Ml/day), and such treatment as is required is often provided by a facility at the well-head; they are often unmanned, being remotely monitored and serviced as necessary by area-wide teams.
>
> *Surface water*, on the other hand, is generally of lower quality, being obtained either from impounding reservoirs or river intakes. Treatment is then provided in relatively large scale water treatment works (WTWs), ranging in capacity from about 20 Ml/day to over 300 Ml/day. These facilities typically occupy quite large sites and have a permanent workforce. Whereas with boreholes, acquisition and treatment are more or less a single integrated process, with WTWs the water comes from separate facilities, such as reservoirs or river pumping stations, which may themselves involve substantial investment and operating costs.

Analysis of water production costs needs to try to take into account all these complications.

b. Specification

The general specification developed for water production in **Section 1 (b)**, after dropping p_L, is:

$$VCP = VC_P(QP, \overline{K_P}, Z_P) \qquad \ldots\ldots\ldots \qquad (4.7)$$

Where *VCP* is variable cost of water production (i.e. water acquisition and treatment), *QP* is volume of water produced, $\overline{K_P}$ is a measure of water production capital and Z_P is a vector of control variables. Ideally, this specification would be estimated in a flexible form (such as translog) but this would require more observations than are available in

[41] In England & Wales, such limits are reflected in the annual licensed volume in the abstraction licences granted by the Environment Agency, for an annual fee.

[42] Or sometimes natural springs.

[43] However, some borehole water is of low quality and may have to be sent to a WTW for treatment.

the applications reported below. Another constraint, using BWC or AWWA data, is lack of information on $\overline{K_P}$. However, there is scope for testing the effect of several possible components of Z_P . The specification adopted is therefore more restrictive:

$$\ln VCP = \alpha_0 + \alpha_1 \ln QP + \alpha_2 (\ln QP)^2 + \sum_i \alpha_i \ln Z_{Pi} \qquad \ldots. \qquad (4.8)$$

Despite its limitations, this specification should be adequate to give an indication of operating cost economies of scale in water production, subject to the available controls. On the face of it, the absence of any term for capital cost is a drawback.

It is worth noting here that in the engineering literature[44] it is generally accepted that the costs of water treatment at plant level can reasonably be represented by a function of the form:

$$TCP = \beta(QP)^\alpha \qquad \ldots\ldots\ldots\ldots.. \qquad (4.9)$$

where *TCP* is total production costs and *QP* is volume of water produced, with α<1, reflecting scale economies in both the capital and operating costs of water treatment plant. This basic specification can be refined in various ways. To test for the possibility that economies of scale peter out as scale increases, a term in $(QP)^2$ can added. If there are other known factors leading to differences in costs between the cases being investigated, control variables for these can also be added. In the case of water treatment, such factors might include types of water being treated, standards of treatment, technology used and age of plant. So the end result is a specification rather similar to (4.8).

3. Application to BWC data

a. Data issues

The water supply operations of BWC cover the full range of supply sources, types of treatment works and distribution arrangements. Sources include boreholes, reservoirs, river abstractions and bulk imports from other companies. Information on the

[44] See for example Clark & Stevie (1981), p.20 or Grigg (1986), p.67. The latter includes the following table (last column calculated from Grigg's data, indicating an α value of 0.75):

Size of treatment plant	**Population served**	**Total project cost ($m, 1978)**	**Annual capital cost per person served ($, 1978)**
700 gpm package	4,500	0.710	27.6
5 mgd conventional	20,000	2.364	19.8
40 mgd conventional	125,000	10.334	14.8
130 mgd conventional	575,000	26.050	7.7

proportions from each source is included in the company’s June Return. Broadly, this shows:

Source	**Volume (%)**
Boreholes	32
Impounding Reservoirs	10
River abstractions	38
Bulk imports	20

Table 4.1: BWC water sources by type

The return also indicates that most of the bulk imports come from impounding reservoirs. Thus less than half the company’s water comes from the relatively demanding (in terms of treatment) river sources.

Information provided by the company shows that fewer than 20 of the more than 150 treatment works reported to Ofwat are large conventional WTWs. However, these large works account for two-thirds of BWC water production, having an average flow of about 74 Ml/day. For the other works (all borehole sites), the average flow is about 5 Ml/day. One further point to note here is that about 20 plants (all boreholes) which are counted in BWC’s June Return produced no output and are recorded as being for emergency use only and have therefore been excluded from the tables below. Presumably, the operating costs associated with these plants are negligible but they will no doubt have a capital value.

An analysis of BWC works by size band, based on this information, shows their size distribution to be as in **Table 4.2.**

Size of plant[45]	**WTWs**		**Boreholes**[a]	
	% of plants	**Av output (Ml/d)**	**% of plants**	**Av output (Ml/d)**
Band 1 (≤ 1 Ml/day)	-	-	7.5	0.30
Band 2 (>1 to ≤ 2.5 Ml/day)	-	-	19.4	1.13
Band 3 (>2.5 to ≤ 5 Ml/day)	-	-	21.6	2.38
Band 4 (>5 to ≤ 10 Ml/day)	-	-	25.4	4.61
Band 5 (>10 to ≤ 25 Ml/day)	5.9	18.60	23.9	9.91
Band 6 (>25 to ≤ 50 Ml/day)	41.2	28.27	2.2	16.42
Band 7 (>50 to ≤ 100 Ml/day)	17.6	42.44	-	-
Band 8 (>100 to ≤ 175 Ml/day)	17.6	104.56	-	-
Band 9 (>175 Ml/day)	17.6	217.68	-	-

Note: (a) Excluding zero output (emergency use) works.

Table 4.2: BWC size of treatment plants

This analysis underlines the relatively small size of borehole supplies compared with WTWs – in fact, all the WTWs show a larger output than any of the boreholes despite the small overlap in their range. This helps to explain why two thirds of output comes from WTWs although they only account for 11.3% of the number of plants (see **Table 4.3**).

As regards type of treatment, the June Return to Ofwat distinguishes 5 categories[46]. For BWC, analysis using the same company information as for **Table 4.2** yields the figures shown in **Table 4.3.**

[45] Ofwat guidance states that works should be allocated to size bands according to each work's peak hydraulic capacity, not its distribution input in a particular year.

[46] Ofwat guidance defines these as: SD - Simple disinfection; W1 – SD + simple physical treatment (e.g. filtration); W2 – Single stage complex physical or chemical treatment (e.g. filtration + coagulation/flocculation); W3 – More than one stage of complex treatment (e.g. orthophosphate dosing); W4 – Other processes with high operating costs (e.g. ozone addition, UV treatment, arsenic removal).

Type of treatment	WTWs		Boreholes	
	% of Plants	% of Output	% of Plants	% of Output
Simple disinfection	-	-	22.4	15.0
W1	-	-	2.2	2.0
W2	-	-	23.9	19.3
W3	17.6	40.1	25.4	37.1
W4	82.4	59.9	26.1	26.6
% of total WTWs + BHs	**11.3**	**67.7**	**88.7**	**32.3**

Table 4.3: BWC type of treatment and plant size

This shows that all the water produced by WTWs is treated to level W3 or W4; for boreholes however, nearly half provide only the simpler kinds of treatment (disinfection, W1 or W2). Other things equal one would expect unit costs to be higher for the higher levels of treatment but also that this extra cost might be offset to a greater or lesser extent in WTWs by economies of scale in these larger plants.

The cost information for WTWs provided by BWC for this research shows operating costs for each works. Among “other water supply costs” not allocated to WTWs or boreholes are large amounts for Rates (35% of total water supply costs), Environment Agency abstraction licence fees (9.6%), Bulk imports (7.6%) and Aqueducts (1.4%). In the cost analysis below, the last three items were attributed to works on what appeared to be a reasonable basis but there is no obvious way to do this with local authority rates, which are therefore excluded (they are also excluded from Ofwat’s cost analyses). No information was provided by BWC on either asset values or capital maintenance by works and this has limited the analysis that can be carried out. **Figure 4.1** shows a plot of average (or unit) cost (*UVCP*) against output (in Ml/d) for BWC’s WTWs.

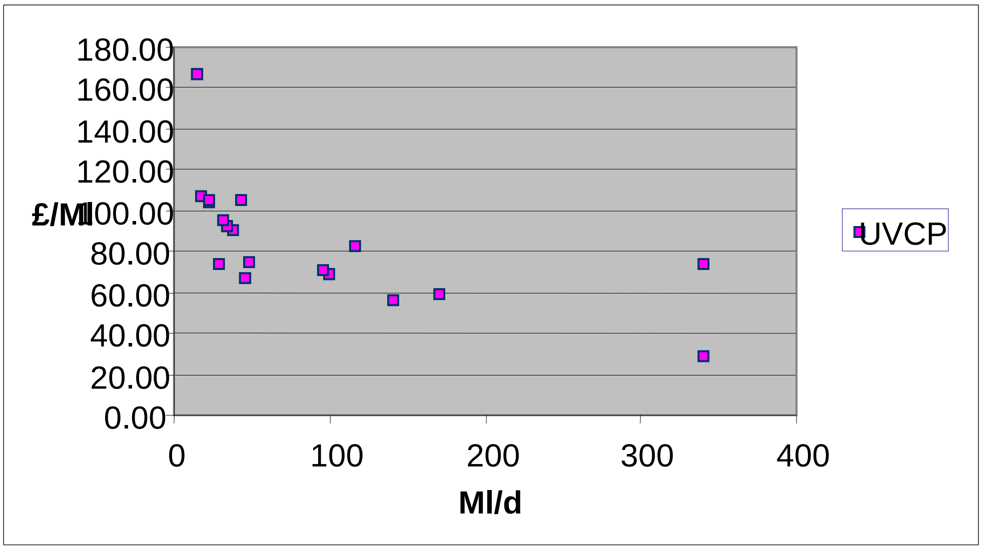

Figure 4.1: BWC water treatment works average (unit) costs

There are two markers for the largest works in this plot: the higher one includes the cost of the bulk imports which are treated at this works, which more than doubles the average cost. The effect of this on the estimates of economies of scale will be discussed in **Section b** below.

For boreholes, operating cost data is aggregated at county level and there is insufficient information to enable an assessment of economies of scale for this type of works. Indeed, boreholes are generally unmanned, being serviced by area-wide teams, so that allocating costs to individual boreholes may be difficult. However, it was observed that the average cost of borehole supplies is about £76.5/Ml compared with about an average of about £75/Ml for WTW supplies[47], indicating that these relatively small sources are relatively high cost, despite in general requiring less treatment.

b. Specification and results

For BWC's WTWs, the specification based on (4.8) is:

$$\ln(VCP) = \alpha_0 + \alpha_1 \ln(QP) + \alpha_2 (\ln QP)^2 + \alpha_3 W4D \quad \ldots\ldots\ldots\ldots.. \qquad (4.10)$$

Where *VCP* is operating costs, *QP* is quantity treated and *W4D* is a dummy for level 4 treatment (as **Table 4.3** shows, all WTWs operate to either level 3 or level 4).

The results obtained using (4.10), dropping the term in (ln*QP*)^2 where this was not significant, are as shown in **Table 4.4**:

	17 WTWs (excl. imports)	**17 WTWs (incl. imports)**	**16 WTWs (excl. largest)**	**15 WTWs (excl. largest and smallest)**
α_0 (Constant)	4.385	6.858	4.316	4.065
S.E.	*0.284*	*0.811*	*0.265*	*0.214*
α_1 (ln*QP*)	0.684***	-0.594	0.724***	0.781***
S.E.	*0.055*	*0.406*	*0.055*	*0.045*
α_2 (ln*QP*)^2	Dropped	0.167***	Dropped	Dropped
S.E.		*0.049*		
α_3 (W4D)	0.290**	0.179	0.208*	0.213**
S.E.	*0.121*	*0.107*	*0.120*	*0.091*
R^2	0.9255	0.9659	0.9316	0.9623

Table 4.4: regression results for BWC's WTWs using (4.10)
(Significance levels: * = 1%; ** = 5%; * = 10%; relative to 1 for α_1)**

[47] £75/Ml includes cost of imports; without imports the cost is £63/Ml.

The first two columns of **Table 4.4** compare the results for all 17 works, with and without imports to the largest works. There is some uncertainty about the amount of imports to attribute to this works but it is certainly a large amount and its inclusion makes a big difference, changing the sign of the coefficient on ln*QP* and producing a significant positive coefficient on (ln*QP*)^2, suggesting rather large economies of scale for smaller WTWs which diminish as the size of works increases. Leaving out the largest works (on the grounds that it is not typical, as well as uncertainty about the imported supply) moderates this result – see third column of **Table 4.4**. Finally, if the smallest works is also considered to be an outlier (see **Figure 4.1**), the results in the fourth column are obtained. For present purposes, where the objective is to arrive at a reasonable representation of economies of scale at plant level, the results in column 4 are the most appropriate ones to adopt. They indicate returns to scale of about 1.28 (1/0.781) for a typical WTW. They also show a significant extra cost associated with level 4 treatment.

Unfortunately, the limitations of the data mean that it is not possible to carry out a similar analysis for BWC's boreholes. However, as noted above, the average cost of borehole supplies is about £76.5/Ml, while the average size of boreholes is only 4.6 Ml/day. Referring back to **Figure 4.1**, this suggests that a similar plot for boreholes would lie below that for WTWs, as depicted (in log form) in **Figure 4.2**:

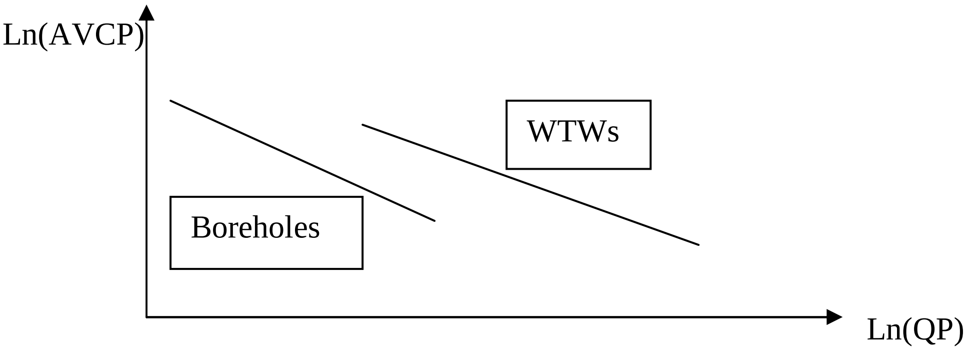

Figure 4.2: Sketch of relationship between average cost and size of works for boreholes and surface treatment works

One implication of this is that carrying out analysis of water production costs without regard to type of works is likely to be misleading. If possible, it would be desirable to try to identify the effect on costs of each type of plant separately.

4. Application to AWWA data

a. Data issues

Of the 897 utilities in the general information table form of the AWWA 1996 survey, only 548 provided information for the annual O&M expenses table. These provided the starting point for further investigation. Separate samples for the analysis of production costs and distribution costs respectively were then developed.

For **production costs**, the samples used in the regressions were obtained as follows:

Reason for dropping cases	Numbers affected
Starting point: Utilities in O&M table	548
No figure for water produced	-10
Supply + treatment cost = 0	-12
Outlier: UVCST > $2/’000galls	-7
Outlier: UVCST < $0.01/’000galls	-2
TreatQS sample	**517**
Omit utilities taking purchased water	-129
TreatQP sample	**388**
Of which: Groundwater only (**TreatGW sample**)	**161**
Of which: Surface water only (**TreatSW sample**) Of which: No of plants not reported or not clear	**145** -30
TreatSWN sample	**115**

Table 4.5: Selection of water production cases from AWWA 96 data

To give a visual impression of the data, **Figure 4.3** below plots the unit cost of production (supply + treatment) (*UCST*) against quantity supplied (*QS*) for the 517 cases in the **TreatQS** sample. Because of the wide dispersion in the data, this is shown in log form. Even in this form there is still considerable dispersion around the central tendency although a generally negatively sloped relationship is just about discernible, consistent with economies of scale in water production. A more precise assessment is given in **(b)** below.

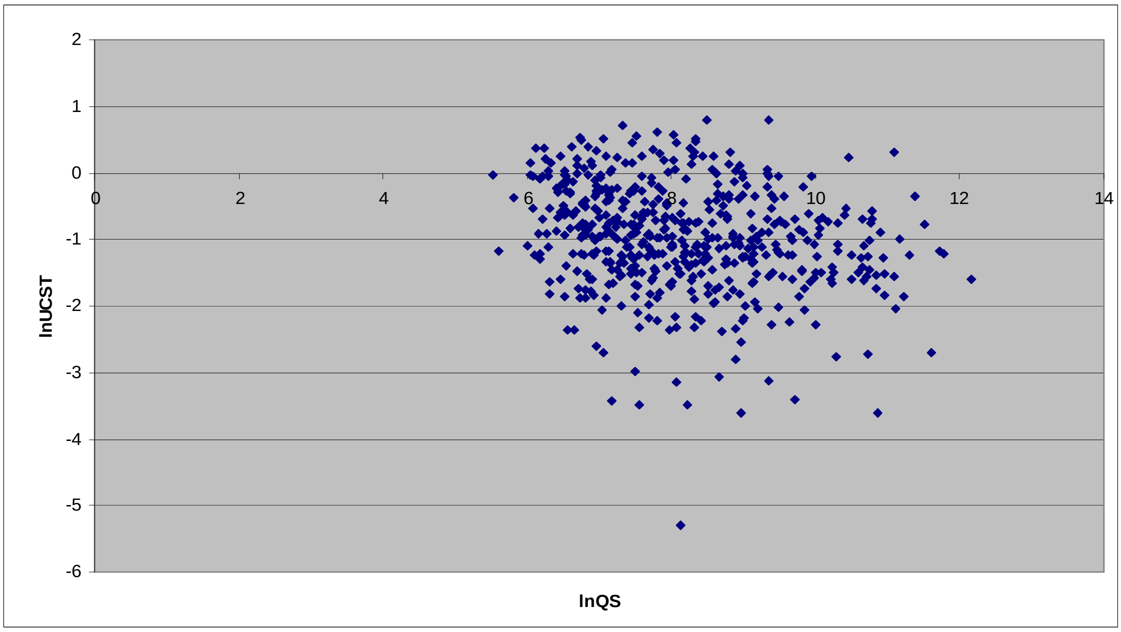

Figure 4.3: Log plot of unit water supply and treatment costs against volume supplied

b. Specification and results

The specification used in this section is based on (4.8), adapted to take account of the data available for US water undertakings. In comparison with the situation in England & Wales, many US undertakings operate only one production unit serving a single settlement; and many are either wholly surface water or wholly groundwater. These circumstances should facilitate the kind of analysis we are trying to carry out. On the other hand, the cost information does not include capital maintenance or depreciation (although current maintenance is included) so that analysis can only be done for operating costs. A further issue is that many US undertakings purchase considerable volumes of water from other undertakings, and it is not clear from the data whether this water is treated or untreated, so that the volumes to which the recorded treatment costs relate is also often unclear. To deal with this issue, our main analysis puts supply and treatment costs together. This leads to the specification:

$$\ln(CST) = \alpha_0 + \alpha_1 \ln QS + \alpha_2 (\ln QS)^2 + \alpha_3 \ln(1+SP) + \alpha_4 \ln(1+PP) \,\ldots.. \quad (4.11)$$

Where *CST* is the variable cost of water supplied (i.e. cost of purchased water as well as the cost of own water treatment), *QS* is the quantity of water supplied (including purchased water); *SP* is proportion of surface water and *PP* is proportion of purchased water. Using the **TreatQS** sample, this leads to the results reported in the first column of **Table 4.6** below. To see whether purchased water is distorting the results, we also carry out regressions leaving out cases where any of the water supplied is purchased (using the **TreatQP** sample) – see third column of **Table 4.6**. It can be seen that the coefficients on ln*QS* and (ln*QS*)^2 are not significant when both are included but dropping the (ln*QS*)^2 term leaves the coefficient on ln*QS* highly significant and of a plausible value, as shown in the columns marked (b) in **Table 4.6**.

Coefficient	**Using TreatQS sample**		**Using TreatQP sample**	
	(a)	**(b)**	**(a)**	**(b)**
α_0 (Const)	-4.970	-6.865	-3.681	-6.920
S.E.	*1.101*	*0.196*	*1.217*	*0.210*
α_1 (ln*QS*)	0.394	0.851***	0.076	0.855***
S.E.	*0.262*	*0.024*	*0.289*	*0.026*
α_2 (ln*QS*)^2	0.027*	Dropped	0.046***	Dropped
S.E.	*0.015*		*0.017*	
α_3 (l n(1+*SP*))	0.258**	0.259**	0.285***	0.278***
S.E.	*0.106*	*0.106*	*0.105*	*0.106*
α_4 (l n(1+*PP*))	1.021***	1.012***	n.a.	n.a.
S.E.	*0.152*	*0.152*		
R^2	0.7309	0.7293	0.7598	0.7552
No of cases	517	517	388	388

Table 4.6: Regression results using (4.11), AWWA data
(Significance levels: * = 1%; ** = 5%; * = 10%; relative to 1 for α_1)**

On the basis of the (b) columns in **Table 4.6**, the AWWA data provides evidence of plant level returns to scale in water production of about 1.18 (1/0.85). However, the US results are for operating costs only and one can only speculate what effect the inclusion of capital costs would have on these figures.

As it is possible to identify in the AWWA data a substantial number of utilities which supply only groundwater (from boreholes) (the **TreatGW** sample) or only surface water (the **TreatSW** sample), it seemed worth carrying out separate analyses for these cases using (4.11) when neither the *SP* control nor the *PP* control is required. Information about number of treatment plants (*TN*) in the AWWA data relates only to utilities supplying surface water (and therefore operating water treatment plants) but is missing for some of these utilities. For those for which this information is available, the effect of controlling for number of plants can be tested using the **TreatSWN** sample and specification (4.12) below:

$$\ln(CST) = \alpha_0 + \alpha_1 \ln QS + \alpha_2 (\ln QS)^2 + \alpha_5 \ln TN \qquad \ldots.. \qquad (4.12)$$

The results are reported in **Table 4.7** below:

Coefficient	Using TreatGW (Boreholes)		Using TreatSW (WTWs)		Using TreatSWN (single WTWs)	
	(a)	(b)	(a)	(b)	(a)	(b)
α_0 (Const)	-4.396	-7.373	-3.665	-6.304	-5.405	-5.832
S.E.	*2.143*	*0.376*	*1.949*	*0.332*	*2.037*	*0.354*
α_1 (lnQS)	0.178	0.911*	0.179*	0.800***	0.630	0.731***
S.E.	*0.522*	*0.048*	*0.453*	*0.039*	*0.476*	*0.043*
α_2 (lnQS)^2	0.044	Dropped	0.036	Dropped	0.006	Dropped
S.E.	*0.031*		*0.026*		*0.028*	
α_5 (lnTN)	n.a.	n.a.	n.a.	n.a.	0.864***	0.874***
S.E.					*0.184*	*0.176*
R^2	0.6992	0.6954	0.7510	0.7477	0.8254	0.8253
No of cases	161	161	145	145	115	115

Table 4.7: Regression results using (4.12) for groundwater only and surface water only cases, AWWA data
(Significance levels: * = 1%; ** = 5%; * = 10%; relative to 1 for α_1)**

These results again suggest that it is preferable to drop the term in (lnQS)^2, relying on the coefficients estimated in the (b) columns. It appears that there is a significant difference in scale economies between (groundwater) boreholes and (surface water) treatment works, with the US data suggesting both that operating costs for boreholes are on average lower than for treatment works and that returns to scale for boreholes (at about 1.10) are well below what they are for water treatment works (about 1.25). Indeed, when the number of works is controlled for (using the **TreatSWN** sample), returns to scale for the latter rise to 1.37.

It is revealing to examine these data in scatter plot form, as in **Figures 4.4** and **4.5**.

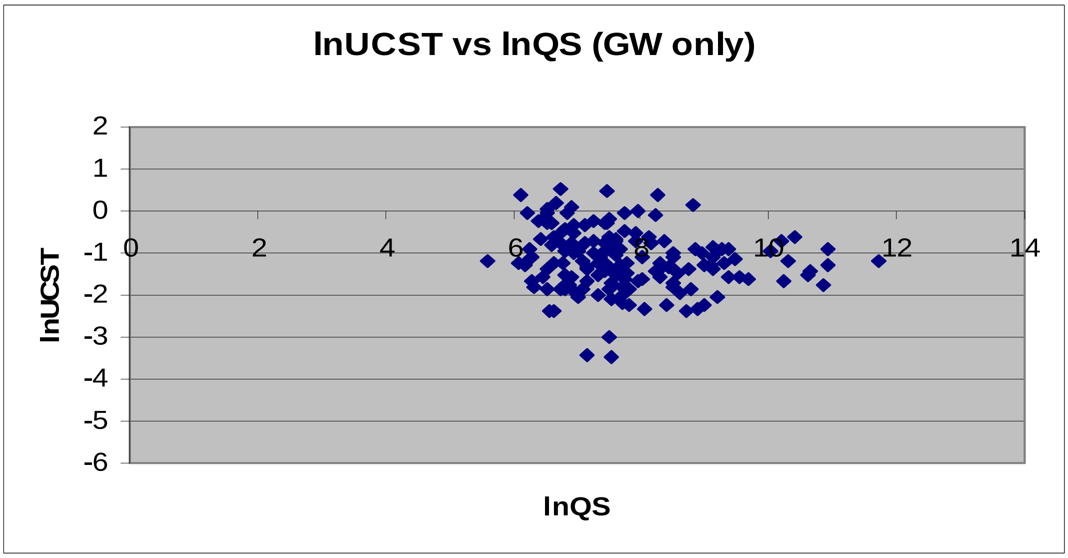

Figure 4.4: Log plot of average operating costs against quantity produced for US boreholes

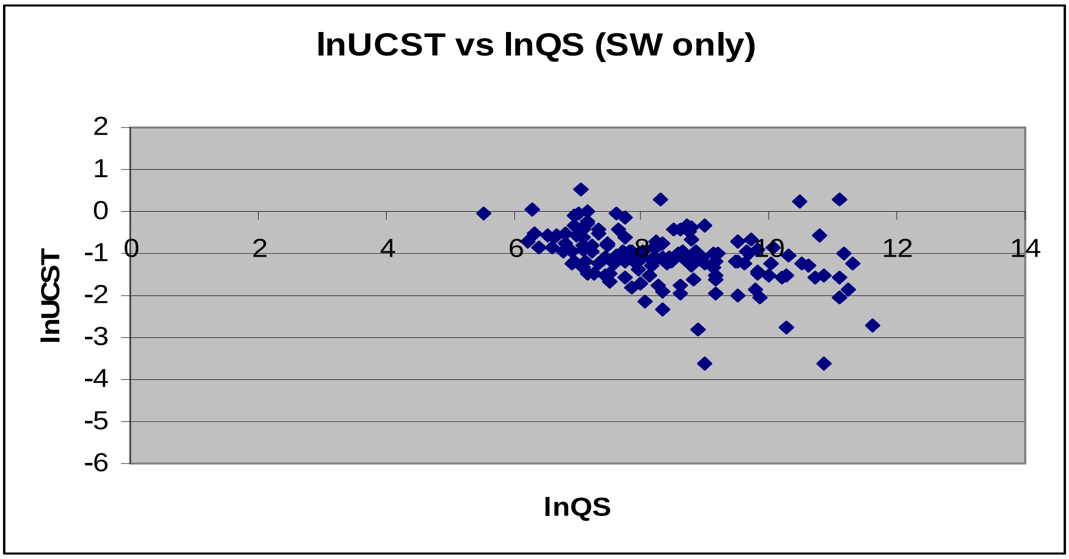

Figure 4.5: Log plot of average operating costs against quantity produced for US treatment works

From **Figure 4.5** it is evident that the relationship for (surface water) treatment works is reasonably coherent, which is reflected in the highly significant coefficient on ln(*QS*) in **Table 4.7**. For (groundwater) boreholes on the other hand, **Figure 4.4** shows much less structure. It appears that with boreholes, scale of output is not the major factor in determining costs. Although we do not have information to throw light on what these other factors might be, it would be reasonable to suppose that they include, *inter alia*, borehole depth (with its effect on pumping costs) and the level of treatment required, as is the case with UK water companies.

5. Using Ofwat data to differentiate between boreholes and WTWs

a. Data issues

The analyses in **Sections 2 - 4** above suggest that it would be very desirable to try to carry out separate investigation of production from boreholes and production from WTWs. In this section, a method for doing this using Ofwat data is developed.

As a first step, it is necessary to have information about the number (*TN*) and average output (*AQP*) of each type of treatment works for each company. Table 12 of the June Return does not quite give this degree of detail (and the information may not be wholly reliable[48]). However, by assuming that borehole works are all smaller than surface

[48] In a private communication, a member of Ofwat staff commented: “While we review these annually as part of the June Return process, because we do not use these variables in our modeling we do not subject them to the same level of scrutiny and checking as model variables. We do not consider that they are robust or consistently reported.” On the other hand, this information, like all that in the June Returns, has been certified by independent auditors appointed by Ofwat.

treatment works (as was the case for BWC) and then summing the number of works in each size band from the smallest until the proportion of distribution input from borehole sources is all accounted for, an approximate split between borehole works and surface treatment works can be made. Information provided directly by BWC showed that many of their smaller reported works (all boreholes) are not currently operational, being held for emergency or standby use. It is likely that the position is similar for other companies. To reduce the impact of this problem, works in the size 1 band (< 1 Ml/d) have been omitted from the analysis[49]. The resulting data set is shown in **Tables 4.8A** and **4.8B**:

Company[50]	**TN (No)**	**Boreholes[a] (No)**	**Treatment works (No)**	**AQP_B (Ml/day)**	**AQP_T (Ml/day)**
BWH	7	4	2	6.62	65.72
BRL	23	16	7	2.40	36.12
CAM	14	14	0	5.23	0
DVW	9	4	5	1.11	12.55
FLK	18	18	0	2.75	0
MKT	29	27	2	4.60	8.09
PRT	20	19	1	6.39	55.82
MSE	65	57	5	4.34	21.53
SST	29	24	2	5.87	94.97
SES	11	17	1	16.93	41.41
THD[51]	2	1	1	25.9	4.2
TVW	99	86	7	5.08	50.94

Note: (a) Excluding size band 1 and zero output works.

Table 4.8A: Estimated data on boreholes and treatment works for WoCs

Company[50]	**TN (No)**	**Boreholes[a] (No)**	**Treatment works (No)**	**AQP_B (Ml/day)**	**AQP_T (Ml/day)**
ANH	143	129	10	4.56	56.12
WSH	105	30	48	0.97	17.72
YKY	90	51	21	5.35	48.68
NNE	67	34	18	3.50	59.85
SWT	40	18	20	2.53	20.05
SVT	173	136	18	4.58	73.97
SRN	102	83	5	5.03	35.22
TMS	99	88	10	7.19	218.99
NWT	137	81	40	1.86	44.65
WSX	119	87	5	3.09	23.92

Note: (a) Excluding size band 1 and zero output works.

Table 4.8B: Estimated data on boreholes and treatment works for WaSCs

[49] This resulted in dropping 12 works (0.13% of output) for WOCs and 148 works (0.34% of output) for WaSCs.

[50] For key to company acronyms, see **Tables 3.1A** and **3.1B** in **Chapter III**.

[51] THD reported only 2 works and a borehole proportion of 0.834 but the June Return explains that water from all its 7 borehole sources is treated at Horsley Cross WTW while its other supply is shared with Anglian Water. The THD figures are therefore not fully comparable with those for other companies.

b. Specification and results

Now, a procedure to separately estimate scale effects for boreholes and treatment works, in a cross-company analysis, can be developed in the following way. Based on **Figure 4.2** and the results reported in **Table 4.7**, and assuming that companies all use the same borehole and WTW technologies, unit production costs (*UCP*) for boreholes can be modeled as:

$$UCP_B = \beta_B (AQP_B)^{\alpha_B - 1} \quad \ldots\ldots\ldots \quad (4.13)$$

and for surface works as:

$$UCP_T = \beta_T (AQP_T)^{\alpha_T - 1} \quad \ldots\ldots\ldots \quad (4.14)$$

Then the observed *UCP* for each company will be a weighted average of these two components:

$$UCP = BH_{prop}.UCP_B + (1 - BH_{prop}).UCP_T \quad \ldots\ldots\ldots \quad (4.15)$$

Hence

$$UCP = \beta_B.BH_{prop}(AQP_B)^{\alpha_B - 1} + \beta_T (1 - BH_{prop})(AQP_T)^{\alpha_T - 1} \quad \ldots\ldots \quad (4.16)$$

Where *BH_{prop}* is the proportion of production from borehole supplies. Note that it is necessary to work with unit production costs here for the averaging in (4.15) and (4.16) to be valid; these costs may be either unit variable costs (*UVCP*) or unit total costs (*UTCP*).

While the information in the Ofwat data does not enable the proportion of W4 treatment and resource pumping head to be linked directly to types of works, controls for these factors can be introduced by assuming that pumping head mainly affects boreholes while the proportion of W4 treatment applies generally, leading to:

$$UCP = (1 + W4P)^{\gamma}\{PHR^{\delta}\beta_B.BH_{prop}(AQP_B)^{\alpha_B - 1} + \beta_T (1 - BH_{prop})(AQP_T)^{\alpha_T - 1}\} \ \ldots \ (4.17)$$

This can be estimated using NLS. The results of so doing are reported in the first two columns of **Table 4.9** (in the first column the dependent variable is unit variable costs – *VCP/QDI* from **Table E.1** in **Appendix E**; in the second column it is unit total costs, *TCP/QDI*). Data for one company, THD, has been omitted for the reason given in **Footnote 51**.

There is a risk, when using average costs (costs divided by number of works) to assess economies of scale, that the results will be misleading. This is because if the size distribution of works across companies is very different, it is possible to get a finding of economies of scale using average costs although data for the individual works would not show this[52]. To check whether the results obtained using (4.17) may be vulnerable on this score, a more sophisticated specification can be constructed to make use of the information in the Ofwat June Returns on the number of treatment plants by size band, and the proportion of output from each size band. Modified non-linear versions of (4.13) and (4.14) to exploit this data are:

$$UCP_B = \beta_B \left\lfloor \sum_{i=2}^{9} p_{B\,i} \left(AQP_{B\,i}\right)^{\alpha_B - 1} \right\rfloor \qquad \ldots\ldots\ldots\ldots\ldots \qquad (4.18)$$

And

$$UCP_T = \beta_T \left\lfloor \sum_{i=2}^{9} p_{T\,i} \left(AQP_{T\,i}\right)^{\alpha_T - 1} \right\rfloor \qquad \ldots\ldots\ldots\ldots\ldots \qquad (4.19)$$

Where the *p*’s are proportions of output and the *i*'s indicate size bands.

These then lead to an amended version of the non-linear specification (4.17), which takes into account the size distribution of works and should therefore be more reliable. The resulting specification is[53]:

$$UCP = (1 + W4P)^{\gamma} \left\{ PHR^{\delta} \beta_B \left\lfloor \sum_i p_{B\,i} \left(AQP_{B\,i}\right)^{\alpha_B - 1} \right\rfloor + \beta_T \left\lfloor \sum_i p_{T\,i} (AQP_{Ti})^{\alpha_T - 1} \right\rfloor \right\} \quad \ldots. (4.20)$$

The results of running this specification are also shown in **Table 4.9**, in the last two columns.

Coefficients	**Using (4.17)**		**Using (4.20)**	
	Variable costs	**Total costs**	**Variable costs**	**Total costs**
β_B	14.1	61.8	14.1	70.4
S.E.	*9.5*	*50.0*	*10.9*	*65.8*
$\alpha_B - 1$	-0.02	-0.18	-0.04	-0.21
S.E.	*0.13*	*0.18*	*0.15*	*0.20*
β_T	327***	475**	325**	451**
S.E.	*102*	*167*	*104*	*165*
$\alpha_T - 1$	-0.40***	-0.26**	-0.36***	-0.22*
S.E.	*0.10*	*0.11*	*0.10*	*0.11*
γ	0.31	0.31	0.21	0.28
S.E.	*0.28*	*0.31*	*0.30*	*0.33*
δ	0.46**	0.35*	0.48**	0.34
S.E.	*0.16*	*0.18*	*0.17*	*0.20*
R^2	0.9782	0.9704	0.9757	0.9669
No of cases	21	21	21	21

Table 4.9: Results of non-linear regressions using (4.17) and (4.20), Ofwat data (Significance levels: * = 1%; ** = 5%; * = 10%)**

First, it may be noted that the results using (4.20) are not greatly different from those obtained using (4.17), indicating that the problem of heterogeneous plant sizes giving

[52] I am grateful to David Saal for pointing this out to me, with a constructed example.

[53] Terms in BH_{prop} are not required here as the *p*’s are measured as proportions of total output.

misleading results has not arisen in this case. However, the results using (4.20) are probably more accurate and therefore to be preferred.

Before turning to the scale parameters, brief comment on the control variables ($1 + W4P$) and PHR is in order. Tested on their own, the coefficient on the first (γ) was comfortably significant and the coefficient on the second (δ) not far off. Although running them together has rendered the coefficient on ($1 + W4P$) insignificant, it seemed best to retain it as in earlier parts of this chapter it has been found to have explanatory value.

Focusing then on the (4.20) results in **Table 4.9**, it may be seen that there is a rather low value for the constant β_B for boreholes (although the terms in ($1 + W4P$) and PHR will push it higher) while the scale parameter $\alpha_B – 1$ although negative is not significantly different from zero, so that constant returns to scale for this type of works cannot be rejected. For WTWs on the other hand, the constant term β_T is large and $\alpha_T – 1$ indicates returns to scale of about 1.56 (1/(1 – 0.36)) for variable costs (larger than was found for BWC's works in **Table 4.4**) and about 1.28 (1/(1 – 0.22)) for total costs. It thus appears that bringing in capital costs raises the value of the constant term (unsurprisingly) while reducing returns to scale.

6. Discussion of findings

The results of the investigations reported in this Chapter throw useful light on the economics of water production. These results are summarized in **Table 4.10** below, with the coefficients converted to returns to scale form:

Data source	No of cases	Speci-fication	WTWs		Boreholes	
			Operating costs	Total costs	Operating costs	Total costs
Ofwat companies (see Table 4.9)						
All Cos	21	(4.20)	1.56***	1.28*	1.04	1.27
(S.E.)			*(0.16)*	*(0.14)*	*(0.16)*	*(0.25)*
AWWA (see Table 4.7)						
TreatSW	145	(4.12)	1.25***	n.a.	-	-
(S.E.)			*(0.05)*			
TreatSWN	115	(4.12)	1.37***	n.a.	-	-
(S.E.)			*(0.06)*			
TreatGW	161	(4.12)	-	-	1.10*	n.a.
(S.E.)					*(0.05)*	
BWC (see Table 4.4)						
WTWs	15	(4.10)	1.28***	n.a.	n.a.	n.a.
(S.E.)			*(0.06)*			

Table 4.10: Estimated plant level returns to scale in water production (Significance levels, relative to 1: * = 1%, ** = 5%, * = 10%)**

Generally, there is strong evidence for plant level scale economies in WTWs although for boreholes the evidence is much weaker. The results for AWWA cases (see **Table 4.7** in particular) bring out quite well the difference between boreholes and WTWs, with plant level returns to scale of about 1.10 for the former and about 1.25 (or more) for the latter, for operating costs only. The method used in **Section 5** to derive similar plant level results using the Ofwat data has required some simplifying assumptions but they again indicate (**Table 4.9**) returns to scale for WTWs of about 1.28 on a full cost basis (considerably higher, 1.56, for operating costs only). To obtain better estimates would require fuller information for a reasonably large and representative sample of works. The closest we have to this ideal is the information for BWC's WTWs leading to the results reported in **Table 4.4.** Taking the last column of **Table 4.4** as the most appropriate to rely on, this shows a well-determined value of about 1.28 for operating cost returns to scale for WTWs in the size range 20-200Ml/day. It seems that bringing in capital costs would reduce this figure but by quite how much is difficult to say. For boreholes, positive returns to scale cannot be confirmed because of the wide confidence interval on the estimates.

What needs to be decided, in the light of these findings, is what figures would provide a reasonable representation of water production costs to use in modeling urban water supply. In **Chapter VI**, illustrative calculations of water supply costs for urban districts served by BWC, for the areas served by WOCs and for the areas served by US utilities are carried out. For the first two, the estimates of full cost scale effects obtained using

(4.20) in **Table 4.9** look suitable, while for the US, the estimates obtained using the **TreatQP** sample in **Table 4.6** will be adopted. With these parameters, the cost of water production for different levels of output can be calculated assuming average values for the relevant control variables. Boreholes do not in fact provide a good model for other types of urban infrastructure (their costs apparently depending mainly on factors other than scale), whereas WTWs (for which economies of scale appear to be significant) offer the prospect of a productive exploration of the trade-off between production economies and distribution diseconomies. But first, it is necessary to investigate scale effects in water distribution, and this is taken up in **Chapter V**.

V. ECONOMIES OF SCALE AND SPATIAL COSTS IN WATER DISTRIBUTION: EMPIRICAL INVESTIGATION

1. Introduction

Water distribution costs are more significant than water production costs. For example, in the case of BWC, although distribution operating costs are about the same in total as production operating costs, distribution capital costs are about twice as large. Scale effects in distribution therefore merit careful attention. The purpose of this chapter is to estimate the effect of settlement size and population density on water distribution costs.

If it is assumed that the technical options for water distribution can be represented by a standard production function, cost minimisation (or profit maximisation) would lead to a cost function for water distribution having the general form (see **Chapter IV, section 1(c)**):

$$VCD = VC_D(DO, p_{LD}, \overline{K_D}, Z_D) \qquad \ldots\ldots\ldots\ldots \qquad (5.1)$$

Where *VCD* is the variable cost of water distribution, *DO* is a measure of distribution output, p_{LD} is a price for variable inputs, $\overline{K_D}$ is a measure of water distribution capital[54] and Z_D is a vector of control variables. However, when this specification was applied – whether in simple or translog form – to data for 184 BWC zones, the results were inconclusive – see **Appendix G**.

This led to a fundamental reconsideration of what might be the characteristics of a production function for water distribution. The discussion in **Chapter III, section 3 (b)** concludes that there is a good case for two innovations in the analysis of water distribution costs:

> (i) Adoption of a Leontief-type production function, in view of the limited choice of technology and the lack of variance in input prices (particularly within a single company);
>
> (ii) Measurement of distribution output (*DO*) as the product of quantity consumed (*QC*) and the average distance to properties (φ). This measure is analogous to the tonne-km or passenger-miles used in transport studies.

[54] Following Garcia & Thomas (2001), capital in this formulation is taken to be “quasi-fixed”.

In **Section 2** of this chapter specifications are developed to enable these innovations to be implemented. In **Section 3**, the results of applying these specifications to BWC data are set out. In **Section 4** the same methods are applied to data for 11 of the smaller water companies in England & Wales; and in **Section 5** they are applied to data for 305 US retail only water utilities. The implications of the results are developed in **Section 6** and conclusions on scale effects in water distribution are drawn in **Section 7**.

2. Implementation of a Leontief-type production function for water distribution

a. Measuring distribution output

At first sight it might seem that the output of the distribution system is simply the volume of water delivered. But the essence of the distribution function is to deliver water to many different places, in the amounts and at the times when it is required. Not all these wider functions can easily be measured but the most important is the spatial aspect, the distance over which the water needs to be transported to reach customers.

To reflect this aspect, distribution output will be taken to be the quantity of water used at each property weighted by its distance from a central point. With some simplifying assumptions, it can then be shown (see **Chapter III, Section 3** for a fuller discussion) that for a circular settlement with density declining exponentially at a rate λ away from the centre, distribution output can be expressed as:

$$DO = \frac{2}{\lambda} w.N \frac{\left[1 - e^{-\lambda R}\left(1 + \lambda R + \frac{\lambda^2 R^2}{2}\right)\right]}{\left[1 - e^{-\lambda R}(1 + \lambda R)\right]} \qquad \ldots\ldots\ldots\ldots\ldots. \qquad (5.2)$$

Where *w* is consumption per property, *N* is number of properties, and *R* is the radius of the settlement.

(5.2) shows this measure of distribution output to be the product of two components, total consumption (*QC* = *w.N*) and a measure of average distance to properties (φ) which is a function of λ and *R* given by:

$$\phi(\lambda, R) = \frac{2}{\lambda} \frac{\left[1 - e^{-\lambda R}\left(1 + \lambda R + \frac{\lambda^2 R^2}{2}\right)\right]}{\left[1 - e^{-\lambda R}(1 + \lambda R)\right]} \qquad \ldots\ldots\ldots\ldots \qquad (5.3)$$

The implications of this expression are sketched in **Figure 5.1** which indicates how, for given *N*, higher values of λ will be associated with a larger settlement radius *R* if the central density d_0 is the same.

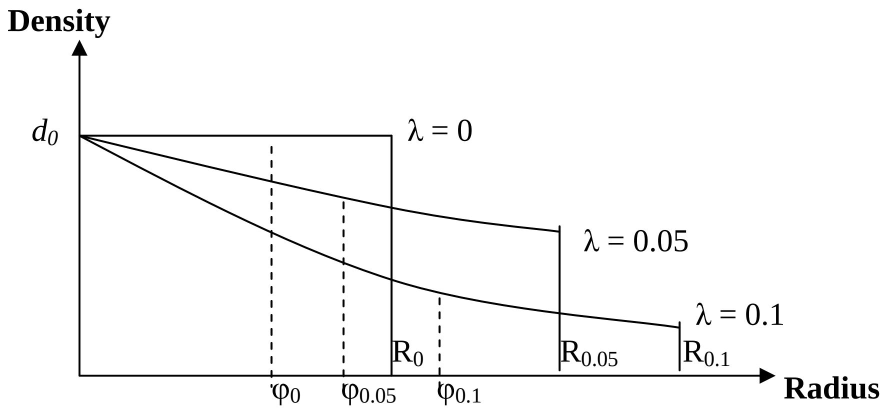

Figure 5.1: Relationship between density and settlement radius for different values of λ (not to scale)

In **Figure 5.1,** the average distance to properties, *φ(λ,R)*, is indicated by the dotted lines: when λ = 0, it is 2/3 R; with higher values of λ, it increases as determined by (5.3).

b. Cost function specification

The implication of a Leontief-type production function is that there is a particular amount of variable input associated with any particular level of output, i.e:

$$V = V(DO) \qquad \ldots\ldots\ldots \qquad (5.4)$$

If *V* is measured as *VCD* this becomes (5.1) shorn of the additional variables on the RHS. And for any particular level of output and variable input, there will be an associated amount of capital input, which is why a capital variable is not needed in (5.4)[55]. Returns to scale can be estimated from (5.4) alone.

[55] It is in this respect that the approach here differs from the ‘quasi-fixed’ capital approach of Garcia & Thomas (2001).

A note on Leontief production functions

If technology is such that Q units of output require $u.Q^{\alpha}$ units of fixed capital input and $v.Q^{\beta}$ units of variable inputs, three distinct cases arise:

4. α = β = 1: This is the textbook Leontief production function, which has the two properties: (a) K/V = u/v (i.e. a constant); and (b) $\frac{\partial(\ln K)}{\partial(\ln Q)} = \frac{\partial(\ln V)}{\partial(\ln Q)} = 1$; i.e. constant returns to scale.
5. α = β = γ (γ≠1): This can be called a Leontief-type production function. It has the two properties: (a) K/V = u/v (i.e. a constant); and (b) $\frac{\partial(\ln K)}{\partial(\ln Q)} = \frac{\partial(\ln V)}{\partial(\ln Q)} = \gamma$, i.e. increasing or decreasing returns to scale depending whether γ is <1 or >1.
6. α ≠ β: This is a new case, which does not seem to be discussed in the literature. It has the properties: (a) $\frac{K}{V} = \frac{u}{v} Q^{\alpha-\beta}$ (i.e. varies with the level of output); and (b) returns to scale also varies with output, being a function of $\frac{\partial(\ln K)}{\partial(\ln Q)}$ and $\frac{\partial(\ln V)}{\partial(\ln Q)}$.

To check whether the data for the 35 BWC "urban districts" (see **section 3(b)** of this chapter) are consistent with a Leontief-type production function (Type 2 above), the following regressions were carried out (using *VCD* as *V* and capital maintenance *CMD* as *K*):

i. ln(*K/V*) = 1.177 – 0.012ln*DO* , showing that this is not Type 3 above.
(0.045)

ii. ln*K* = 1.452 + 0.941ln*V* , suggesting that this is not Type 1 above.
(0.070)

iii. ln*K* = 4.001 + 0.608ln*DO* , and
(0.041)
ln*V* = 2.824 + 0.620ln*DO*
(0.026)
Confirming that the two coefficients are not significantly different from each other, with $\gamma \approx 0.61$. Returns to scale can be estimated from either relationship.

A simple specification for (5.4), convenient for assessing elasticities, would be:

$$\ln VCD = \alpha + \beta \ln DO \qquad \ldots\ldots\ldots\ldots \qquad (5.5)$$

However, noting from (5.2) that *DO* is the product of *QC* (= *w.N*) and φ, so that ln*DO* = ln*QC* + lnφ, the specification (5.6) below would help to expose the different effect on distribution costs of variations in volume and variations in average distance to properties:

$$\ln VCD = \alpha + \beta_1 \ln QC + \beta_2 \ln \phi \qquad \ldots\ldots\ldots.. \qquad (5.6)$$

It is for consideration whether there are control variables that it would be desirable to add to the above specifications. One possibility is distribution pumping head, which may reflect to some extent differences in hilliness between areas. However, this information is not available below company level in England & Wales. Although higher leakage rates might be expected to add to distribution costs[56], a control for this factor is not appropriate for the reasons given in **Chapter III, section 6**. A further possibility is the proportion of urban land in an area. Where data is available, a control for this factor can be tested.

c. Estimating distribution elasticities

Although specification (5.6) will provide an indication of the different effect on distribution costs of changes in volume and changes in average distance to properties, the estimated coefficients do not provide direct measures of distribution elasticities. This is because *N* and φ are both functions of λ and *R* and so are not independent of each other. Three elasticities are of particular interest:

(i) ε_w, measuring the response of distribution costs to changes in water consumption per property;

(ii) ε_A, measuring the response of distribution costs to changes in distribution area;

(iii) ε_N, measuring the response of distribution costs to changes in the number of properties.

To evaluate these elasticities, it is necessary to start from a variant of (5.6).

We can re-write *DO* as:

$DO = w.\psi$, where $\psi = N.\varphi$ is total distance to properties ……. (5.7)

(5.6) can then be re-stated as:

$$\ln VCD = \alpha + \beta_1 \ln w + \beta_2 \ln \psi \qquad \text{……….} \qquad (5.8)$$

Now, from **Chapter III**, (3.21) and (3.20), we have:

$$\psi = \frac{4\pi d_0}{\lambda^3}\left[1 - e^{-\lambda R}\left(1 + \lambda R + \frac{\lambda^2 R^2}{2}\right)\right], \text{ and} \qquad \text{…………….} \qquad (5.9)$$

$$N = \frac{2\pi d_0}{\lambda^2}\left[1 - e^{-\lambda R}(1 + \lambda R)\right] \qquad \text{…………….} \qquad (5.10)$$

Evaluating ε_w is straightforward:

[56] Although there is some ambiguity here: higher distribution costs may be incurred to keep leakage rates low.

$$\varepsilon_w = \frac{\partial(\ln VCD)}{\partial(\ln w)} = \beta_1 \qquad \text{……………} \qquad (5.11)$$

This can be viewed as a pure quantity effect, measuring the response of distribution costs to changes in water consumption per property (numbers of properties and other distribution area characteristics held constant).

The other elasticities are more difficult to evaluate and are not constants but vary with scale. It is helpful to start with a visual representation of what it is that the estimated elasticities might be trying to measure. In the monocentric urban model underlying our measure of distribution output, the configuration of a settlement is reflected in the four parameters: d_0, λ, N and R. The data used suggest a value for d_0 of about 30 properties/Ha and, although there could be cases with a higher value (e.g. high rise city centres) or a lower value (e.g. towns lacking a centre), 30 properties/Ha has been assumed throughout. The relationship between the parameters is then such that if any two of the remaining three is fixed, the third is also determined. Cases of particular interest then are:

(a) ***Densification***[57]: Number of properties (N) varies, while settlement radius (R) is held constant (λ also therefore varying);

(b) ***Dispersion***: Coefficient of dispersion (λ) varies, holding number of properties (N) constant (R also therefore varying);

(c) ***Suburbanisation***: Number of properties (N) varies, holding λ constant (R also therefore varying) ;

(d) ***Constant density***: Number of properties (N) varies, holding density (N/A) constant (when both λ and R vary).

The resulting variations in settlement configurations are portrayed in **Figure 5.2.**

[57] It is recognised that this term has acquired particular policy connotations in the urban planning context; here it is simply adopted as a convenient descriptive label.

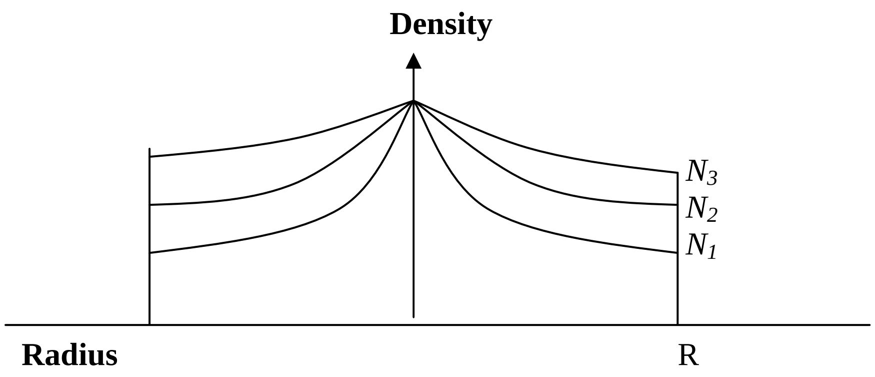

Figure 5.2: (a) Settlement cross-sections: *R* constant, *N* varies (‘*densification*’)

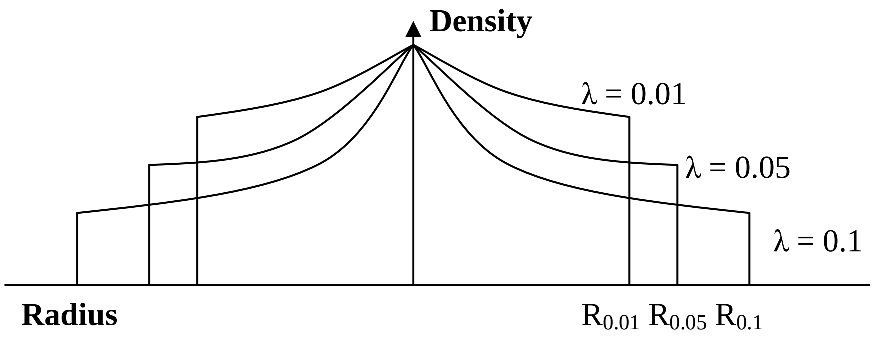

Figure 5.2: (b) Settlement cross-sections: *N* constant, λ varies (‘*dispersion*’)

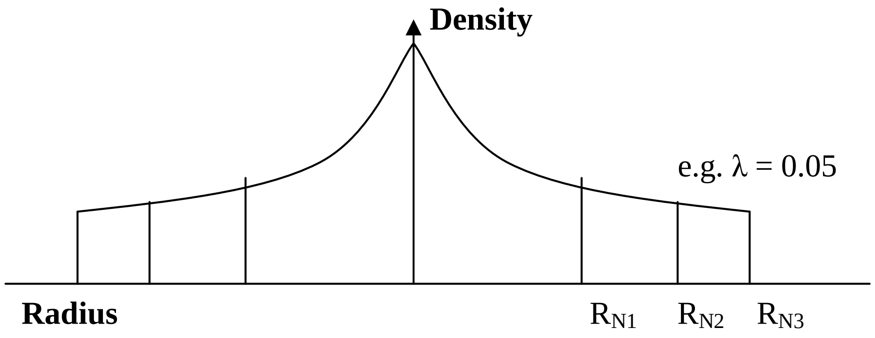

Figure 5.2: (c) Settlement cross-sections: λ constant, *N* varies (‘*suburbanisation*’)

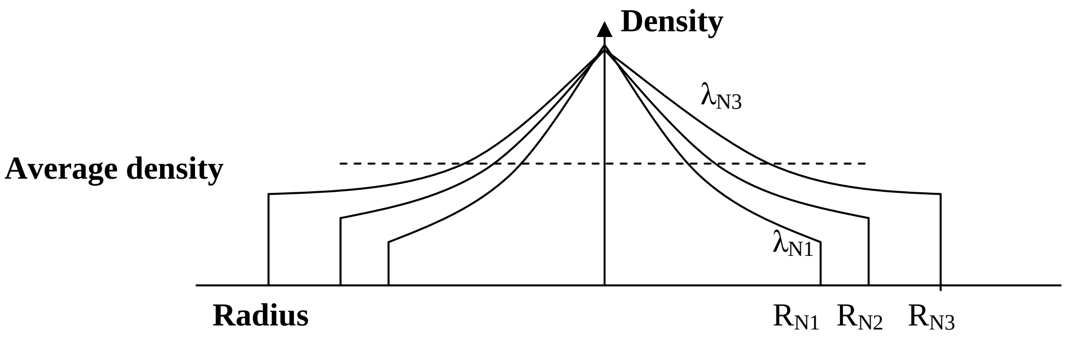

Figure 5.2: (d) Settlement cross-sections: Density constant, *N* varies (‘*constant density*’)

The complex form of equations (5.9) and (5.10) makes the derivation of expressions for the elasticities corresponding to these cases rather tricky[58]. The least mathematically awkward case is (c) ("*suburbanization*"). In this case λ is constant, say $\bar{\lambda}$. An expression for $\varepsilon_{R/\bar{\lambda}}$ (the elasticity of cost with respect to variations in *R*, conditional on $\bar{\lambda}$) can then be derived as follows:

$$\varepsilon_{R/\bar{\lambda}} = \frac{\partial(\ln VCD)}{\partial(\ln\psi)}.\frac{\partial(\ln\psi)}{\partial(\ln R)} = \beta_2.\frac{R}{\psi}.\frac{\partial\psi}{\partial R}$$

$$= \beta_2.\frac{R}{\psi}\left\{\frac{4\pi d_0}{\lambda^3}\left[-e^{-\lambda R}\left(\lambda+\lambda^2 R\right)+\left(1+\lambda R+\frac{\lambda^2 R^2}{2}\right)\lambda.e^{-\lambda R}\right]\right\}$$

$$= \beta_2.\frac{R}{\psi}.2\pi R^2 d_0.e^{-\lambda R} \qquad \ldots\ldots\ldots\ldots \qquad (5.12)$$

Which can alternatively be expressed in area form, using $\frac{d(\ln R)}{d(\ln A)} = \frac{1}{2}$, as

$$\varepsilon_{A/\bar{\lambda}} = \beta_2.\frac{R}{\psi}.\pi R^2 d_0.e^{-\lambda R} \qquad \ldots\ldots\ldots\ldots \qquad (5.13)$$

This is the elasticity of cost with respect to area served, conditional on $\bar{\lambda}$. Evidently, it is a (rather complex) function of *R* and λ but is clearly positive.

From (5.10), number of properties (*N*) varies with *R* (and *A*), so that there is a related elasticity $\varepsilon_{N/\bar{\lambda}}$, the elasticity of cost with respect to variations in *N*, conditional on $\bar{\lambda}$. It can be derived as follows:

$$\varepsilon_{N/\bar{\lambda}} = \varepsilon_{R/\bar{\lambda}.}\frac{N}{R}.\frac{\partial R}{\partial N} = \varepsilon_{R/\bar{\lambda}}.\frac{N}{R}.\frac{1}{\left\{\frac{2\pi d_0}{\lambda^2}\left[-e^{-\lambda R}.\lambda+\left(1+\lambda R\right).\lambda.e^{-\lambda R}\right]\right\}}$$

$$= \beta_2.\frac{N}{\psi}.2\pi R^2.d_0.e^{-\lambda R}.\frac{\lambda^2}{2\pi.d_0\lambda^2 R.e^{-\lambda R}} = \beta_2.\frac{N}{\psi}.R = \beta_2.\frac{R}{\varphi} \qquad \ldots\ldots\ldots \qquad (5.14)$$

This elasticity simplifies quite nicely but it also is a function of *R* and λ. Since volume rises in line with *N* (if *w* is constant), a value for $\varepsilon_{N/\bar{\lambda}}$ = 1 would indicate constant returns to scale. However, higher values are to be expected because of diseconomies associated with expansion into lower density suburbs.

[58] I am grateful to George Fane (Australian National University, Canberra) for helping me to come to grips with this point.

The algebra involved in deriving elasticities corresponding to cases (a) ("*densification*"), (b) ("*dispersion*") and (d) ("*constant density*") proved intractable (the last two involving simultaneous variation in both λ and R)[59]. Evaluation for these cases is therefore carried out by means of illustrative calculations for hypothetical urban areas using average data values, as described in **Section 6** of this chapter. In case (a), a value of 1 for $\varepsilon_{N/\bar{R}}$ would indicate constant returns to scale, if w is held constant. However, the expectation is of a value between 0 and 1, as more properties in a given area should give rise to density economies. In case (b) N is fixed, so a positive value for $\varepsilon_{A/\bar{N}}$ would indicate diseconomies (higher unit distribution costs), if w is also held constant. In case (d), N, λ and R move in tandem and while a value of 1 for $\varepsilon_{N/\bar{D}}$ would indicate constant returns to scale, there is no *a priori* reason why observed values should not be greater or less than 1.

3. Application to BWC data

a. Data issues

A full description of the data on water distribution obtained from BWC can be found in **Appendix H.** In brief**,** information on numbers of properties, length of mains, water consumption, leakage and geographical area for some 3000 District Metering Areas (DMAs) was aggregated and combined with information on operating costs to enable the relationships developed in **Section 2** above to be estimated, first for 184 Water Quality Zones (WQZs) and then for 35 Urban Districts. For the purposes of this research, DMAs are too small, having little relationship to urban areas; WQZs are better but large urban areas may still comprise several WQZs, while in other cases more than one urban area is included in a WQZ. The 35 urban districts (omitting the more rural parts of BWC's supply area) have been selected to try to overcome these difficulties.

b. Results for BWC's 184 zones

To get a feel for the results obtainable by the application of our approach, we start by considering BWC's 184 zones. The key question is how to obtain a measure of distribution output (*DO*) for these zones. To be able to use (5.2) some simplifying assumptions are required:

[59] However, it can be noted that in case (b) "*dispersion*", the coefficient on lnφ in (5.6) is related to the elasticity $\varepsilon_{A/\bar{N}}$.

i. First, it is supposed that each zone can be treated as if it were a circular settlement;

ii. Next, a measure of area is needed. Actual areas include unoccupied or unserviced areas; but only areas having access to water mains can be serviced. The area of accessible land in each zone (A_o) can be estimated as $M/0.15$, where M is length of mains. This is because M/A is observed to be approximately 0.15 in fully urban zones; the argument then is that a similar ratio of mains to land with access to a supply will prevail in less urbanized zones – density of properties in terms of properties per km of mains is however generally much lower outside urban areas;

iii. Now the effective radius (R) for each zone can be estimated as $R = \sqrt{A_o / \pi}$, where A_o is the area of accessible land;

iv. λ can then be estimated from the observed property density N/A_o by interpolation in a table which calculates density in properties/Ha for different values of R and λ (see **Appendix I** for an extract from this table);

v. Density at the centre of each zone (d_0) is taken to be 30 properties/Ha (a little above the highest value observed for BWC's zones);

vi. Finally, by using water consumed, i.e. $w.N = QC$, in (5.2), that part of distribution costs attributable to leakage will be reflected in a higher unit distribution cost (the cost of producing the water lost to leakage is a separate matter, not relevant to this part of the analysis – although it will be relevant when water production and water distribution are brought together in **Chapter VI**.)

With these assumptions, distribution output (DO) for each zone can be calculated as:

$$DO = QC.\phi(\lambda, R) \qquad \ldots\ldots\ldots\ldots \qquad (5.15)$$

Where $\varphi(\lambda,R)$ is given by (5.3) above.

Equations (5.5) and (5.6) can then be estimated giving:

$$(5.5) \quad \ln VCD = 2.702 + 0.645^{***} \ln DO \qquad \ldots\ldots\ldots\ldots \qquad (5.16)$$

(S.E. 0.013) *(R^2 = 0.9314)*

$$(5.6) \quad \ln VCD = 1.630 + 0.363^{***} \ln QC + 1.298^{***} \ln \phi \qquad \ldots.. \qquad (5.17)$$

(S.E. 0.079) *(S.E. 0.178) (robust)* *(R^2 = 0.9386)*

The result in (5.16) indicates economies of scale in distribution, since the coefficient on ln*DO* < 1 (very significantly so) but (5.17) then puts a rather different perspective on this result. The interpretation of the coefficient on ln*QC* in (5.17) is that higher consumption in a zone, whether due to greater usage per property or more properties on the existing network has a less than proportionate effect on costs (e.g. a 10% increase in *QC* would increase operating costs by about 3.6%).

The interpretation of the coefficient on $\ln\phi(\lambda, R)$ is less obvious. $\phi(\lambda, R)$ is a measure of the average distance to properties. Therefore a higher value for $\phi(\lambda, R)$, if *QC* is fixed[60], indicates that properties are more dispersed, implying a higher value for λ and hence also for *R*, as shown in the "dispersion" case in **Figure 5.2(b)**[61]. Any positive value for the coefficient on φ indicates that greater dispersion adds to the cost of distributing a given volume of water and is therefore a diseconomy. In fact this effect appears to be rather large here with (e.g.) a 10% increase in φ increasing operating costs by about 13%)[62]. This can be interpreted as a form of density effect, with lower density adding to distribution costs and higher density reducing costs[63].

A control for the proportion of urban land in each zone, *UAP*, is available with the BWC zone level data but when tested this was found not to be significant[64].

c. Results for 35 BWC "urban districts"

Although the BWC zones provide reasonably coherent units for analyzing distribution costs, they do not correspond very well with urban areas. In some cases, a large urban area is divided into several zones, while in other cases there is more than one urban area in a zone.

As a first step towards refinement, the maps defining the company's supply area held by Ofwat were examined to identify all the urban areas (as defined by ONS) with

[60] E.g. because average consumption and the number of properties is unchanged.

[61] If *N* is fixed, λ and *R* cannot vary independently of each other as they are linked through the relationship (5.10).

[62] However, the dispersion variable φ is relatively insensitive to changes in area served, as can be seen in **Table 5.2** below.

[63] But not all changes in density have this effect: if density increases or decreases without changing the average distance to properties (i.e. if the increase or decrease has exactly the same dispersion as existing properties) there will be no additional effect on costs from $\phi(\lambda, R)$ but there will still be a *QC* effect, since such a change implies an increase or decrease in *N*.

[64] The coefficient on ln(1+*UAP*) was 0.024 (*S.E.* 0.121); and tests for heteroscedasticity and omitted variables did not indicate any cause for concern.

population over 5,000 within the area[65]. This required checking several ONS regions as the company's boundaries, being based on river catchments, do not match those of the ONS regions. The company area was found to contain over 100 urban areas with more than 5k population, accounting for about 88% of the company population but only about 8.6% of the company area. The average density of these urban areas worked out at 38 persons/Ha, so that by difference the average density of the remaining 91% of the company's area is only about 0.5 persons/Ha[66]. It is indeed one of the more abiding impressions from looking at maps, even of such a densely populated country as England, how much of the surface area is not occupied by settlements. It is very evident that the mechanics and economics of providing services (such as water supply) to the relatively small numbers of people in isolated rural communities must be different from those of supplying large populations in densely settled areas.

With the assistance of a member of BWC staff, a relationship between BWC zones and ONS urban areas with population over 5,000 was established. The result was a list of 54 areas which we will call "urban districts" as they generally differ from urban areas as defined by ONS by including greater or lesser amounts of non-urban land. In 28 cases, a single ONS urban area was contained within a single zone so that there is a one to one correspondence between zone and "urban district"; in another 7 cases, a single urban area comprised more than one zone (including one urban area comprising 23 zones). In the other cases, the urban districts (comprising between 1 and 10 zones) included several urban areas (the number varying between 2 and 8). 49 zones could not be related to urban areas: for about 6 of these, this was because the ONS boundary for the urban area concerned was not clear, the rest were zones which did not appear to contain any urban areas with population over 5,000.

A new data set was created by amalgamating the zone data for the 54 "urban districts". Attention was then focused on the 35 "urban districts" (28 + 7) which covered a single urban area, as these were judged most likely to correspond reasonably well to the circular settlement model of (5.2). Adopting the same simplifying assumptions that were applied above to BWC's 184 zones to estimate φ, regressions were then run matching (5.16) and (5.17). The results obtained are reported below.

[65] In ONS (2004), "urban areas" are defined as areas of built up land of at least 20 Ha, having a population of 1,500 or more. Urban areas with population between 1,500 and 5,000 have been left out here.

[66] 38 persons/Ha is equivalent to about 16.5 properties/Ha; 0.5 persons/Ha is equivalent to about 0.22 properties/Ha.

$$\ln VCD = 2.824 + 0.620^{***} \ln DO \qquad \ldots\ldots\ldots\ldots \qquad (5.18)$$

(S.E. 0.036) robust (R² = 0.9444)

$$\ln VCD = 2.047 + 0.393^{***} \ln QC + 1.095^{***} \ln \varphi \qquad \ldots\ldots\ldots\ldots \qquad (5.19)$$

(S.E. 0.161) (S.E. 0.329) robust (R² = 0.9474)

These results are not very different from those obtained for BWC's zones, and indicate significant economies of scale with respect to volume (*QC*) and significant diseconomies with respect to the average distance measure (*φ(λ,R)*).

Re-estimating (5.19) in the (5.8) form gave:

$$\ln VCD = -4.572 + 0.432^{**} \ln w + 0.617^{***} \ln \psi \qquad \ldots\ldots.. \qquad (5.20)$$

(S.E 0.219) (S.E. 0.037) robust (R² = 0.9455)

From (5.20), the distribution elasticities identified at (5.11) – (5.14) above can be evaluated for these urban districts as:

$$\varepsilon_w = \beta_1 = 0.432$$

This is significantly less than 1 (at 5% level), indicating quite large increasing returns to this dimension of scale, although with a relatively high standard error.

$$\varepsilon_{A/\bar{\lambda}} = \beta_2 . \frac{R}{\psi} . \pi R^2 d_0 . e^{-\lambda R}$$

Taking β_2 = 0.617 from (5.20), values for this elasticity calculated from the 35 BWC urban districts data range from about 0.8 to about 0.2, with a tendency for higher values of $\varepsilon_{A/\bar{\lambda}}$ to be associated with lower values of *λ* (See **Figure 5.3**).

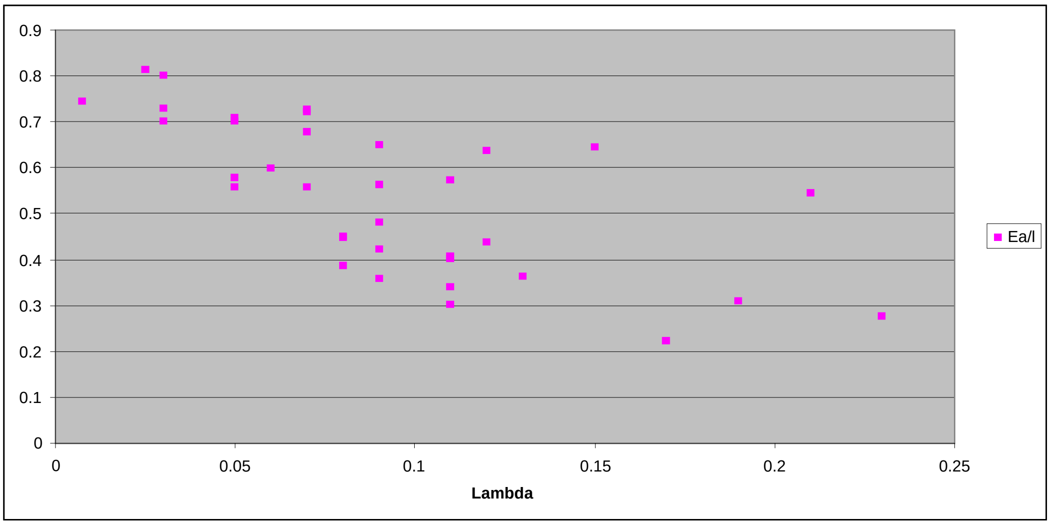

Figure 5.3: Relationship between $\varepsilon_{A/\bar{\lambda}}$ and *λ* for 35 BWC urban districts

In all cases this elasticity is < 1, so that with suburbanisation the proportionate increase in costs is generally less than the proportionate increase in area at the margin. Whether this implies scale economies in the usual sense (higher unit cost) will depend on the relationship between increase in area and increase in numbers of properties. This is best assessed by considering $\varepsilon_{N/\bar{\lambda}}$, as is done next.

$$\varepsilon_{N/\bar{\lambda}} = \beta_2 . \frac{R}{\varphi}$$

The values for *R/φ* observed in the 35 BWC urban districts' data range between about 1.6 and 2.4[67]. In conjunction with the estimated value for β_2 of 0.617 from (5.22) above, this gives values for $\varepsilon_{N/\bar{\lambda}}$ in the range 0.99 to 1.48, indicating roughly constant returns to scale for less dispersed districts but decreasing returns to scale for the more dispersed districts (See **Figure 5.4**)

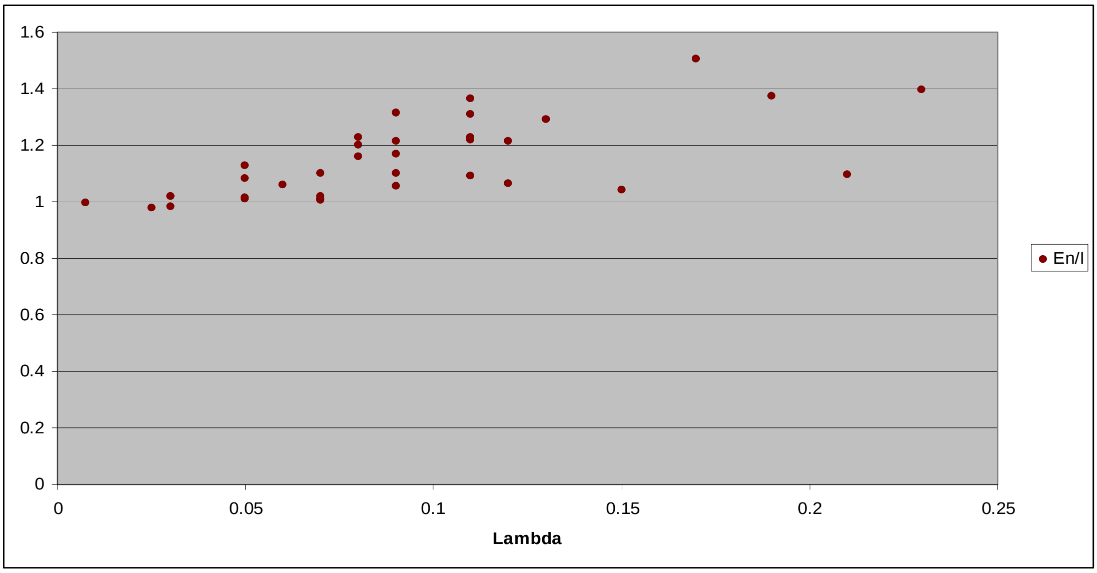

Figure 5.4: Relationship between $\varepsilon_{N/\bar{\lambda}}$ and λ for 35 BWC urban districts

Before moving on, attention needs to be given to capital costs, which are very substantial in water distribution. In the case of BWC, capital costs (on an annualised basis) are made up of £76.6m of capital maintenance expenditure plus £68.5m return on regulatory value (6.4% on an estimated capital value of £1,070.49m – see **Appendix F**), making £145.1m in all. Allocating this amount to urban districts in proportion to length of mains provides a value for the capital cost of distribution (*CCD*). Using this as a measure of capital input, a regression parallel to (5.20) then gave:

$$\ln CCD = -10.65 + 1.617^{**} \ln w + 0.622^{***} \ln \psi \quad …….. \quad (5.21)$$

(S.E. 0.299)   (S.E. 0.032) robust   (R^2 = 0.8981)

[67] The minimum value for *R/φ* is 1.5 as *φ = 2R/3* when *λ = 0*.

It might have been expected that the influence of ψ on capital costs would be larger than on variable costs, but this result indicates a similar value, so that the elasticities $\varepsilon_{A/\bar{\lambda}}$ and $\varepsilon_{N/\bar{\lambda}}$ are about the same. The influence of *w* on the other hand appears very large, indicating that consumption per property has a strong effect on capital requirements, to the extent that there are scale diseconomies, with $\varepsilon_w = \beta_1 = 1.617$.

It had been hoped that that the urban districts identified above would turn out to match well with water production facilities operated by BWC so that production and distribution would be found to be largely self-contained within these districts, facilitating analysis of the interaction between production and distribution at this level. However, this turns out not to be generally the case. More commonly, because water is supplied to consumers from several sources (presumably for security of supply and water quality reasons), self-containment is only evident for rather larger areas. This will affect the applicability of the above results to distribution in these larger areas. Discussion of this issue is deferred to **Chapter VI**.

4. Application to Ofwat data for 11 WOCs

Application of the methods developed in this chapter to all the water companies in England and Wales would be inappropriate as many of them are very large, serving large numbers of settlements. They are therefore far from matching the kind of distribution systems modelled in **Chapter III**, on which (5.2) is based. However, the WOCs operate on a smaller scale and a number of them appear to serve a single large urban area (albeit including some smaller satellite towns and villages), e.g. Bristol Water, Cambridge Water, Dee Valley Water (Chester), Folkestone & Dover Water, Portsmouth Water. An exception is Three Valleys & North Surrey, the largest WOC, whose supply area straggles over several parts of outer London and is therefore far from being a single settlement company. It was therefore decided to omit it from the analysis. For the remaining 11 companies, it seemed worth testing whether they might show similar characteristics to those found for BWC's zones and urban districts. It should be emphasised at the outset that with only 11 cases, the statistical significance of the results is bound to be weak .

With these caveats in mind, data for these 11 WOCs was assembled and the relationships (5.20) and (5.21) were estimated. First however the relationships (5.22) and (5.23) below were estimated to check for consistency with the Leontief-type production function, which was found to be the case as the coefficients on ln*DO* in (5.22) and (5.23) are not significantly different:

$$\ln VCD = -3.791 + 0.625^{***}\ln DO \quad \ldots\ldots\ldots \quad (5.22)$$
$$\textit{(0.064) robust} \quad (R^2 = 0.9251)$$

$$\ln CMD = -2.694 + 0.581^{***}\ln DO \quad \ldots\ldots\ldots \quad (5.23)$$
$$\textit{(0.084) robust} \quad (R^2 = 0.8832)$$

The regressions matching (5.20) and (5.21) then gave:

$$\ln VCD = -3.213 - 0.211\ln w + 0.659^{***}\ln\psi \quad \ldots\ldots. \quad (5.24)$$
$$\textit{(S.E. 0.342)} \quad \textit{(S.E. 0.049) robust} \quad (R^2 = 0.9574)$$

$$\ln CCD = -0.657 - 0.453\ln w + 0.624^{***}\ln\psi \quad \ldots\ldots\ldots\ldots \quad (5.25)$$
$$\textit{(S.E. 0.394)} \quad \textit{(S.E. 0.057) robust} \quad (R^2 = 0.9378)$$

In comparison with the results obtained for BWC's 35 urban districts, it may be seen that in (5.24) the coefficient on lnψ is well-determined although somewhat lower but that the coefficient on ln*w* has turned negative[68] (although not significantly different from zero as the standard error is large). Of course, the sample size is small with only 11 cases, but on this evidence, there is some corroboration of the picture found in the BWC case.

The indications from (5.25) are similar. The influence of distance (ψ) on capital costs is a little lower (and close to the BWC value), while that of volume is again negative (but not significant). The broad conclusion that capital costs for WOCs are driven almost entirely by distance to properties (and hence length of mains) does not seem unreasonable.

As with the BWC urban districts, the coefficients from (5.24) can be used to calculate the "suburbanisation" elasticities $\varepsilon_{A/\bar{\lambda}}$ and $\varepsilon_{N/\bar{\lambda}}$. The resulting values are plotted in **Figure 5.5** below. As may be seen, the values for $\varepsilon_{N/\bar{\lambda}}$ are all above 1, indicating suburbanisation diseconomies. The values for $\varepsilon_{A/\bar{\lambda}}$ are lower than for the BWC urban districts but this is probably a reflection of the larger size (and lower density) of the

[68] The reasons for this have not been determined but it could possibly be due to lower distribution costs for large industrial customers.

WOCs. As with the BWC estimates, higher values for $\varepsilon_{N/\bar{\lambda}}$ and lower values for $\varepsilon_{A/\bar{\lambda}}$ are associated with higher λ.

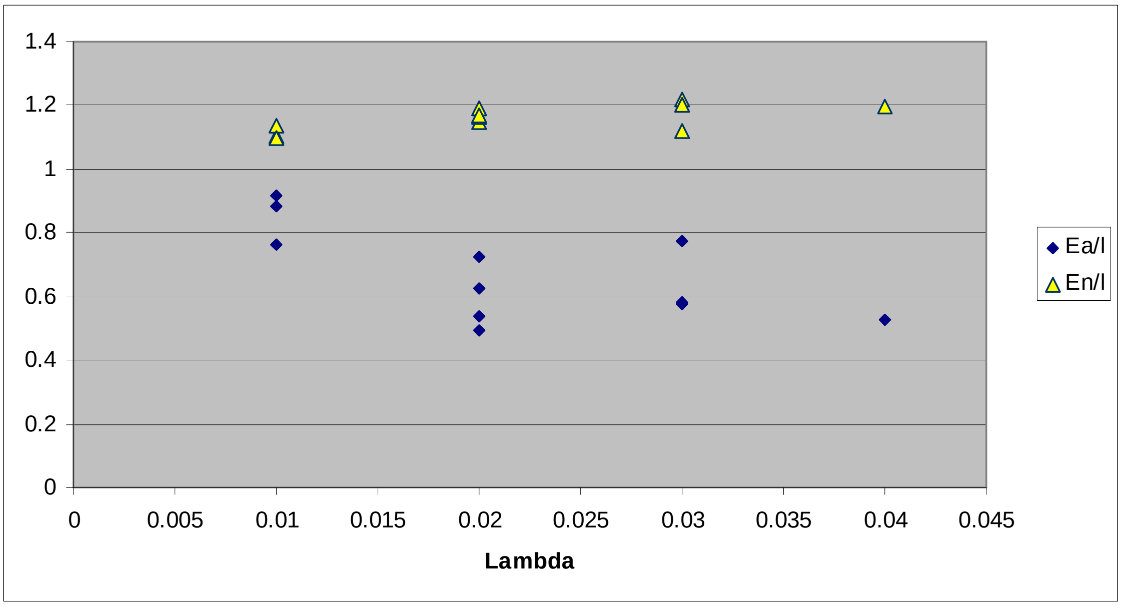

Figure 5.5: Relationship between $\varepsilon_{A/\bar{\lambda}}$, $\varepsilon_{N/\bar{\lambda}}$ and λ for 11 WOCs

5. Application to AWWA data

a. Data issues

The information collected by the American Water Works Association (AWWA) in its 1996 survey does not provide information on capital maintenance costs so that it is not possible to assess how far the assumption of a Leontief-type production function represented by (5.4) corresponds to the actual situation. On the other hand, as most US water undertakings are relatively small scale, with each undertaking generally serving a single settlement or community, the situation is thus often close to that envisaged in the models developed in **Chapter III** above.

Despite the large size of the AWWA sample, there are a number of problems with the data. Many smaller utilities did not respond to the 1996 survey. Of the 897 utilities that did respond, only 548 provided information for the annual O&M expenses table and in some of these cases some of the data items were missing. A further issue is that in the USA, it is quite common for water utilities to sell water in bulk to other utilities. In the AWWA data, this appears as an estimated figure for “wholesale population” rather than a volume. Because of uncertainty about what distribution costs might be associated with these sales, attention was focused on utilities serving “retail populations” only. After the other adjustments shown in **Table 5.1**, this left 305 cases.

Reason for dropping cases	Numbers affected
Starting point: Utilities in O&M table	548
No figure for water produced	-10
No figure for distribution costs	-22
No figure for length of mains	-16
No retail population	-6
Effective radius > 10 km	-11
Density > 30 properties/ Ha	-3
Usable cases	**480**
Utilities serving wholesale as well as retail populations	-175
Retail only utilities	**305**

Table 5.1: Selection of distribution cases from AWWA 96 data

As with the BWC data, the effective service area was represented by length of mains (converted from miles to km) divided by 0.15, except for a few cases where this value was greater than the service area reported by the company, when the latter figure was used[69]. Population numbers were divided by 2.25 to provide an estimate of numbers of properties, and then property density and service area radius were calculated for each utility in relation to the effective service area. Information on leakage rates is not included in the AWWA data so the quantity variable is water put into distribution (*QDI*) rather than water consumed (*QC*).

The data is illustrated in **Figure 5.6**:

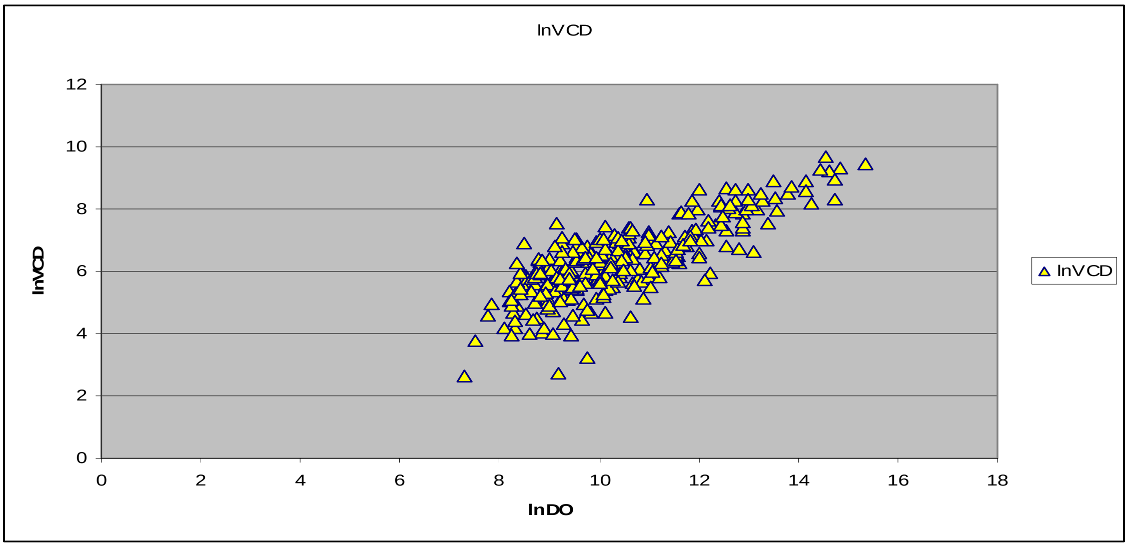

Figure 5.6: Log plot of distribution output against distribution costs for 305 US retail only water utilities

[69] In a few cases, no service area was reported so that the value derived from length of mains was the only one available. (All areas were converted from square miles to hectares for consistency with the earlier analyses.)

b. Results

Repeating the regressions (5.5) and (5.6) with the 305 AWWA retail only cases (and using *QDI* rather than *QC* because information on leakage is lacking in the AWWA data) gave the results shown in (5.26) and (5.27):

$$\ln VCD = 0.03 + 0.605{***}\ln DO \qquad \ldots\ldots\ldots\ldots \qquad (5.26)$$

(S.E. 0.026) *(R^2 = 0.6464)*

$$\ln VCD = 0.193 + 0.489{***}\ln QDI + 0.885{***}\ln\varphi \qquad \ldots\ldots\ldots\ldots \qquad (5.27)$$

(S.E. 0.091) *(S.E. 0.211)* *(R^2 = 0.6484)*

These results are encouragingly similar to those obtained earlier for BWC's 35 urban districts. However, measuring volume as *QDI* rather than *QC* will have affected the coefficients but probably not to a large extent[70].

Re-running (5.27) in (5.8) form then gave:

$$\ln VCD = -5.253 + 0.369{***}\ln W + 0.629{***}\ln\psi \qquad \ldots\ldots\ldots\ldots \qquad (5.28)$$

(S.E. 0.096) *(S.E. 0.027)* *(R^2 = 0.6538)*

As with the earlier results, certain elasticities can be estimated from (5.28). First, from the coefficient on ln*W*, we have ε_W = 0.369, giving returns to scale of 2.7 for consumption per property (note that *W* here is water put into distribution per property, i.e. consumption *plus* leakage). From the coefficient on lnψ, values for the "suburbanisation" elasticities $\varepsilon_{A/\bar{\lambda}}$ and $\varepsilon_{N/\bar{\lambda}}$ can be calculated. They are plotted in **Figure 5.7**, and can be seen to have the same characteristics as were found for BWC urban districts in **Figures 5.3** and **5.4**, and for the WOCs in **Figure 5.5**.

[70] Re-running (5.17) for BWC urban districts using *QDI* in place of *QC* produced $\ln VCD = 1.791 + 0.365\ln QDI + 1.169\ln\phi$ so that in this case the coefficients are changed by less than 10%.

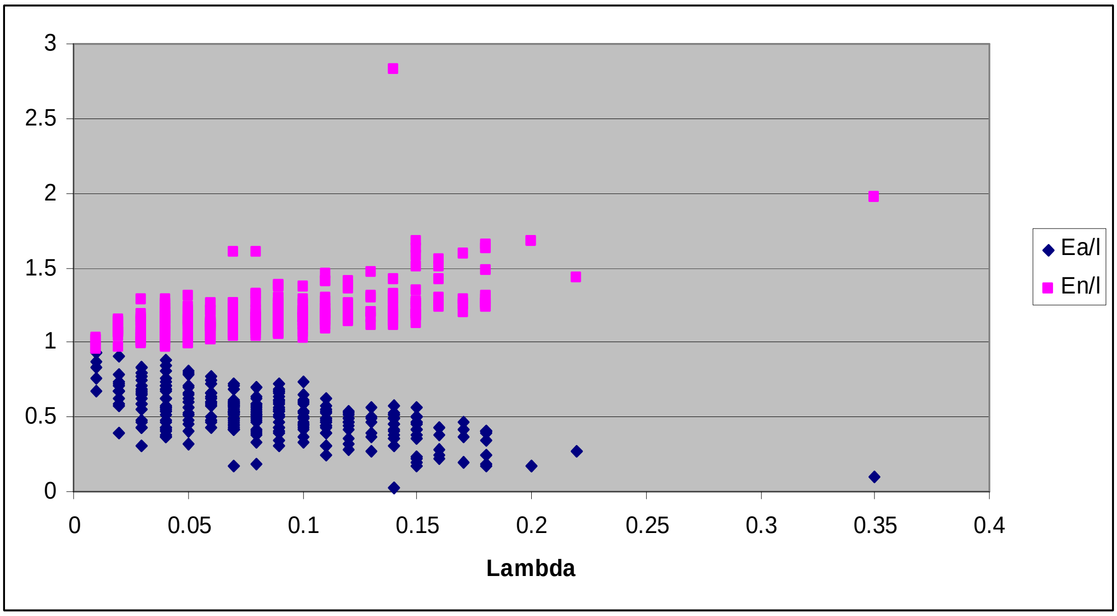

Figure 5.7: Relationship between $\varepsilon_{A/\bar{\lambda}}$, $\varepsilon_{N/\bar{\lambda}}$ and λ for 305 US retail only utilities

6. Implications of results

In this section, the estimated relationships for distribution costs obtained in **sections 3, 4** and **5** above are used to carry out illustrative calculations for settlements or companies with different distribution characteristics. These show that distribution costs depend strongly on the spatial configuration of the distribution area. Thus, although there are differences of detail, the calculations all agree that with a monocentric structure "*densification*" reduces unit distribution costs whereas greater dispersion of properties (higher λ) raises them. The calculations also suggest that more properties (higher *N*) with λ held constant ("*suburbanisation*") would also raise distribution costs but to a much smaller extent. (The implication of higher *N* with λ held constant is lower density and a larger settlement area.) With density rather than λ held constant, more properties lead to lower unit distribution costs. The story is a bit more complicated if there is more than one settlement in the service area (polycentric structure). The details and results of these calculations are set out below using first the estimated relationships for the 35 BWC urban districts, then the estimated relationships for the 11 WOCs and finally the estimated relationships for the 305 US retail only utilities.

a. Basis for calculations

Following the schema in **Figure 5.2**, four kinds of illustrative calculations are presented:

a. ***"Densification"***: The effect on distribution costs of varying the number of properties (*N*), holding settlement radius (*R*) constant;

b. ***“Dispersion”*:** The effect on distribution costs of varying λ, *N* held constant;

c. ***“Suburbanisation”*:** The effect on distribution costs of varying *N*, λ held constant;

d. ***“Constant density”***: The effect on distribution costs of varying *N*, density held constant.

The results are expressed as unit costs as the implications are most easily appreciated in this form.

For (a), the steps in the calculation are:

i. Take *N* to be 18,000 properties for a typical BWC urban district[71], 200,000 properties for a typical WOC and 50,000 properties for a typical US retail water utility. These are roughly the average values observed in the 3 data sets;

ii. Use the relationship (5.11) $N = \frac{2\pi . d_0}{\lambda^2}\left[1 - e^{-\lambda R}(1 + \lambda R)\right]$ between *N*, d_0, λ and *R* to estimate λ for each value of *R* (d_0 = 30 properties/Ha in all cases);

iii. Assume *w* to be 420 litres/property/day for BWC, 520 litres/property/day for the WOCs and 1500 litres/property/day in the US, which are approximately the values observed in the data sets. (Note that the US figure includes distribution losses, whereas the others do not – but even allowing for this difference, consumption per property in the USA still appears to be about twice what it is in England & Wales.);

iv. Use (5.9) to calculate ψ;

v. Calculate ln*VCD* using (5.20), (5.24) or (5.28) as appropriate. Convert the result to £/Ml (England & Wales) or $/million US gallons to give the required unit costs (*UVCD*);

vi. Calculate ln*CCD* using (5.21) and (5.25) to derive unit capital costs (*UCCD*) for BWC urban districts and the WOCs respectively.

For (b), the procedure is very similar but at step (ii) the relationship is used to estimate *R* for each value of λ. For (c), λ is held constant at step (ii) while *N* is varied; for (d), the starting point is an assumed density, which when combined with varying *N* leads to changes in the values for *R* and λ at this step.

[71] This average excludes the largest urban district which has some 600,000 properties.

b. Calculations for 35 BWC "urban districts"

The calculations using the estimated relationships for these districts lead to the figures shown in **Table 5.2**. The numbers to focus on are in the last 5 columns, where *VCD* and *CCD* are respectively the annual variable and capital costs of distribution, *UVCD* and *UCCD* are the related unit costs and *UTCD* is the total unit cost.

N	λ	***R***	$\varphi(\lambda,R)$	***VCD***	***UVCD***	***CCD***	***UCCD***	***UTCD***
		('00m)		(£m)	(£/Ml)	(£m)	(£/Ml)	(£/Ml)
a. Varying *N*, *R* constant ('densification')								
5,000	0.19	26.8	9.7	0.109	142.23	0.339	441.90	584.13
10,000	0.12	26.8	12.5	0.196	127.94	0.612	399.40	527.34
15,000	0.095	26.8	13.6	0.266	115.59	0.832	361.73	477.32
20,000	0.075	26.8	14.6	0.331	107.81	1.036	337.99	445.79
25,000	0.06	26.8	15.3	0.390	101.86	1.225	319.76	421.62
40,000	0.03	26.8	16.6	0.550	89.64	1.730	282.17	371.81
50,000	0.015	26.8	17.3	0.646	84.24	2.035	265.54	349.78
b. Varying λ, *N* constant ('dispersion')								
18,000	0	13.8	9.2	0.233	84.60	0.730	264.48	349.08
18,000	0.02	15.3	9.9	0.245	88.63	0.765	277.17	365.80
18,000	0.04	17.3	10.8	0.258	93.53	0.807	292.62	386.15
18,000	0.06	20.3	12.1	0.276	99.91	0.863	312.74	412.65
18,000	0.08	25.8	13.9	0.301	109.22	0.944	342.13	451.35
18,000	0.10	48.7	18.1	0.354	128.30	1.111	402.45	530.75
c. Varying *N*, λ constant ('suburbanisation')								
5,000	0.06	8.6	5.5	0.077	100.27	0.238	310.64	410.91
10,000	0.06	13.3	8.2	0.152	98.97	0.473	308.31	407.28
15,000	0.06	17.7	10.7	0.229	99.44	0.715	310.80	410.24
20,000	0.06	22.0	12.9	0.307	100.16	0.962	313.82	413.98
25,000	0.06	26.8	15.3	0.390	101.86	1.225	319.76	421.62
40,000	0.06	46.2	22.9	0.669	109.14	2.110	344.10	453.24
50,000	0.06	81.1	30.1	0.911	118.88	2.880	375.76	494.64
d. Varying *N*, density=10 ('constant density')								
5,000	0.15	12.6	7.0	0.089	116.17	0.276	360.33	476.50
10,000	0.1	17.8	10.0	0.170	111.18	0.531	346.67	457.85
15,000	0.08	21.9	12.2	0.249	108.15	0.778	338.25	446.40
20,000	0.07	25.2	14.1	0.324	105.72	1.016	331.36	437.08
25,000	0.065	28.2	15.7	0.397	103.48	1.245	324.87	428.35
40,000	0.05	35.7	19.9	0.615	100.27	1.937	315.92	416.19
50,000	0.045	39.9	22.2	0.755	98.54	2.384	310.99	409.53

Table 5.2: Illustrative calculations to show the effect of different values of λ and *N* on unit distribution costs (using relationships estimated for BWC urban districts)

Section (a) of **Table 5.2** shows how adding properties within a fixed urban boundary substantially reduces unit distribution costs. This is because volume economies of scale in distribution outweigh the effect of a small increase in dispersion as measured by φ. Section (b), on the other hand, shows that for a settlement of a given size in terms of numbers of properties, greater dispersion leads to diseconomies in distribution. In this case, although the number of properties (and hence total consumption) does not change, higher λ leads to a larger service area with distribution costs rising by 50% as λ rises from zero to 0.1. These two cases provide good illustrations of density economies in distribution, as in both cases higher density leads to lower distribution costs. In section (c), increasing the number of properties with λ constant results at first in economies of scale with respect to volume more or less offsetting the effect of greater dispersion, although above 10,000 properties, the latter effect increasingly dominates, leading again to diseconomies in distribution. In contrast, section (d), which compares settlements of similar density but different size, shows scale economies, particularly in capital costs. In this case, although more properties result in a larger radius settlement, this is accompanied by reduction in λ and hence less dispersion, leading to savings in the unit cost of distribution.One way of viewing the section (c) figures is as showing the effect of extending water supply from an urban core first to the suburbs and then to a rural fringe. The first 10,000 properties (the urban core) occupy only about 556 Ha at an average density of 18.0 properties/Ha. The next 15,000 properties (the suburbs) occupy about 1700 Ha (average density 8.8 properties/Ha). The next 15,000 properties (the rural fringe) occupy about 4450 Ha (average density 3.4 properties/Ha); and another 10,000 properties would add about 14,000 Ha at an average density of 0.7 properties/Ha. The effect on distribution costs is plotted in **Figure 5.8** below. Compared with the total unit cost of distribution in the urban core, £407/Ml, adding the suburbs raises this cost by about 4% to £422/Ml; adding the rural fringe adds another 7% bringing the cost to £453/Ml; with the outer fringe (bringing the total number of properties to 50,000) the cost rises further to £495/Ml, over 20% above the figure for the urban core alone. Clearly, the marginal cost of distribution to these more remote and highly dispersed properties is high[72].

[72] For the last 10,000 properties, the unit cost is £660/Ml, some 60% higher than the £408/Ml unit cost for the 10,000 properties in the urban core.

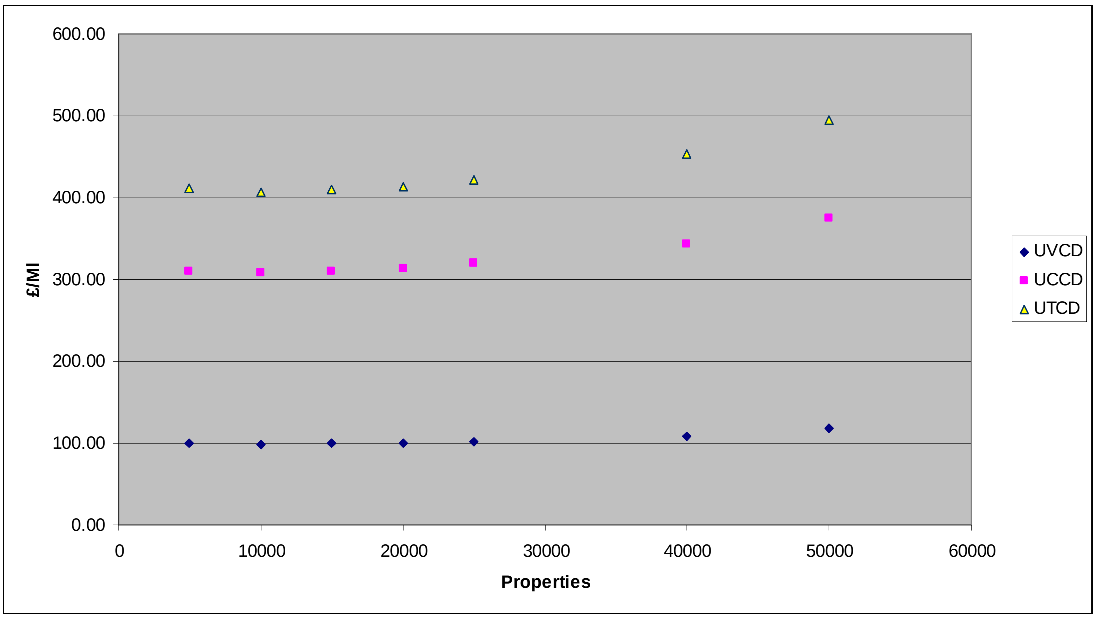

Figure 5.8: Effect of increasing settlement size with constant λ ("*suburbanisation*") (from section (c) of Table 5.2)

c. Calculations for 11 WOCs

Similar calculations were then carried out for the 11 WOCs. Although the estimated relationships differ somewhat the pattern of the results is very similar, as may be seen in **Table 5.3**.

N	λ	R('00m)	$\varphi(\lambda,R)$	*VCD* (£m)	*UVCD* (£/Ml)	*CCD* (£m)	*UCCD* (£/Ml)	*UTCD* (£/Ml)
a. Varying *N*, *R* constant ('densification')								
50,000	0.055	57.3	27.0	1.181	119.81	2.761	290.98	410.79
100,000	0.03	57.3	32.2	2.160	109.57	4.727	249.06	358.63
150,000	0.02	57.3	34.3	2.987	101.03	6.310	221.63	322.66
200,000	0.01	57.3	36.3	3.792	96.20	7.804	205.58	301.78
250,000	0.005	57.3	37.3	4.507	91.47	9.101	191.81	283.28
b. Varying λ, *N* constant ('dispersion')								
200,000	0	46.1	30.7	3.377	85.66	7.038	185.40	271.06
200,000	0.01	55.1	35.0	3.697	93.79	7.630	200.99	294.78
200,000	0.02	72.4	42.0	4.193	106.36	8.534	224.80	331.16
200,000	0.03	162.5	60.3	5.392	136.78	10.675	281.22	418.00
c. Varying *N*, λ constant ('suburbanisation')								
50,000	0.025	29.2	17.5	0.900	91.33	2.169	228.52	319.85
100,000	0.025	47.5	26.4	1.974	100.17	4.365	229.95	330.12
150,000	0.025	66.8	34.3	3.193	108.01	6.696	235.21	343.22
200,000	0.025	91.0	42.0	4.596	116.59	9.260	243.95	360.54
250,000	0.025	128.0	49.7	6.259	127.02	12.191	256.92	383.94
d. Varying *N*, density=10 ('constant density')								
50,000	0.045	39.9	22.2	1.033	104.83	2.452	258.36	363.19
100,000	0.0325	56.4	31.3	2.117	107.41	4.644	244.69	352.10
150,000	0.025	69.1	38.8	3.254	110.07	6.810	239.20	349.27
200,000	0.0225	79.8	44.5	4.365	110.73	8.845	233.01	343.74
250,000	0.02	89.2	49.8	5.510	111.81	10.883	229.35	341.16

Table 5.3: Illustrative calculations to show the effect of different values of λ and *N* on unit distribution costs (using relationships estimated for 11 WOCs)

d. Calculations for 305 US retail only utilities

The pattern of results for the 305 retail only US utilities is also similar (see **Table 5.4**) although in this case, no data is available to enable capital costs to be estimated (at the same time there is no reason to suppose that distribution capital costs in the US would not also be 2 or 3 times as large as operating costs and would follow a similar pattern).

N	**λ**	***R*****('00m)**	***φ(λ,R)***	***VCD*** $m	***UVCD*** $/m.galls[73]	***CCD***	***UCCD***	***UTCD***
a. Varying *N*, *R* constant ('densification')								
10,000	0.13	30.2	12.8	0.129	89.24	n.a.	n.a.	n.a.
25,000	0.07	30.2	16.2	0.267	73.73	n.a.	n.a.	n.a.
50,000	0.03	30.2	18.5	0.448	62.00	n.a.	n.a.	n.a.
75,000	0.0065	30.2	19.8	0.603	55.62	n.a.	n.a.	n.a.
b. Varying λ, *N* constant ('dispersion')								
50,000	0	23.0	15.3	0.391	54.06	n.a.	n.a.	n.a.
50,000	0.02	27.4	17.4	0.423	58.54	n.a.	n.a.	n.a.
50,000	0.04	36.2	21.0	0.476	65.86	n.a.	n.a.	n.a.
50,000	0.06	81.1	30.1	0.598	82.71	n.a.	n.a.	n.a.
c. Varying *N*, λ constant ('suburbanisation')								
10,000	0.04	12.1	7.7	0.092	63.85	n.a.	n.a.	n.a.
25,000	0.04	21.5	13.3	0.231	63.78	n.a.	n.a.	n.a.
50,000	0.04	36.2	21.0	0.476	65.86	n.a.	n.a.	n.a.
75,000	0.04	54.1	28.9	0.751	69.25	n.a.	n.a.	n.a.
d. Varying *N*, density=10 ('constant density')								
10,000	0.1	17.8	10.0	0.108	74.87	n.a.	n.a.	n.a.
25,000	0.063	28.2	15.8	0.257	71.13	n.a.	n.a.	n.a.
50,000	0.045	39.9	22.2	0.494	68.31	n.a.	n.a.	n.a.
75,000	0.0325	48.9	27.9	0.735	67.76	n.a.	n.a.	n.a.

Table 5.4: Illustrative calculations to show the effect of different values of λ and *N* on unit distribution costs (using relationships estimated for 305 US retail only utilities)

e. Effect of multiple settlements

The same calculations can be used to throw light on the effect on distribution costs if there are two or more settlements in an area. For this purpose we use the estimated relationships for BWC's 35 urban districts, comparing distribution costs for a monocentric settlement of 50,000 properties with:

- 2 settlements with 25,000 properties;
- 1 settlement of 40,000 properties and 1 settlement of 10,000 properties;
- 5 settlements of 10,000.

In each case, total area (5000 Ha) and average density (10 properties/Ha) are held constant (with λ varying in consequence). These comparisons are set out in **Table 5.5**.

[73] 1 US gallon = 3.786 litres, so $1/m.galls = £0.176/Ml if £1 = $1.5.

N	**λ**	***R*** **(‘00m)**	**Area (Ha)**	***QC=w.N*** **(Ml/d)**	***VCD*** **(£m)**	***UVCD*** **(£/Ml)**	***CCD*** **(£m)**	***UCCD*** **(£/Ml)**	***UTCD*** **(£/Ml)**
a. Single settlement									
50,000	0.045	39.9	5000	21.0	0.755	**98.54**	2.384	**310.99**	**409.53**
b. Two equal settlements									
25,000	0.065	28.2	2500	10.5	0.397	103.48	1.245	324.87	
25,000	0.065	28.2	2500	10.5	0.397	103.48	1.245	324.87	
Total			5000	21.0	0.794	**103.48**	2.490	**324.87**	**428.35**
c. Two unequal settlements									
40,000	0.05	35.7	4000	16.8	0.615	100.27	1.937	315.92	
10,000	0.1	17.8	1000	4.2	0.170	111.18	0.531	346.67	
Total			5000	21.0	0.785	**102.41**	2.528	**329.81**	**432.22**
d. Five equal settlements									
10,000	0.1	17.8	1000	4.2	0.170	111.18	0.531	346.67	
10,000	0.1	17.8	1000	4.2	0.170	111.18	0.531	346.67	
10,000	0.1	17.8	1000	4.2	0.170	111.18	0.531	346.67	
10,000	0.1	17.8	1000	4.2	0.170	111.18	0.531	346.67	
10,000	0.1	17.8	1000	4.2	0.170	111.18	0.531	346.67	
Total			5000	21.0	0.850	**111.18**	2.655	**346.67**	**457.85**

Table 5.5: Calculations to show the effect of multiple settlements on unit distribution costs (using relationships estimated for BWC urban districts)

From **Table 5.5**, it may be seen, comparing (b) with (a), that splitting the population into two equal settlements has the effect of increasing the unit distribution cost by 5% from £409.53 to £428.35; splitting into 2 unequal settlements (c) also increases distribution costs, to a somewhat greater extent, because of a higher capital cost. Splitting into 5 smaller settlements of 10,000 properties each (d) results in a rather larger increase of 12% to £457.85. The reason for these results is that the compensating variation in λ has the effect of increasing the dispersion of properties in the smaller settlements. In consequence, the large single settlement in (a) shows distribution costs which are lower (by about 12%) compared with the five settlements in (d), showing how greater dispersion leads to diseconomies in distribution. On the other hand, if the smaller settlements had the same λ value as the large settlement (0.045 in this case), they would occupy a smaller area in total, at higher density, leading to a small saving in distribution costs.

f. Derived elasticities

The calculated results in **Tables 5.2, 5.3** and **5.4** can be used to derive estimated elasticities corresponding to those discussed above in **Section 2 (c)** of this chapter.

Being estimated from intervals rather than by continuous variation, these values are approximations with uncertain confidence intervals. The values in **Table 5.6** are for an average sized urban district or company from the middle of the range of calculated values, using variable costs (*VCD*)[74]. The elasticities shown are:

(a) ***Densification***: $\varepsilon_{N/\bar{R}}$, the elasticity of costs as the number of properties (*N*) varies, while settlement radius is held constant. If $\varepsilon_{N/\bar{R}} < 1$, there are scale economies;
(b) ***Dispersion***: $\varepsilon_{A/\bar{N}}$, the elasticity of costs as the coefficient of dispersion (λ) varies, holding number of properties constant. If $\varepsilon_{A/\bar{N}} > 0$, there are scale diseconomies;
(c) ***Suburbanisation***: $\varepsilon_{N/\bar{\lambda}}$, the elasticity of costs as the number of properties (*N*) varies, holding λ constant; and the related elasticity $\varepsilon_{A/\bar{\lambda}}$. If $\varepsilon_{N/\bar{\lambda}} > 1$, there are scale diseconomies;
(d) ***Constant density***: $\varepsilon_{N/\bar{D}}$, the elasticity of costs as the number of properties (*N*) varies, holding density (*N/A*) constant (which is equal in value to $\varepsilon_{A/\bar{D}}$). If $\varepsilon_{N/\bar{D}} < 1$, there are scale economies.

	Average BWC urban district	**Average WOC**	**Average US retail utility**
No. of properties	18,000	200,000	50,000
(a) Densification			
$\varepsilon_{N/\bar{R}}$ *(range)*	0.73 *(0.80 – 0.70)*	0.81 *(0.83 – 0.75)*	0.68 *(0.71 – 0.69)*
(b) Dispersion			
$\varepsilon_{A/\bar{N}}$ *(range)*	0.18 *(0.21 – 0.07)*	0.19 *(0.22 – 0.07)*	0.17 *(0.20 – 0.06)*
(c) Suburbanisation			
$\varepsilon_{N/\bar{\lambda}}$ *(range)*	1.03 *(0.97 – 1.45)*	1.32 *(1.19 – 1.45)*	1.07 *(1.00 – 1.16)*
$\varepsilon_{A/\bar{\lambda}}$ *(range)*	0.63 *(0.70 – 0.17)*	0.51 *(0.73 – 0.37)*	0.58 *(0.69 – 0.47)*
(d) Constant density			
$\varepsilon_{N/\bar{D}} = \varepsilon_{A/\bar{D}}$ *(range)*	0.91 *(0.92 – 0.90)*	1.02 *(1.02 – 1.07)*	0.92 *(0.92 – 0.98)*

Table 5.6: Spatial effect elasticities derived from calculated values in Tables 5.2, 5.3 and 5.4

7. Conclusions on scale effects in water distribution

The approach in this chapter to assess scale effects in water distribution breaks new ground in that distribution output is measured as the product of consumption and average distance. Implementation of this approach has required that distribution areas be modelled as monocentric settlements with density declining away from the centre at a

[74] Similar values would be obtained using capital costs (*CCD*) or total costs (*TCD*) because of the similarity of the values for the coefficient on lnψ.

rate consistent with the data for length of mains and numbers of properties[75]. This is better than using raw areas which include (in many cases) significant amounts of unoccupied or unserviced land, while the circular shape is a reasonable way to capture the spatial aspect of distribution.

Bringing together the results for the quantity elasticity ε_w obtained earlier, together with the spatial elasticities from **Table 5.6** as is done in **Table 5.7** below, it can be seen that the findings for BWC urban districts and US retail utilities are consistent. There are quite large scale economies with respect to consumption per property in water distribution (returns to scale of about 1/0.4 = 2.5). Among the spatial elasticities, densification and constant density expansion are also characterised by scale economies (returns to scale about 1/0.7 = 1.4 and 1/0.9 = 1.1 respectively). On the other hand there are diseconomies associated with dispersion and suburbanisation. The WOC results are in reasonable agreement as regards densification and dispersion but show higher diseconomies for suburbanisation and (small) diseconomies for constant density – possibly a reflection of the relatively large size of the WOCs so that there are a number of subsidiary settlements around the main centre.

[75] While exponential decline in density is the standard assumption in the urban literature, other specifications are possible, e.g. a bell-shaped curve based on the normal distribution might better capture the actual density gradient of some settlements. Whether this is the case has not been investigated in this research but it is considered that the character of the results obtained using a different specification would not be very different from those reported here.

	Average BWC urban district	**Average WOC**	**Average US retail utility**[a]
No. of properties	18,000	200,000	50,000
1. Quantity effect			
ε_W *(S.E.)*	0.43 *(0.23)*	-0.21 *(0.34)*	0.37 *(0.10)*
2. Spatial effects			
(a) Densification			
$\varepsilon_{N/\bar{R}}$ *(range)*	0.73 *(0.80 – 0.70)*	0.81 *(0.83 – 0.75)*	0.68 *(0.71 – 0.69)*
(b) Dispersion[b]			
$\varepsilon_{A/\bar{N}}$ *(range)*	0.18 *(0.21 – 0.07)*	0.19 *(0.22 – 0.07)*	0.17 *(0.20 – 0.06)*
(c) Suburbanisation			
$\varepsilon_{N/\bar{\lambda}}$ *(range)*	1.03 *(0.97 – 1.45)*	1.32 *(1.19 – 1.45)*	1.07 *(1.00 – 1.16)*
$\varepsilon_{A/\bar{\lambda}}$ *(range)*	0.63 *(0.70 – 0.17)*	0.51 *(0.73 – 0.37)*	0.58 *(0.69 – 0.47)*
(d) Constant density			
$\varepsilon_{N/\bar{D}} = \varepsilon_{A/\bar{D}}$ *(range)*	0.91 *(0.92 – 0.90)*	1.02 *(1.02 – 1.07)*	0.92 *(0.92 – 0.98)*

[a] In this case the volume variable was ln*QDI*.

[b] For this elasticity, a value > 0 implies diseconomies.

Table 5.7: Comparison of distribution cost elasticities across three data sets

Which effect then predominates depends on the spatial characteristics of the distribution area. The implications can be seen in **Tables 5.2**, **5.3** and **5.4** and **Figure 5.8**. Sections (a) and (b) of the tables bring out the benefits of higher densities in terms of lower unit distribution costs. **Figure 5.8**, on the other hand, provides a good illustration of the diseconomies associated with extending supply to the lower density periphery of an urban area. The illustrative calculations in **Table 5.5** are also of interest. The higher distribution costs incurred when an area is occupied by smaller more dispersed settlements draws attention to one aspect of the interaction between production and distribution costs being investigated in this thesis. If each settlement operates its own water production facilities, it risks a double cost penalty, on the production side from smaller plant size and on the distribution side from greater dispersion.

In **Chapter VI** we move on to examine interactions of this kind more systematically in the light of the results obtained in this chapter and in **Chapter IV**.

VI. BRINGING TOGETHER WATER PRODUCTION AND DISTRIBUTION: THE VOLUME/SPACE TRADE-OFF IN URBAN WATER SUPPLY

1. Introduction

The purpose of this chapter is to use the results obtained in **Chapters IV** and **V** to examine the interaction between water production and distribution costs, to see what the implications are for scale effects. A comparison can then be made with results obtained when production and distribution are not treated separately, and with the findings of other researchers. Thus **Section 2** examines these implications using the relationships obtained with BWC data, applied first to 35 urban districts, then to 31 water supply areas and finally to a case where 4 towns are supplied by a single works. **Section 3** then takes the relationships estimated with AWWA data, using them first to compare the effect of estimating production and distribution separately with joint estimation, and then in a comparison with the results obtained by Torres & Morrison Paul (2006). **Section 4** moves on to carry out a similar exercise using Ofwat data for 10 WOCs, comparing the results with those obtained by Stone & Webster Consultants (2004).

Generally, these investigations indicate that there are volume economies of scale in water distribution as well as in water production. However, the ability to exploit economies of scale in water production is constrained in practice by the capacity and location of suitable water resources and by the often small size of settlements. Density effects also need to be taken into account, with low density adding substantially to distribution costs. An important feature of the situation, conditioning these results, is that water suppliers generally have to take the size and location of the settlements they serve as given. They are not able to pursue cost savings by organizing the merger or relocation of small towns or awkwardly located customers; and it is unlikely even in the longer term, that differential water supply costs have much effect on the evolution of settlement patterns.

By way of background, it is helpful to have a feel for the size and density of urban settlements. A starting point is provided by the ONS 2001 Census statistics for urban areas[76]. Here we take the figures for the South East region of England in 2001 to illustrate certain general features. They show:

Size band (Popn)	No of UAs	Av. Area (Ha)	Av. No of properties[77]	Av density (Props/Ha)	Density (range) (Props/Ha)
> 1 million	1	161,724	3,147,171	19.46	19.46
500k – 1 m	0	-	-	-	-
100k – 500k	17	5,197	81,774	15.51	11.09 – 21.28
50k – 100k	21	1,895	28,351	15.01	11.36 – 19.78
20k – 50k	47	858	12,212	14.71	8.49 – 26.19
10k – 20k	52	470	5,932	13.49	5.80 – 18.98
5k – 10k	66	211	2,699	13.78	7.16 – 20.98

Table 6.1: Size and density of urban areas, SE England 2001

Apart from the overwhelmingly dominant position of the Greater London UA, this demonstrates the relatively small size of most English settlements, even in the South East. It also shows that average density varies much less between size bands than it does within each size band. It appears that within each size band there is a range of configurations.

A somewhat different picture emerges if the 54 "urban districts" formed by combining zones within the BWC supply area to better match ONS urban areas and their peripheries are examined. It is evident from **Table 6.2** that even the most densely populated of these urban districts must include large areas of non-urban land:

Size band (Properties)[78]	No of urban districts	Av. Area (Ha)	Av. No of properties	Av density (Props/Ha)	Density (range) (Props/Ha)
>200k	3	73,846	368,759	4.99	2.47 -9.06
50k – 200k	10	28,046	97,212	3.47	2.37 – 5.46
20k – 50k	17	30,656	29,560	0.96	0.16 – 5.36
10k – 20k	11	20,133	14,815	0.74	0.34 – 4.51
5k – 10k	9	8,948	7,637	0.85	0.23 – 3.27
<5k	4	6,879	3,206	0.47	0.39 – 2.60

Table 6.2: Size and density of "urban districts", BWC supply area 2004

Similarly, average property densities for whole company areas in England & Wales, which include all land within the company boundary, whether urban or non-urban, give

[76] ONS (2004). In this report, "urban areas" are areas of built up land of at least 20 Ha, with a population of 1,500 or more.

[77] The ONS counts "household spaces" which are more numerous than household "properties" counted by water companies because of properties comprising more than one household. On the other hand, the ONS figures exclude commercial and industrial properties.

[78] Note that in this table, unlike **Table 6.1**, the size bands are numbers of properties, not population.

rather low property densities, ranging from 0.65 properties/Ha (Welsh Water) up to 4.24 properties/Ha (Thames Water), as shown in **Table 6.3**:

Water only companies (WOCs)				**Water and sewerage companies (WaSCs)**			
Company[79]	**Area ('000 Ha)**	**Props ('000)**	**Density (Props/Ha)**	**Company[79]**	**Area ('000 Ha)**	**Props ('000)**	**Density (Props/Ha)**
BWH	104.1	188	1.81	**ANH**	2,209.0	1,930	0.87
BRL	239.1	483	2.02	**WSH**	2,040.0	1,317	0.65
CAM	117.5	120	1.02	**YKY**	1,424.0	2,109	1.48
DVW	83.1	117	1.41	**NES**	1,184.3	1,899	1.60
FLK	42.0	72	1.71	**SWT**	1,030.0	726	0.70
MKT	205.0	242	1.18	**SVT**	1,974.5	3,279	1.66
PRT	86.8	290	3.34	**SRN**	445.0	1,007	2.26
MSE	360.7	590	1.64	**TMS**	820.0	3,474	4.24
SST	150.7	548	3.64	**NWT**	1,441.5	3,120	2.16
SES	83.3	270	3.24	**WSX**	735.0	537	0.73
THD	35.2	70	1.99				
TVW	372.7	1,224	3.28				

Table 6.3: Property densities for whole company areas, England & Wales, 2003

2. Bringing water production and distribution together: (i) BWC

a. 35 "urban districts"

It may be recalled from **Chapter V** that 35 "urban districts" within the BWC supply area were selected for analysis because they seemed to provide a reasonable approximation to the kind of monocentric settlement envisaged in our distribution model. Ideally, to assess the effect of bringing together water production and water distribution, one would use direct information about the relevant costs for each of the 35 districts. However, BWC's supply arrangements are mostly not self-contained within these districts[80]. Instead, to calculate water production costs, it is assumed that in each case water production is from a single WTW of the appropriate size, using the parameters obtained using (4.20) in **Table 4.9**, and assuming level 4 treatment[81]. Illustrative cost calculations for hypothetical settlements of varying sizes and densities can then be carried out for the same scenarios as in **Chapter V, section 6(b)** ("*densification*", "*dispersion*", "*suburbanization*" and "*constant density*"), with distribution costs taken directly from **Table 5.2**.

[79] For key to company acronyms, see **Tables 3.1A** and **3.1B** in **Chapter III**.

[80] **Section 2(i)(b)** below will present results for distribution areas within which production and distribution are largely self-contained, although these areas no longer approximate monocentric settlements.

[81] These parameters are for total production costs, including capital costs.

Thus, for water production, starting from (4.20):

$$UCP = (1 + W4P)^{\gamma} \left\{ PHR^{\delta} \beta_B \left\lfloor \sum_i p_{B\,i} \left(AQP_{B\,i} \right)^{\alpha_B - 1} \right\rfloor + \beta_T \left\lfloor \sum_i p_{T\,i} (AQP_{Ti})^{\alpha_T - 1} \right\rfloor \right\} \quad \ldots\ldots \quad (6.1)$$

With the parameters from the last column of **Table 4.9**, if there are no boreholes and only one WTW, with *W4P* = 1, the average (or unit) cost (£/Ml) of production for a WTW producing *QP* Ml/day can be calculated as:

$$UCP = 2^{0.31}.474.QP^{-0.24} \qquad \ldots\ldots\ldots.. \qquad (6.2)$$

If, in addition, for the purposes of these illustrative calculations, a leakage rate of 20% is assumed, then:

$$QP = QC / 0.8 \qquad \ldots\ldots\ldots\ldots.. \qquad (6.3)$$

The calculations in this section thus give a somewhat stylized view of the effect on production costs of different settlement characteristics. They do however help to show up such trade-offs as there are between economies of scale in production and diseconomies in distribution, without too many extraneous factors complicating the comparisons. More complex situations, with multiple works, including borehole supplies, feature in **Section 2(i)(b)** below.

Now, the distribution costs shown in **Table 5.2** can be brought together with production costs obtained using (6.2) and (6.3) to give illustrative total costs of water supply for the scenarios considered previously (the values for *N*, *λ* and *w* are chosen to be reasonably representative of the values observed among BWC urban districts) leading to the results shown in **Table 6.4**. In this table, *TCP* is the total cost of water production, *TCD* is the total cost of water distribution and *TC(P+D)* is the total cost of water supply, comprising production and distribution. *UTCP*, *UTCD* and *UTC(P+D)* are the related unit costs, obtained by dividing by *QC* converted to an annual rate.

Illustrative values			Unit costs (£/Ml)			Total costs (£m pa)		
N	**λ**	**QC=w.N (Ml/d)**	**UTCP**	**UTCD**	**UTC(P+D)**	**TCP**	**TCD**	**TC(P+D)**
a. Varying *N*, *R* constant (‘densification’)								
5,000	0.19	2.1	582.66	584.13	1166.79	0.447	0.448	0.895
10,000	0.12	4.2	493.37	527.34	1020.71	0.756	0.808	1.564
15,000	0.095	6.3	447.62	477.32	924.94	1.029	1.098	2.127
20,000	0.075	8.4	417.76	445.79	863.55	1.281	1.367	2.648
25,000	0.06	10.5	395.97	421.62	817.59	1.518	1.615	3.133
40,000	0.03	16.8	353.73	371.81	725.54	2.169	2.280	4.449
50,000	0.015	21.0	335.29	349.78	685.07	2.570	2.681	5.251
b. Varying λ, *N* constant (‘dispersion’)								
18,000	0	7.56	428.45	349.08	777.53	1.182	0.963	2.145
18,000	0.02	7.56	428.45	365.80	794.26	1.182	1.010	2.192
18,000	0.04	7.56	428.45	386.15	814.60	1.182	1.065	2.247
18,000	0.06	7.56	428.45	412.65	841.10	1.182	1.139	2.321
18,000	0.08	7.56	428.45	451.35	879.80	1.182	1.245	2.427
18,000	0.10	7.56	428.45	530.75	959.21	1.182	1.465	2.647
c. Varying *N*, λ constant (‘suburbanisation’)								
5,000	0.06	2.1	582.66	410.91	993.58	0.447	0.315	0.762
10,000	0.06	4.2	493.37	407.28	900.65	0.756	0.625	1.381
15,000	0.06	6.3	447.62	410.24	857.85	1.029	0.944	1.973
20,000	0.06	8.4	417.78	413.98	831.74	1.281	1.269	2.55
25,000	0.06	10.5	395.97	421.62	817.59	1.518	1.615	3.133
40,000	0.06	16.8	353.73	453.24	806.97	2.169	2.779	4.948
50,000	0.06	21.0	335.29	494.64	829.93	2.570	3.791	6.361
d. Varying *N*, density=10 (‘constant density’)								
5,000	0.15	2.1	582.66	476.50	1059.15	0.447	0.365	0.812
10,000	0.1	4.2	493.37	457.85	951.22	0.756	0.701	1.457
15,000	0.08	6.3	447.62	446.40	894.01	1.029	1.027	2.056
20,000	0.07	8.4	417.76	437.08	854.84	1.281	1.34	2.621
25,000	0.065	10.5	395.97	428.35	824.32	1.518	1.642	3.160
40,000	0.05	16.8	353.73	416.19	769.92	2.169	2.552	4.721
50,000	0.045	21.0	335.29	409.53	744.81	2.570	3.139	5.709

Table 6.4: Illustrative calculations to show the effect of different values of λ and *N* on water supply costs for 35 BWC urban districts, assuming a single WTW

***Densification*: Section (a)** of **Table 6.4** shows the two-fold advantage of densification, leading to lower unit costs for both production and distribution. The unit cost of supply for a settlement of 50,000 properties is about 40% lower than for a settlement of 5,000 properties covering the same area. Returns to scale estimated from the last column are about 1.5.

***Dispersion*:** In **section (b)**, the unit cost of water production does not vary between cases so that this cost (about £428/Ml) is simply added to distribution costs. As in **Table 5.2**, greater dispersion (higher λ) leads to higher distribution costs (the increase in the unit cost of distribution is about 52% as λ increases from λ = 0 to λ = 0.1) and hence total costs which also rise, from about £778/Ml when λ = 0 to about £959/Ml when λ = 0.1.

Suburbanisation: **Section (c)** of the table is more interesting: here the higher volumes produced as N increases result in savings in unit production costs, which fall by about 40% from £583/Ml when N = 5,000 to £335/Ml when N = 50,000, thus offsetting the increase in distribution costs associated with serving less dense suburbs and rural areas. The effect is shown in **Figure 6.1**. Whereas distribution cost alone is minimized at about 10,000 properties, the minimum for production and distribution costs together in this case occurs at about 35,000 properties. The elasticity $\varepsilon_{N/\bar{\lambda}}$ for the combined cost is less than 1 below 35,000 properties (indicating scale economies) whereas for distribution alone it is greater than 1 if there are more than 10,000 properties (indicating scale diseconomies).

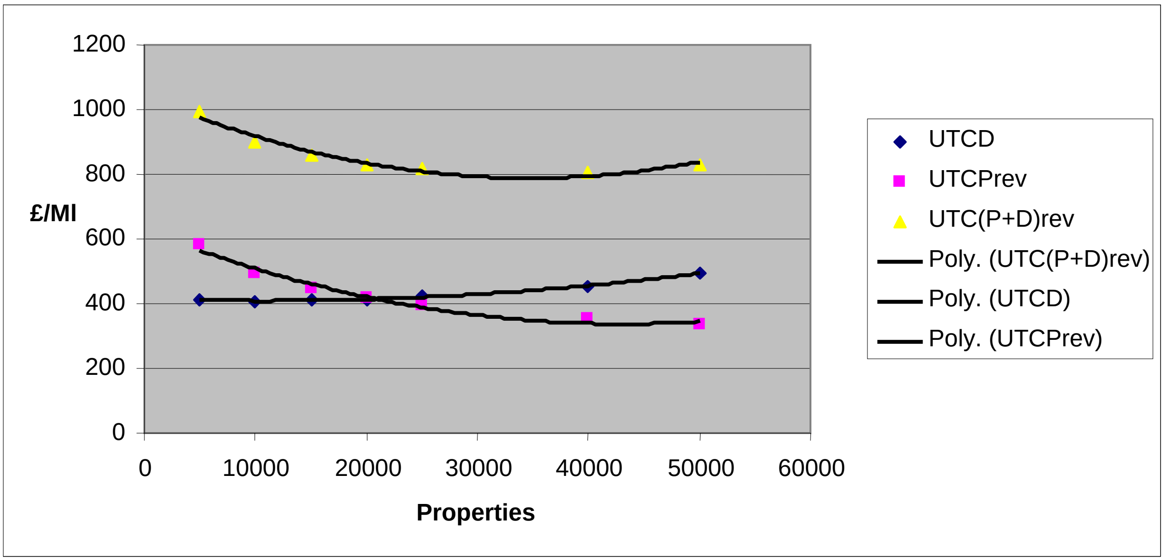

Figure 6.1: Unit production cost (UTCP), distribution cost (UTCD) and total cost (UTC(P+D)) from section (c) of Table 6.4

***Constant density*: Section (d)** of **Table 6.4** then shows how economies of scale in production reinforce the decline in distribution costs when property numbers increase but density remains constant, so that combined unit cost falls by about 30% from £1059/Ml when N = 5,000 to £745/Ml when N = 50,000. Returns to scale, estimated from the last column are about 1.25 (compared with about 1.10 for distribution alone).

These results indicate that the benefits of more compact settlement will be clearest when comparing towns of similar area or similar population but differing in density, as in sections (a) and (b) of **Table 6.4**. Adding population by expanding into peripheral areas (suburbanization) introduces a trade-off between volume economies (in both production and distribution) and diseconomies of average distance, which may on balance be favourable, despite lower average density, at least for moderate expansion, as shown in **Figure 6.1**. Constant density expansion, on the other hand, is unequivocally favourable so that in comparing towns of similar density but different populations, the larger towns should benefit from scale economies in both production and distribution, as in section (d) of **Table 6.4**.

b. Water supply areas

Data provided by BWC included information on the proportion of water supplied to each zone coming from each WTW or borehole source. By combining this with information on the output of each source, it was possible to assemble a new data set in which zones are grouped into 39 "water supply areas" which are more or less self-contained for water supply purposes. This enables the actual costs of water production in these areas to be estimated. Adding these costs to actual distribution costs for each area then provides an estimate of the actual total costs of water supply for these areas, which can be compared with the illustrative calculations of **Section (a)** above.

For water production, production costs are again estimated using (6.1) with the parameters from **Table 4.9** but now the calculated costs are for the numbers and sizes of production facilities, including boreholes, actually operating and their reported levels of treatment, rather than assuming that each area is served by a single level 4 WTW. The effect is to reduce the extent of economies of scale in water production, particularly in areas where boreholes predominate as returns to scale are lower for these sources. Distribution costs are obtained by summing the relevant operating costs and capital costs (allocated in proportion to length of mains) across the zones making up each distribution area.

The resulting combined unit costs[82] for 31 water supply areas are shown in **Figures 6.2** and **6.3** (data problems led to 8 areas being excluded from the results[83]). **Figure 6.2** shows the unit costs of production (*UTCP*), distribution (*UTCD*) and combined (*UTC(P+D)*) plotted against numbers of properties on the *x*-axis. First, it may be noticed that economies of scale in water production are very muted, due to multiple

[82] Calculated using water consumed (*QC*) as the divisor, as in **Table 6.4**.

[83] The production data is for a later year than the distribution data leading to discrepancies in quantities and changes in the boundaries of some areas. However, using unit costs, the effect on the results is small (even for the excluded cases).

works and the lack of variance in the size of boreholes – for example, the largest area is substantially reliant on numerous relatively small borehole supplies, whereas the second largest is mostly served by a single large WTW. Then although water distribution costs show some evidence of scale economies (the volume effect), there is a very wide range of costs among the smaller areas.

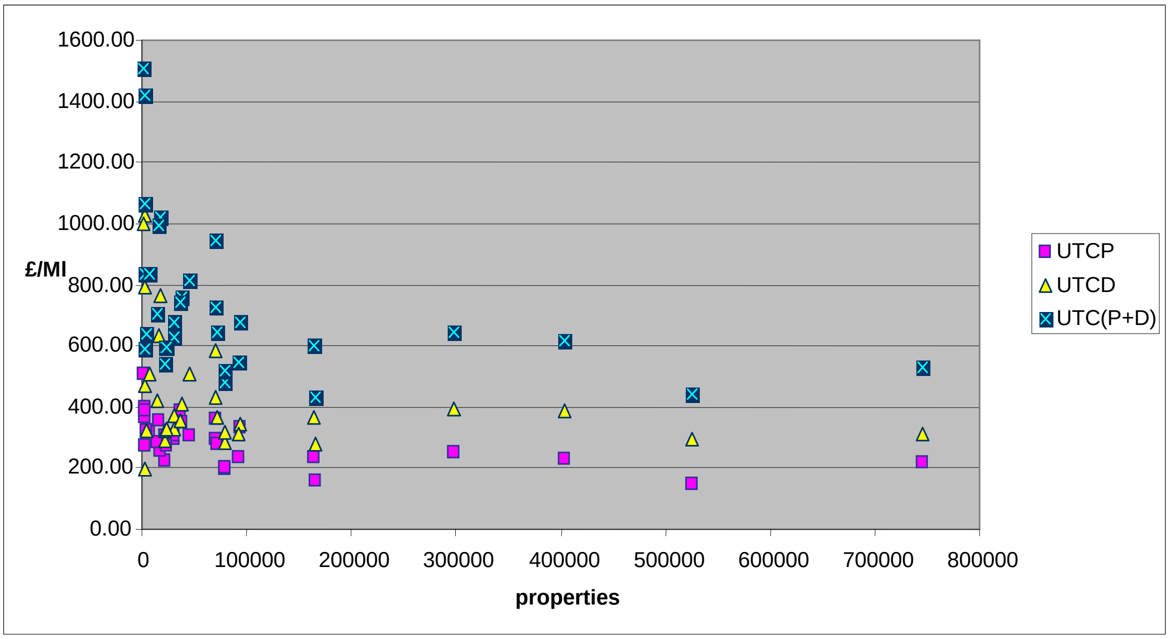

Figure 6.2: Unit costs of water production and distribution for 31 BWC water supply areas, plotted against number of properties

Much clearer is the picture that emerges in **Figure 6.3** when the same unit costs are plotted against density (measured as properties/km mains). Production costs are pretty much flat but with distribution a strong negative relationship between density and unit distribution costs is evident: low density leads to high distribution costs.

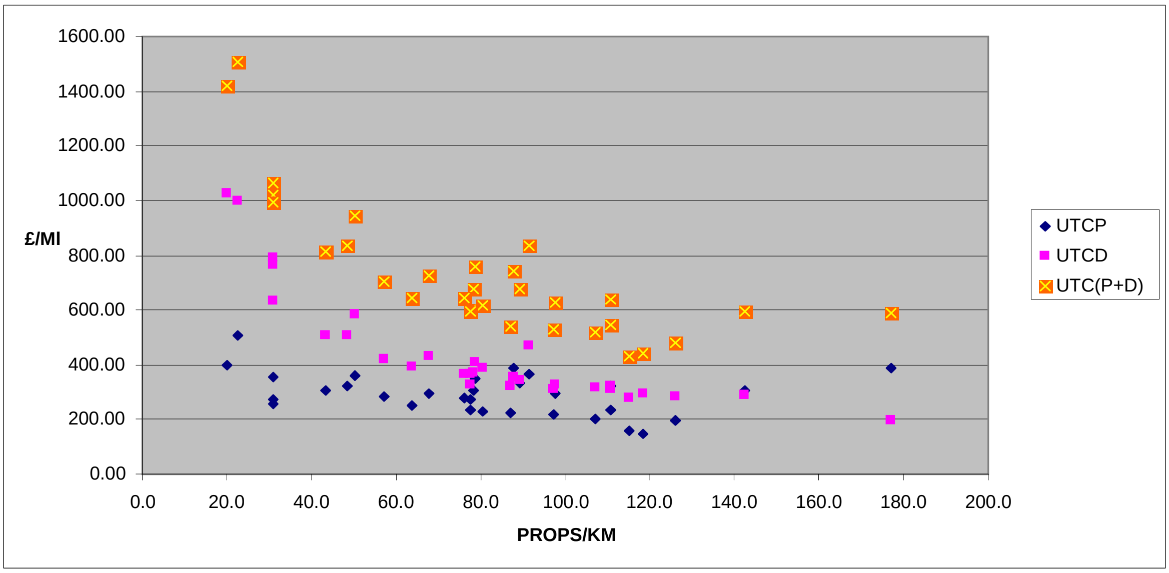

Figure 6.3: Unit costs of water production and distribution for 31 BWC distribution areas, plotted against density (properties/km mains)

Comparing these results with those in **Section 2(i)(a)**, the implication is that in practice, variations in distribution costs due to density effects (as in sections (a) and (b) of **Table 6.4**) are likely to be more important than quantity effects.

c. The 4 towns case

Inspection of the information used in **section 2(i)(b)** found 12 cases where a single source (WTWs in 3 cases and boreholes in 9 cases) provides the whole supply for that area. Much more common was the situation where each area receives supplies from several sources. It is likely that security of supply and water quality considerations rather than cost minimization explains this pattern of supply. In one interesting case, the water supply area consists of a single works serving 4 towns (with small amounts going to 2 other towns), and this works is the sole source for these towns. This case provides an opportunity to test the impact on costs if each town were to have its own treatment works compared with the arrangement actually in place. The set-up is sketched in **Figure 6.4** below, where **WTW** is the water treatment works, and **A**, **B**, **C** and **D** are towns. Each town has a suburban or rural periphery, as indicated by the dotted lines, which is also part of the supply area.

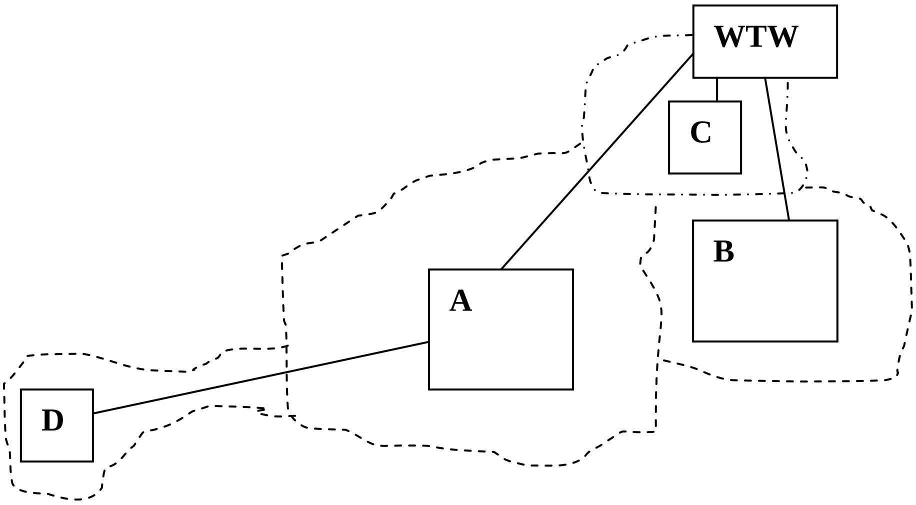

Figure 6.4: Sketch of the 4 towns set-up

Using the same relationships as in **section 2(i)(a)** above, this case can be used to estimate what the costs of supply would be under a variety of urban configurations. Starting with the existing set-up (1 WTW, 4 towns), this is compared below with:

- Each town having its own WTW of the appropriate size;
- The population of Town D migrating to Town C;

- The population of Town D migrating to Town A;
- All four towns combining to form a single town covering the same total area.

Basic data for the various areas is set out in **Table 6.5**:

Town	QP (Ml/d)	QC (Ml/d)	Props (No)	A_0 (Ha)	R ('00m)	λ	Density (props/km)
Town A	41.4	32.6	71998	5461	41.7	0.03	13.2
Town B	28.6	21.5	58446	3795	34.8	0.03	15.4
Town C	13.8	10.7	23214	3461	33.2	0.08	6.7
Town D	7.7	5.4	10756	1427	21.3	0.11	7.5
Town C+D	21.8	16.1	33970	3461	33.2	0.055	9.8
Town A+D	50.2	38.0	82754	5461	41.7	0.015	15.2
A+B+C+D	91.5	70.2	164414	14144	67.1	0.0225	11.6

Table 6.5: Basic data for the 4 Towns case

Now, to estimate water production costs, we use (6.2), taking the treatment to be level 4 (as is the case for the single treatment works here) so that *W4P* = 1, i.e:

$$UTCP = 2^{0.31}.474.QP^{-0.24} \quad \ldots\ldots\ldots.. \quad (6.4)$$

While to estimate water distribution costs, we use (5.22) and (5.23), i.e:

$$\ln VCD = -4.572 + 0.432\ln w + 0.617\ln\psi \quad \ldots\ldots\ldots\ldots\ldots \quad (6.5)$$

(S.E 0.234) (S.E. 0.027) (R^2 = 0.9455)

And

$$\ln CCD = -10.65 + 1.617\ln w + 0.622\ln\psi \quad \ldots\ldots.. \quad (6.6)$$

(S.E. 0.328) (S.E. 0.037) (R^2 = 0.8981)

This procedure leads to estimates of unit costs[84] and total costs of water production and distribution for different configurations of the 4 towns, the results of which are shown in **Table 6.6**:

[84] In this case, unit costs have been calculated using quantity produced (*QP*) as the divisor.

Configuration	**Unit costs (£/Ml)**			**Total costs (£m pa)**		
	Prodn (UTCP)	**Distn (UVCD+UCCD)**	**Total (UTCS)**	**Prodn (TCP)**	**Distn (TCD)**	**Total (TCS)**
(a) Existing set-up						
Town A	198.76	307.27	506.03	6.640	4.646	7.651
Town B		268.92	467.68		2.806	4.880
Town C		366.40	565.16		1.846	2.848
Town D		356.10	554.86		1.001	1.560
Total				**6.640**	**10.300**	**16.939**
(b) Separate supplies (‘autonomy’)						
Town A	240.41	307.27	547.68	3.635	4.646	8.281
Town B	262.79	268.92	531.71	2.742	2.806	5.549
Town C	312.95	366,40	679.35	1.577	1.846	3.423
Town D	359.99	356.10	710.10	1.012	1.001	2.013
Total				**8.967**	**10.300**	**19.266**
(c) Town D into C						
Town C/D	280.41	324.82	605.22	2.233	2.586	4.819
Saving vs (b)				**0.356**	**0.261**	**0.617**
(d) Town D into A						
Town A/D	229.58	292.96	522.54	4.206	5.368	9.574
Saving vs (b)				**0.441**	**0.363**	**0.720**
(e) A+B+C+D combined						
A+B+C+D	198.76	283.34	482.11	6.640	9.466	16.106
Saving vs (b)				**2.327**	**0.834**	**3.160**

.**Table 6.6: Estimated costs for supplying different town configurations**

Some caution is in order in interpreting these results, as the relationships being relied on are approximate and no account has been taken of any connecting reticulation between towns that might be required. What this table suggests – comparing (b) with (a) – is that each town having its own WTW would add about £2.3m (35%) to water supply costs, because of the higher costs of the smaller works operated by each town. Having the 4 towns share a single large works is clearly preferable to autonomy[85]. However, starting from (b), a position of autonomy, there are various other ways in which lower water supply costs might be achieved. For example, if the population of town D all migrated to town C, raising the population and density of the latter, this would lead to savings of about £0.356m in production costs and £0.261m in distribution costs, as shown in section (c) of **Table 6.6**. Similarly, migration of town D to town A would also produce savings as shown in section (d). More radically if the 4 towns combined to form a single town covering the same total area, this would yield savings of £0.834m in distribution costs as well as the £2.3m savings in production costs from sharing a single WTW, as shown in section (e) – a 16% reduction in total water supply costs compared with 4 towns, each self-sufficient. Indeed, if the populations of towns B, C and D were all to

[85] This would not be the case if the towns were supplied from boreholes with constant returns to scale.

migrate to town A, raising the density there to about 30 properties/Ha (about the maximum observed in the BWC zone data), additional savings of some £1.8m in distribution costs would be reaped.

What these examples show is that in the absence of some fundamental reorganization of distribution arrangements, economies of scale in production will dominate. It appears that cases of the type suggested by **Figure 3.1**, in which it might be advantageous to serve an area using two or more smaller works because the higher production costs are more than offset by lower distribution costs, are only likely to arise if linked to a consequential densification on the distribution side. For example, if a large rather dispersed settlement were replaced by two more compact settlements (occupying a smaller area in total), it might be the case that the higher cost of smaller separate WTWs could be offset by lower distribution costs within each settlement – although even in this case, the savings from sharing a single WTW would be worth having, provided the cost of connecting the two settlements is not too high[86].

3. Bringing water production and distribution together: (ii) AWWA

It may be recalled that water utilities in the US are generally relatively small, typically serving a single community, and are therefore rather suitable for the purposes of this research. While the results obtained in **Chapters IV** and **V** using data from the AWWA 1996 survey could be used to carry out illustrative calculations on the same lines as those in **Section 2** above, the story would be much the same as the key parameters are similar. Also there is a limitation in that the US data does not include information on capital costs. Instead, the US results are used here to study two rather different questions: (i) How much difference does analyzing production and distribution separately make to estimates of scale effects? (ii) How do our results compare with those obtained by Torres & Morrison Paul (2006), who used the same data source?

a. Effect of analyzing production and distribution separately

To examine this question, the first step was to identify those utilities which feature in both the **TreatQP** sample and in the **Retail only** sample. There proved to be 191 such cases. These are utilities which do not buy in water from other utilities (so that their production costs all relate to their own production) and nor do they sell water to other utilities (so that their distribution costs all relate to distribution to their own customers).

[86] With borehole supplies the case for sharing would be much weaker.

For these 191 utilities, re-estimating the relationship (4.11)[87] for production yielded:

$$\ln CST = -0.236 + 0.877^{***}\ln QS + 0.451^{***}\ln(1+SP) \quad \ldots\ldots\ldots\ldots \quad (6.7)$$

$$\mathit{(0.041)} \qquad \mathit{(0.153)} \qquad R^2 = 0.7176$$

Which is quite similar to the results obtained using all 388 cases in the **TreatQP** sample (see **Table 4.6** in **Chapter IV**). Then re-estimating (5.27) for distribution yielded:

$$\ln VCD = 0.050 + 0.474^{***}\ln QDI + 0.953^{***}\ln\varphi \quad \ldots\ldots\ldots\ldots \quad (6.8)$$

$$\mathit{(0.109)} \qquad \mathit{(0.259)} \qquad R^2 = 0.6878$$

Which is also little different from the previous result in (5.27). Note also that for the utilities in this sample *QS* = *QP* = *QDI*, as the quantity of water supplied is equal to the quantity produced which is equal to the quantity put into distribution.

So what should be the specification to estimate the relationship be if production and distribution are not treated separately? Final output has not changed but the relevant costs are now *CST* plus *VCD*, together making total variable costs of supply (*TVCS*). A control for the proportion of surface water (*SP*) is still appropriate. These considerations lead to the specification and results in (6.9) below:

$$\ln TVCS = 0.875 + 0.529^{***}\ln QDI + 0.832^{***}\ln\varphi + 0.334^{***}\ln(1+SP) \quad \ldots \quad (6.9)$$

$$\mathit{(0.083)} \qquad \mathit{(0.196)} \qquad \mathit{(0.119)} \qquad R^2 = 0.6878$$

Comparing (6.9) with (6.7) and (6.8), it may be seen that the coefficient on the quantity variable lies between the previous values, indicating stronger volume related scale economies than when production is taken on its own. This is perhaps surprising as it might be expected that distribution costs would counteract economies of scale in production but it follows from there being volume related economies in distribution, given average distance to properties, φ. At the same time, it is very likely that higher volumes will mean an increase in the average distance to properties, so adding to costs, as limits to densification are reached, so that considering the coefficient on ln*QDI* on its own is likely to be misleading in practice. Which effect is stronger will depend on the form of the expansion – whether its character is more like "*densification*", "*dispersion*", "*suburbanization*" or "*constant density*" in **Figure 5.2**. To assess this, we need to re-estimate (6.9) using *W* and ψ in place of *QDI* and φ. This produces:

[87] Dropping the terms in $(\ln QS)^2$ (not significant) and $\ln(1+PP)$ (not required).

$$\ln TVCS = 2.503 + 0.339^{***}\ln W + 0.649^{***}\ln\psi + 0.318^{**}\ln(1+SP) \quad \ldots \quad (6.10)$$

$$(0.090) \qquad (0.025) \qquad (0.117) \qquad R^2 = 0.8103$$

The implications are considered in **Section 3(ii)(b)** below.

b. Comparison of results with Torres & Morrison Paul (2006)

The sample of 255 water utilities used by Torres & Morrison Paul (2006) is taken from the same AWWA 1996 survey. There is some interest therefore in comparing our results with those obtained by these authors. It may be recalled (see **Appendix B, section 4(e)**) that Torres & Morrison Paul derive three primary elasticities of cost with respect to scale variables, and three combined elasticities as shown in **Table 6.7** below:

Measure	**Sample mean (8778 Mgal)**	**Small (675 Mgal)**	**Medium (1794 Mgal)**	**Medium-large (5962 Mgal)**	**Large (29590 Mgal)**
Volume (ε_{CY})	0.58 (*)	0.33 (*)	0.46 (*)	0.53 (*)	0.61 (*)
Service area (ε_{CS})	0.16 *	0.16 *	0.17 *	0.15 *	0.30 *
Customer Nos (ε_{CN})	0.49 *	0.49 *	0.53 *	0.51 *	0.54 *
Spatial density ($\varepsilon_{CYS} = \varepsilon_{CY} + \varepsilon_{CS}$)	0.74 (*)	0.49 (*)	0.63 (*)	0.68 (*)	0.91
Customer density ($\varepsilon_{CYN} = \varepsilon_{CY} + \varepsilon_{CN}$)	1.07	0.82 (*)	0.99	1.04	1.15
Size ($\varepsilon_{Size} = \varepsilon_{CY} + \varepsilon_{CN} + \varepsilon_{CS}$)	1.23 (*)	0.98	1.16	1.20 (*)	1.45 (*)

Table 6.7: Estimates of scale and density economies for 255 US water systems (adapted from Torres & Morrison Paul (2006, p.115)
(* = significantly different from 0; (*) = significantly different from 1; both at 1%)

First, we need to establish a correspondence between Torres & Morrison Paul's measures and the elasticities developed in **Chapter V**, taking into account that the latter cover distribution only whereas Torres & Morrison Paul's measures cover both production and distribution.

Torres & Morrison Paul	Definition	Chapter V equivalent
ε_{CY}	Elasticity of cost w.r.t. volume, *N* and *A* held constant.	ε_{w}
ε_{CS}	Elasticity of cost w.r.t. service area, *N* and volume held constant ("dispersion").	$\varepsilon_{A/\overline{N}}$
ε_{CN}	Elasticity of cost w.r.t. *N*, area and volume held constant.	No equivalent (implies falling *w*)
$\varepsilon_{CYS} = \varepsilon_{CY} + \varepsilon_{CS}$	Elasticity of cost w.r.t. service area and volume, *N* held constant.	No equivalent (implies rising *w*)
$\varepsilon_{CYN} = \varepsilon_{CY} + \varepsilon_{CN}$	Elasticity of cost w.r.t. *N*, area and *w* held constant ("densification").	$\varepsilon_{N/\overline{A}}$
$\varepsilon_{Size} = \varepsilon_{CY} + \varepsilon_{CN} + \varepsilon_{CS}$	Elasticity of cost w.r.t. service area and *N*, *w* held constant ("constant density").	$\varepsilon_{A/\overline{D}}$

Table 6.8: Equivalence between Torres & Morrison Paul's elasticities and those developed in Chapter V

Now, the results in (6.10), which are for production and distribution, can be used to estimate values for those elasticities for which there are equivalents, to compare with Torres & Morrison Paul's values. The values are estimated for a mid-sized utility serving a population of 50,000, with average consumption per property (W = 1,500 litres/property/day) and average use of surface water (SP = 0.34).

Torres & Morrison Paul (from Table 6.7, sample mean) (RTS = returns to scale)	Definition	Chapter V equivalent (mid-size utility)
ε_{CY} = 0.58 (RTS = 1.72)	Elasticity of cost w.r.t. volume, *N* and *A* held constant.	ε_{w} = 0.34 (RTS = 2.94)
ε_{CS} = 0.16	Elasticity of cost w.r.t. service area, *N* and volume held constant ("dispersion").	$\varepsilon_{A/\overline{N}}$ = 0.13
$\varepsilon_{CYN} = \varepsilon_{CY} + \varepsilon_{CN}$ = 1.07 (RTS = 0.93)	Elasticity of cost w.r.t. *N*, area and *w* held constant ("densification").	$\varepsilon_{N/\overline{A}}$ = 0.71 (RTS = 1.41)
$\varepsilon_{Size} = \varepsilon_{CY} + \varepsilon_{CN} + \varepsilon_{CS}$ = 1.23 (RTS = 0.81)	Elasticity of cost w.r.t. service area and *N*, *w* held constant ("constant density").	$\varepsilon_{A/\overline{D}}$ = 0.99 (RTS = 1.01)

Table 6.9: Comparison between Torres & Morrison Paul's elasticities and those calculated from (6.10)

It can be seen that there are significant differences between the estimated elasticities shown in **Table 6.9**, with only the dispersion elasticity being close in value. As regards the other measures, there is agreement that variations in volume, if numbers of properties and service area are fixed, are characterized by large returns to scale. This is entirely plausible as the marginal cost of delivering more water to existing customers is unlikely to be high. Our value is similar to that found by Torres & Morrison Paul for small companies but high compared with their values for larger companies. Looking next at the densification elasticity, Torres & Morrison Paul find modest diseconomies on this measure (although this is reversed for smaller companies). In contrast, our results indicate strong economies of scale. Intuitively, the latter seems more likely and a possible source of the difference is Torres & Morrison Paul's odd measure ε_{CN} which requires consumption per property to decline as numbers increase, itself a consequence of a specification which includes both volume of water (*QDI*) and numbers of properties (*N*) as explanatory variables, whereas using consumption per property (*W*) and numbers of properties would avoid this interaction. Finally, for the constant density elasticity, Torres & Morrison Paul find significant diseconomies whereas our result is constant returns to scale. One possible explanation for this difference is that Torres & Morrison Paul's method may imply that service areas have the same density across the whole area whereas our approach has density declining from centre to boundary, which will mitigate diseconomies – see **Figure 5.2 (d)**.

4. Bringing water production and distribution together: (iii) WOCs

In this section, the effect of bringing together the water production and distribution results for WOCs using Ofwat data is examined. For this purpose, attention is focused on 10 of the 12 WOCs – THD being omitted because of non-comparable treatment works information and TVN because of its wide-ranging distribution area. A comparison is then made with the results for WOCs obtained by Stone & Webster Consultants (2004).

a. Water production costs for WOCs

For production, the relationship used here is (4.20) with the estimated parameters from **Table 4.9**, i.e. for unit variable costs:

$$UVCP = (1+W4P)^{0.26}\left\{PHR^{0.56}(7.9)\left\lfloor\sum_i p_{Bi}\left(AQP_{Bi}\right)^{0.11}\right\rfloor + (343)\left\lfloor\sum_i p_{Ti}(AQP_{Ti})^{-0.39}\right\rfloor\right\} \quad \ldots. \quad (6.11)$$

And for unit total costs:

$$UTCP = (1 + W4P)^{0.31}\left\{PHR^{0.43}(39.7)\left[\sum_i p_{Bi}\left(AQP_{Bi}\right)^{-0.07}\right] + (474)\left[\sum_i p_{Ti}(AQP_{Ti})^{-0.24}\right]\right\} \quad \ldots. \quad (6.12)$$

Although these relationships were estimated across 21 water companies, including the WaSCs, they track reasonably well the actual costs for the 10 WOCs under consideration here, as **Table 6.10** shows.

Company	UVCP (£/Ml)			UTCP (£/Ml)		
	Actual	Calc	Diff(%)	Actual	Calc	Diff(%)
BWH	71	89	+25	170	220	+29
BRL	119	92	-23	250	231	-8
CAM	61	67	+10	100	154	+54
DVW	115	132	+15	223	278	+25
FLK	130	125	-4	284	283	0
MKT	117	91	-22	224	201	-10
PRT	60	72	+20	104	163	+57
MSE	93	97	+4	209	219	+5
SST	64	79	+23	130	196	+51
SES	120	136	+13	233	252	+8

Table 6.10: Actual and calculated unit production costs for 10 WOCs (Data: BHandTWnlsRev.xls)

However, what either set of figures shows is that despite apparently quite large economies of scale *at plant level*, as evidenced by the negative coefficients on the quantity variables in (6.11) and (6.12), economies of scale in water production *at company level* are negligible. Informally, this can be seen from a glance at **Figure 6.5** which plots UVCP and UTCP (actual values in £/Ml) against output in Ml/d for the 10 companies (the 7 smaller companies however might be seen as exhibiting economies of scale, although CAM is out of line, with remarkably low costs for its size):

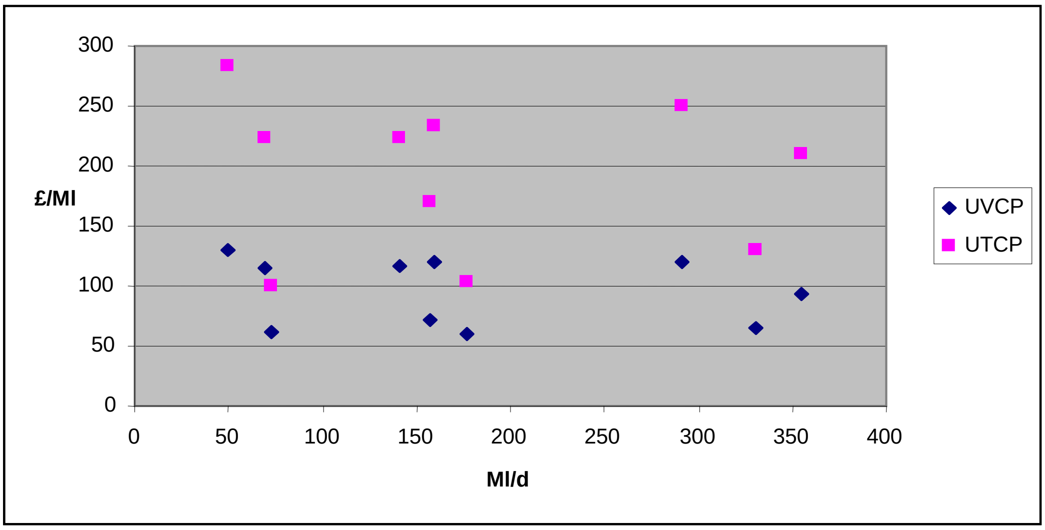

Figure 6.5: Unit production costs, 10 WOCs (actual values)

More formally, regressions of UVCP and UTCP against output (QDI) yield:

$$\ln UVCP = 5.099 - 0.120 \ln QDI \qquad \ldots\ldots.. \qquad (6.13)$$

$$(0.116) \qquad (R^2 = 0.1183)$$

$$\ln UTCP = 5.756 - 0.084 \ln QDI \qquad \ldots\ldots.. \qquad (6.14)$$

$$(0.115) \qquad (R^2 = 0.0637)$$

It can be seen that the coefficients on $\ln QDI$, although negative are not significantly different from zero.

There are two factors at work which help to explain this somewhat paradoxical finding:

(i) different mixes of borehole and surface water (the latter usually in large works but requiring more treatment);

(ii) multiple plant operations (a company operating 10 plants of certain size will show the same unit cost as one operating 5 plants of the same size).

On the first factor, while **Figure 6.6** below shows a high proportion of borehole supplies to be associated with higher cost for some companies, two companies with a high proportion are among the lowest cost producers while three with rather a low borehole proportion show rather high costs, so that across the 10 companies, the influence of this factor is more or less neutral.

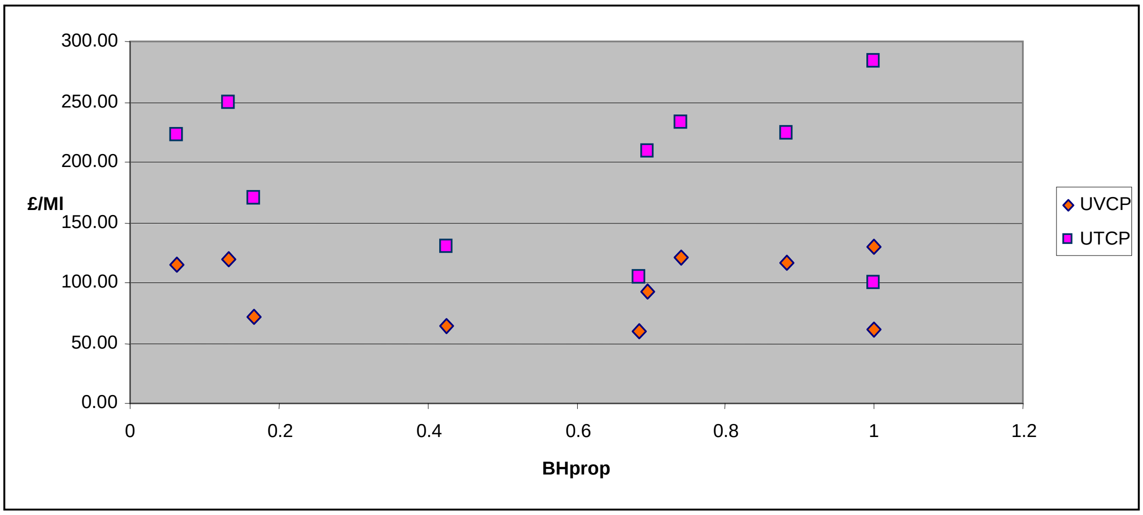

Figure 6.6: Influence of proportion of borehole supplies on unit production costs for 10 WOCs (actual values)

As regards multiple works, the breakdown in **Table 4.8A**, repeated below in **Table 6.11**, confirms that even the smaller companies operate several works while the larger ones operate dozens.

Company[88]	QDI (Ml/d)	TN (No)	Boreholes[a] (No)	WTWs (No)	AQP_B (Ml/day)	AQP_T (Ml/day)
BWH	157.6	7	4	2	6.62	65.72
BRL	291.3	23	16	7	2.40	36.12
CAM	73.2	14	14	0	5.23	0
DVW	69.5	9	4	5	1.11	12.55
FLK	49.5	18	18	0	2.75	0
MKT	140.7	29	27	2	4.60	8.09
PRT	177.2	20	19	1	6.39	55.82
MSE	355.2	65	57	5	4.34	21.53
SST	330.9	29	24	2	5.87	94.97
SES	159.9	11	7	1	16.93	41.41

Note: (a) Excluding size band 1 and zero output works.

Table 6.11: Output, numbers, type and average size of works, 10 WOCs

Now economies of scale at company level will only be apparent if larger companies operate larger works on average. But, as **Figure 6.7** shows, this is not generally the case. Although the 3 smallest companies operate works which are rather small (2 having no WTWs at all), the average size of boreholes is much the same for larger companies; then, while WTWs operated by larger companies are much larger than the average borehole, there is no clear tendency for larger companies to operate larger works on average. So again, economies of scale evident at works level get obscured in the aggregate.

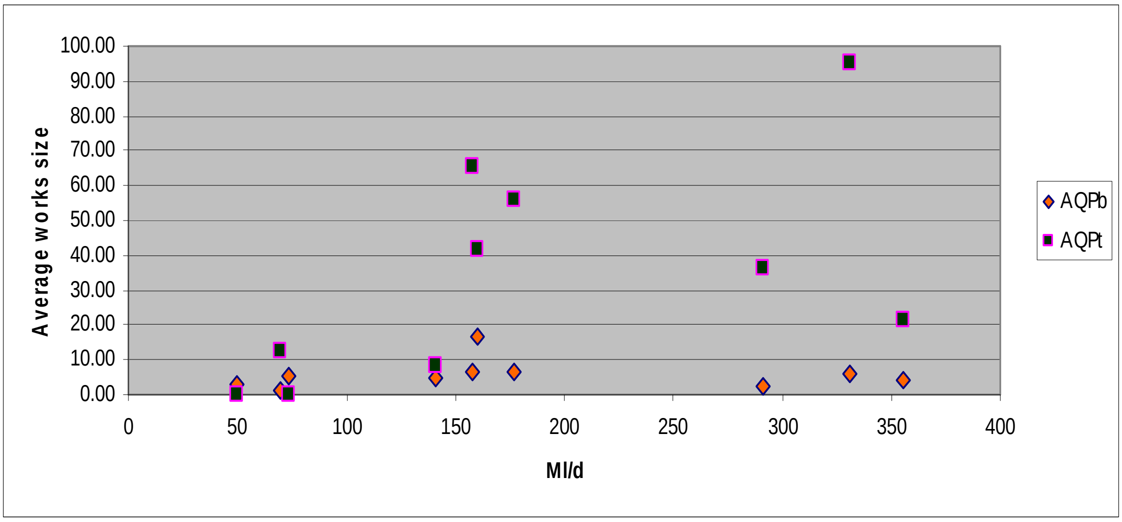

Figure 6.7: Company output and average size of works, 10 WOCs

The most probable reasons why companies do not exploit economies of scale in production to a greater extent are: (a) the location and capacity of the available water resources; and (b) the size and location of centres of demand, and the extra cost of

[88] For key to company acronyms, see **Tables 3.1A** and **3.1B** in **Chapter III**.

distribution if these are widely dispersed. The former is a matter of natural endowments and so of limited analytical interest but the latter can be explored further.

b. Water distribution costs for WOCs

The relationships for distribution costs for WOCs estimated in **Chapter V** (see (5.24) and (5.25)) were[89]:

$$\ln VCD = -3.213 - 0.211 \ln w + 0.659 \ln \psi \qquad \ldots\ldots\ldots\ldots \qquad (6.15)$$

(S.E. 0.342) *(S.E. 0.049) robust* *(R^2 = 0.9574)*

$$\ln CCD = -0.657 - 0.453 \ln w + 0.624 \ln \psi \qquad \ldots\ldots\ldots\ldots \qquad (6.16)$$

(S.E. 0.394) *(S.E. 0.057) robust* *(R^2 = 0.9378)*

Applied to the 10 WOCs here under consideration, these lead to the estimated unit distribution costs shown in **Table 6.12** and illustrated in **Figure 6.8**. As with the production cost estimates, the relationships track the actual values fairly well but offer little evidence of either economies or diseconomies of scale in distribution.

Company	UVCD (£/Ml)			UTCD (£/Ml)		
	Actual	Calc	Diff(%)	Actual	Calc	Diff(%)
BWH	77	66	-13	199	193	-3
BRL	103	99	-4	290	296	+2
CAM	142	103	-27	339	323	-5
DVW	86	103	+20	239	325	+36
FLK	75	84	+11	289	263	-9
MKT	82	105	+27	330	317	-4
PRT	85	89	+5	269	270	0
MSE	94	102	+8	347	302	-13
SST	102	92	-10	252	277	+10
SES	97	95	-2	304	290	-5

Table 6.12: Actual and estimated unit distribution costs for 10 WOCs

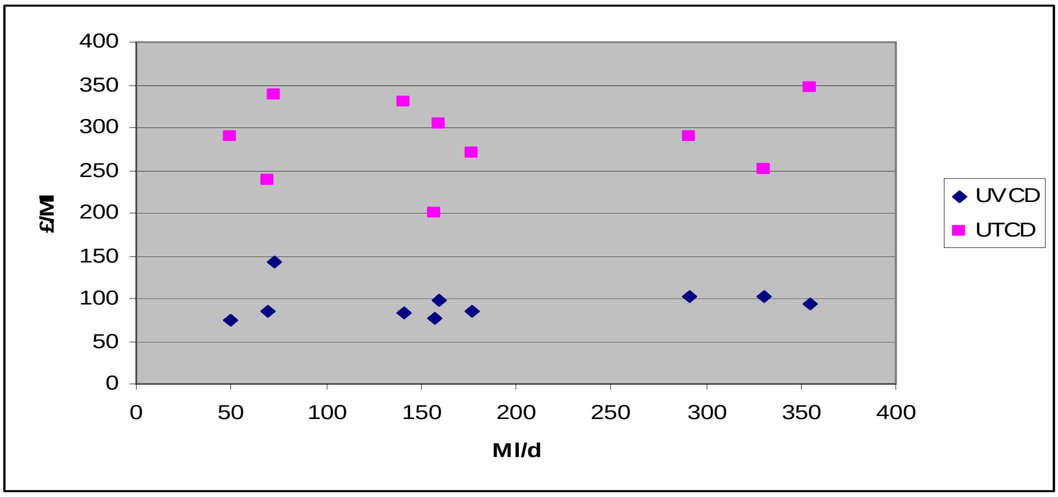

Figure 6.8: Unit distribution costs, 10 WOCs (actual values)

[89] Estimated for 11 WOCs, incl. THD.

From part (c) of **Table 5.3**, it would appear that this kind of outcome is likely if despite their differing sizes, the companies are of similar density. In fact, as **Table 6.3** shows, there is a more than threefold difference between the least dense WOC (CAM with density 1.02 properties/Ha) and the densest (SST with density 3.64 properties/Ha), with some tendency for the larger companies to be relatively dense. Higher density would tend to lower distribution costs. However, there could be a further effect at work: if the larger companies comprise several settlements, there could be an offset from the multi-settlement effect illustrated in **Table 5.5**. In any event, it appears that the net effect of these different influences is broadly neutral, leading to more or less constant returns to scale in distribution as well as water production.

c. Combined costs for WOCs

In the light of these findings, it is not surprising that bringing together the WOC results for water production and water distribution, as is done in **Figure 6.9**, also suggests more or less constant returns to scale in total supply costs.

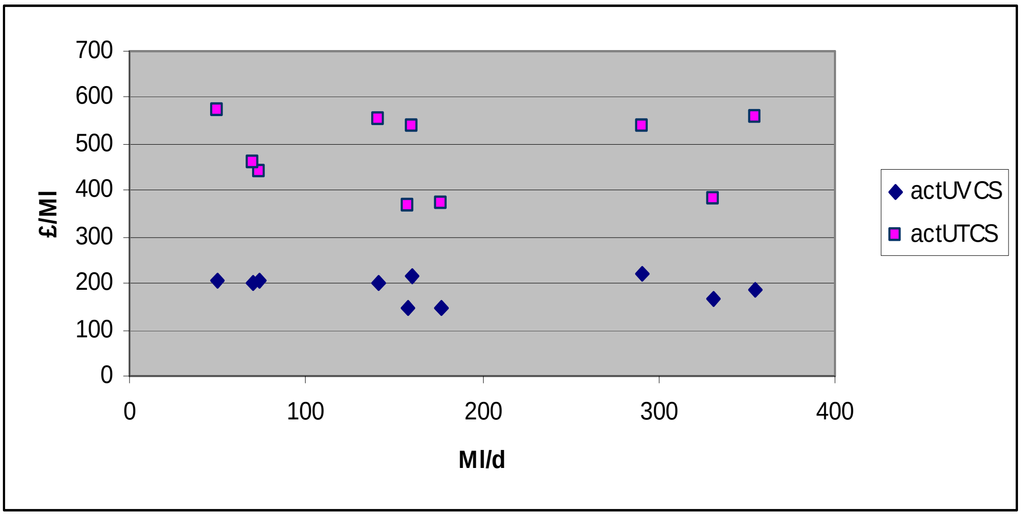

Figure 6.9: Combined unit cost of water supply (production + distribution), Actuals for 10 WOCs

The picture is little different if calculated values of the unit costs - derived using (6.11), (6.12), (6.15) and (6.16) - are plotted, as in **Figure 6.10** below.

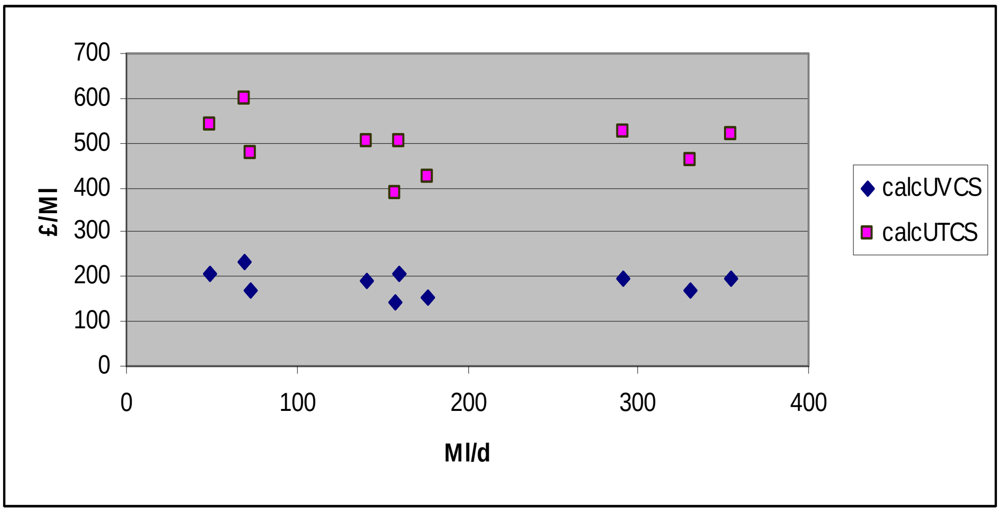

Figure 6.10: Combined unit cost of water supply (production + distribution), Calculated for 10 WOCs

Either way, taken as a whole the 10 companies show roughly constant returns to scale (as simple regressions – not reported here – confirm).

d. Comparison with Stone & Webster Consultants (2004)

In a 2004 report to Ofwat, which was then the most rigorous investigation of scale economies in the water industry of England & Wales to have appeared, Stone & Webster Consultants – hereinafter S&W – use a variable cost model specified in translog form, treating capital as a quasi-fixed input. Their task was complicated by the need to apply their analysis to WaSCs as well WoCs. Here we focus on their results for WOCs.

S&W note (p.10) that:

> "The concept of scale in the context of water service provision has a number of dimensions. Production may be measured in terms of the volumes of water and wastewater delivered and collected, in terms of the number of connections or population served or in terms of the supply area covered. Water companies with a similar scale, as measured by some physical measure such as the number of connected properties, may have very different cost characteristics because of differences in the density of those connections. This means that *economies of density* must be considered simultaneously with economies of scale …"

In their analysis of water supply, S&W take the principal outputs to be volumes of water delivered and number of properties for water supply. However, they felt that additional aspects needed to be considered. S&W addressed this by adopting a graduated approach, starting with a simple output model and then testing for improvements in model significance as additional variables were introduced. S&W conclude that the model specification is improved by adopting a multi-product approach. Their base **Model I**, they suggest, provides estimates of scale economies which are comparable to the estimates of *economies of production density* in Garcia & Thomas (2001). **Model II** in which connected properties feature as an additional output provides estimates of scale economies based on changes in both production and customers served (using numbers of connected properties is intended as a move towards recognition of the different characteristics of water distribution). In **Model III** they follow Garcia & Thomas in treating distribution losses as another output. Finally, in **Models IV** and **V**, a number of "hedonic" variables are introduced to control for "differences in service quality and characteristics of the operating environment for companies". These hedonic variables cover compliance with drinking water standards, water pressure, supply interruptions, % of properties metered, average pumping head and % of water from river sources. Generally, S&W conclude that it is appropriate and necessary to include hedonic variables in the estimated cost functions.

S&W's results for WOCs are summarized in **Table 6.13**. S&W's scale parameter is the inverse of the relevant elasticity (i.e. returns to scale) so that a value greater than one indicates economies of scale; a value less than one indicates diseconomies of scale.

	Short run Economies of scale		**Long run Economies of scale**	
	Parameter	**S.E.**	**Parameter**	**S.E.**
I. Base model (water delivered only)	1.42	0.08	1.25	0.09
II. Base model + connections	1.10	0.08	1.13	0.06
III. Base model + connections + distribution losses	1.09	0.08	1.11	0.07
IV. As III + water quality hedonics	1.04	0.08	1.05	0.07
V. As IV + metering hedonics	1.04	0.10	1.06	0.11

Table 6.13: S&W's estimates of short and long run economies of scale for water supply operations of WOCs
(Adapted from Stone & Webster Consultants (2004), Tables 9 and 11, pp. 40-41)

For WOCs, the preferred model **– Model V** in **Table 6.13 –** produces a result not significantly different from constant returns to scale. It is also noticeable that adding "connections" as an explanatory variable in moving from **Model I** to **Model II** leads to a sharp drop in the estimated scale parameter, which can perhaps be interpreted as some kind of *dis*economy associated with numbers of connections (and, perhaps, density).

While both Stone & Webster and this research find constant returns to scale for WOCs, the differences in approach have some interesting implications. Apart from the greater sophistication of S & W's methods, with a 11 year panel and a flexible form specification, their approach differs in (a) not separating production and distribution; (b) seeking to pick up distribution effects through *N*, the number of properties, rather than *φ* or some other spatial measure, such as service area or density; (c) not distinguishing between WTWs and boreholes, although the hedonic variable for surface water should control at least in part for this.

First, an observation on S&W's base model (line **I** in **Table 6.13**): A simple log regression of costs (production + distribution) against *QDI* using WOC data for 2002/03 yielded coefficients of 0.884 (*SE = 0.056*) (with variable costs, *VCS*) or 0.855 (*SE = 0.075*) (with full costs, *TCS*). In inverse form, these estimates imply returns to scale of 1.13 and 1.17, well below S&W's values of 1.42 and 1.25. (S&W's estimates are closer to our plant level estimates in **Table 4.10** but that is an inappropriate comparison as at company level, plant level scale economies are diluted by multi-plant operations and the WTW/borehole mix.)

Adding a term in numbers of properties (*N*) for comparison with S&W's **Model II** produces:

$$\ln VCS = -2.728 + 0.086 \ln QDI + 0.846 \ln N \qquad \ldots\ldots\ldots \qquad (6.17)$$

$$(0.259) \qquad (0.271) \qquad (R^2 = 0.9817)$$

And

$$\ln TCS = -1.831 - 0.169 \ln QDI + 1.086 \ln N \qquad \ldots\ldots\ldots \qquad (6.18)$$

$$(0.364) \qquad (0.382) \qquad (R^2 = 0.9626)$$

This looks like a big shift but in fact, recalling that ln*QDI* = ln*W* + ln*N* , what has happened is that the quantity variation has all been picked up by ln*N*, while the coefficient on ln*QDI* now reflects consumption per property, which does not vary greatly between companies and is not size related. The sum of the two coefficients is the same as before and the scale measures are now close to those estimated from S&W's **Model II**. In the light of this result, it is not clear why the introduction of numbers of properties should have induced such a large change in S&W's scale measure; it also casts doubt on whether any distribution effect has been picked up by this model.

In contrast, our *φ* variable, which measures average distance to properties, appears more illuminating. Regressions matching (6.17) and (6.18) give:

$$\ln VCS = -3.956 + 0.378 \ln QDI + 1.140 \ln \varphi \qquad \ldots\ldots\ldots \qquad (6.19)$$

$$\textit{(0.204)} \qquad \textit{(0.448)} \qquad (R^2 = 0.9778)$$

And $$\ln TCS = -3.381 + 0.211 \ln QDI + 1.449 \ln \varphi \qquad \ldots\ldots\ldots \qquad (6.20)$$

$$\textit{(0.285)} \qquad \textit{(0.626)} \qquad (R^2 = 0.9554)$$

While this again looks at first sight very different from what has gone before, the implications for scale effects now depend on the interaction between *QDI* and *φ*. For example, in the case of constant density expansion, scale parameters similar to S&W's are obtained, as we now show.

Referring back to **Table 5.3**, it may be seen from **part (d)** of the table that with constant density expansion, the percentage increase in *φ* is about 0.45 of the percentage increase in *N* [90]. The relevant scale measures in this case (if consumption per property does not change) are therefore:

$$RTS_S = \frac{1}{0.378 + (0.45x1.140)} = \frac{1}{0.891} = 1.12 \qquad \ldots\ldots\ldots \qquad (6.21)$$

And $$RTS_L = \frac{1}{0.211 + (0.45x1.449)} = \frac{1}{0.863} = 1.16 \qquad \ldots\ldots\ldots. \qquad (6.22)$$

Similarly, from **part (c)** of **Table 5.3**, if expansion is of the suburbanisation (constant *λ*) type, the percentage increase in *φ* is about 0.63 of the percentage increase in *N*. The relevant scale measures in this case (if consumption per property does not change) are therefore:

$$RTS_S = \frac{1}{0.378 + (0.63x1.140)} = \frac{1}{1.096} = 0.91 \qquad \ldots\ldots\ldots \qquad (6.23)$$

And $$RTS_L = \frac{1}{0.211 + (0.63x1.449)} = \frac{1}{1.124} = 0.89 \qquad \ldots\ldots\ldots. \qquad (6.24)$$

That is to say, with this type of expansion, there will be diseconomies of scale.

Finally, if expansion was in the form of an increase in density within the existing company boundary, as in **part (a)** of **Table 5.3**, application of (6.17) and (6.18) would misleadingly imply the same scale effect as constant density expansion. Application of (6.19) and (6.20), on the other hand, would pick up the point that in this case higher *N* would be associated with a rather small increase in *φ*. In **Table 5.3**, the percentage

[90] Average of piece-wise estimates.

increase in φ is about 0.15 of the percentage increase in *N*, so that relevant scale measures become:

$$RTS_S = \frac{1}{0.378 + (0.15x1.140)} = \frac{1}{0.549} = 1.82 \qquad \ldots\ldots\ldots \qquad (6.25)$$

And $$RTS_L = \frac{1}{0.211 + (0.15x1.449)} = \frac{1}{0.428} = 2.34 \qquad \ldots\ldots\ldots. \qquad (6.26)$$

That is, densification within the existing boundary should lead to rather large economies of scale as the volume effect benefits both production and distribution costs with very little increase in average distance to properties. However, this case should not be extrapolated too far as increasing densification would at some point run up against the assumption that the central density does not exceed 30 properties/Ha – necessitating a reassessment of the relationship between *N*, *R* and *λ* in **Appendix I**.

In short, while S&W are right to say that “The concept of scale in the context of water service provision has a number of dimensions”, it needs a specification which includes a spatial variable to do justice to this point – as Saal & Parker (2005) recognize.

5. Conclusions

The effect of different settlement patterns on the combined costs of water production and water distribution can be seen in the illustrative calculations in **Table 6.4**. Here four types of comparison are set up, characterized as (a) *densification*; (b) *dispersion*; (c) *suburbanization*; and (d) *constant density*. In each case, a single large WTW of the appropriate size is assumed. Now, as numbers of properties are increased in each scenario (leading to higher volumes, given constant usage per property), the key difference is how density is affected.

- With (a) *densification*, because the urban boundary does not change as property numbers increase, density increases in parallel, so that volume economies predominate in distribution as well as production. For example, unit water supply costs for a town doubled in size to 50,000 properties occupying 2,250 Ha (density 22.2 properties/Ha) will, according to these calculations, be 16.2% lower than for a town of 25,000 properties occupying the same area (density 11.1 properties/Ha), about half of the reduction coming from lower unit water production costs and half from lower unit distribution costs.

- With (b) *dispersion*, the number of properties does not increase, so that there is no volume effect, but the more dispersed pattern of settlement means lower density and an increasing average distance to properties, and hence higher distribution costs. For example, unit water supply costs for a town of 18,000 properties spread out over 2,090 Ha (density 8.6 properties/Ha) will be 10.8% higher than for a town of 18,000 properties occupying only 735 Ha (density 24.5 properties/Ha), all due to a 23.4% increase in unit distribution costs.
- With (c) *suburbanization*, the number of properties increases but because the increase is into less dense peripheral areas, average density falls and average distance to properties increases, albeit to a lesser extent than with (b). In this case, volume economies (in both production and distribution) are more or less balanced by average distance diseconomies. For example, unit supply costs for a town which has grown to 50,000 properties occupying over 20,000 Ha (density 2.4 properties/Ha) will be much the same as for the same town when it was only 15,000 properties occupying 985 Ha (density 15.2 properties/Ha) with the 25% reduction in unit production cost due to higher volume largely offset by a similar increase in unit distribution cost (the distance effect outweighing the volume effect in distribution here).
- With (d) *constant density*, the number of properties increases in line with the increase in area so that density is unchanged although the average distance to properties does increase. In this case, volume economies (in both production and distribution) outweigh the average distance effect. For example, unit supply costs for a town of 50,000 properties occupying 5,000 Ha (density 10 properties/Ha) will be 16.7% lower than for a town of 15,000 properties occupying 1,500 Ha (also 10 properties/Ha), about three-quarters of the reduction coming from lower unit production costs and one quarter from lower unit distribution costs.

These examples are enough to illustrate the range of effects that might be observed, but, it might be asked, which are particularly relevant when thinking about urban infrastructure? In studies of agglomeration, it is common to use population as the measure of size. One lesson from these examples is that it may not be sufficient to look at numbers alone. Whereas increase in size through *densification* would, it seems, bring economies of scale (in water supply at least), with a positive influence on agglomeration, as would (to a lesser extent) *constant density* expansion, increase in size through *suburbanization* would be roughly neutral in cost terms. To get the full picture, it would appear necessary to take density explicitly into account, not just size. Moreover, it would be misleading to regard urban areas of similar size, as measured by population, as equivalent from an agglomeration perspective, if they have very different densities. As the '*dispersion'* example suggests, lower density towns or cities are likely to have higher distribution (and access) costs. Put differently, agglomeration by densification would have real cost advantages (at least up to the point where congestion costs become appreciable) whereas suburbanization would not.

Yet another way to look at the matter is to compare water supply costs as between a small town and a large one. Even if they have the same density, the 'constant density' calculations point to lower costs in the latter. If this effect generalizes to other types of infrastructure, it suggests an important reason why large settlements might over time prosper more than small ones; and if the larger one is also denser, the advantage becomes greater still. Of course, infrastructure costs are not the only consideration but if, for example, people have a preference for suburban living, these calculations indicate that there is likely to be a cost penalty (whether or not this is visited on suburbanites through tariffs and connection charges). These wider issues are taken up for further consideration in the next and final **Chapter VII**.

VII. GENERALISATION: APPLICATION TO OTHER URBAN INFRASTRUCTURE AND IMPLICATIONS FOR AGGLOMERATION

1. Focus of this research

It was suggested in **Chapter II** that much of the man-made urban infrastructure can be seen as belonging to one of two broad types:

- **Area-type**: Provides services within a defined area (e.g. utilities, transport systems). In such cases, getting the service to users involves distribution costs;
- **Point-type**: Provides services at a specific point (e.g. hospitals, schools, offices, shops, museums, theatres, etc). In such cases, the equivalent consideration is the cost to users of accessing the facility.

For Area-type infrastructure, it was considered likely that the cost of supply would be driven by:

1. Possible scale economies in production (e.g. water treatment works);
2. Possible diseconomies in distribution costs, which would be likely to increase more than in proportion to the size of the area served;
3. Possible savings in distribution costs related to higher population densities.

For Point-type infrastructure, the equivalent influences were seen as:

1. Any scale economies in the basic facility (e.g. hospital, school, museum);
2. Possible diseconomies in access (e.g. transport) costs, which would be likely to increase more than in proportion to the size of the catchment area (cf. commuting costs – Arnott (1979));
3. Possible savings in access costs related to higher population densities; and, in addition
4. Possible congestion costs, which would be likely to increase with size of catchment area and population density.

Either way, there would be an element of trade-off between economies of scale in production and diseconomies in distribution (or access); and whereas economies of scale in production and density economies would be conducive to agglomeration, diseconomies in distribution would act in the opposite direction. We are now in a position to consider how far this research has been able to assess these effects in the case of water supply (**Section 2**); how far the results might be generalisable to other Area-type infrastructure (**Section 3**); to Point-type infrastructure (**Section 4**) and to Transport (**Section 5**); and what the implications might be for urban agglomeration (**Section 6**). All this is brought together in a summary of conclusions (**Section 7**).

2. Summary of water supply findings

Bearing in mind that the aim of this research is to throw light on scale effects at settlement level, the findings on urban water supply can be summarised as:

- *Economies of scale in water production*: There are economies of scale at plant level for water treatment works (WTWs) – with returns to scale of about 1.25 (or more) – but the evidence for economies of scale for boreholes is less clear. However, these effects may not be observed in practice because large settlements (and large companies) will often exploit multiple water sources, operating numerous works (both WTWs and boreholes), and then aggregate production costs will tend to show more or less constant returns to scale. Only in rather rare cases (such as Birmingham, with a large, dense population, and a large WTW supplied from a large reservoir) will economies of scale in production have an appreciable effect.
- *Diseconomies in distribution*: Modelling urban areas as monocentric settlements and measuring distribution output as the product of volume and distance components, we find that there are volume related scale economies in water distribution but diseconomies related to average distance to properties. Diseconomies are therefore only evident where the distance effect dominates. We have found diseconomies where properties are more spread out ("*dispersion*") and where lower density development around the urban core takes place ("*suburbanisation*"), but not when development takes place within the existing urban boundary ("*densification*") or where a settlement expands without density changing ("*constant density*"). This is because the volume effect dominates in the latter cases. The reasons why these results are not consistent with the Arnott model of urban commuting costs are discussed below.
- *Density savings*: An implication of the distribution findings is that density effects are rather important with higher densities leading to lower unit distribution costs, reinforcing economies of scale in producing for a larger population – so that, for example, the unit cost of water supply for a town doubled in size to 50,000 properties occupying 2,250 Ha (density 22.2

properties/Ha) is about 16.2% lower than for a town of 25,000 properties occupying the same area (density 11.1 properties/Ha), about half of the reduction coming from lower unit water production costs (assuming a single large WTW) and half from lower unit distribution costs.

- *Interaction between economies of scale, distribution costs and density effects*: These effects have been explored in **Chapter VI –** see particularly **Table 6.4**.While there are cases where there is a trade-off between economies of scale in production and diseconomies in distribution, in other cases volume economies in both production and distribution dominate. More generally, changes in the size of a settlement (as measured by numbers of properties, or population) are less important than changes in density in determining whether or not there are economies of scale.

It may be wondered why water distribution costs are not unequivocally rising in settlement size as measured by population, as commuting costs are in Arnott (1979). It is because in Arnott's model, density does not vary within the settlement, each commuter follows a direct radial route from residence to CBD and commuting cost is proportional to distance travelled. In the real world, however, as measurement of density gradients shows and the monocentric urban model requires, density generally declines away from the centre, making the average distance to properties shorter than in the constant density case. More importantly, water is not channelled in individual pipes to each house but is usually carried collectively in larger mains for most of the distance, with consequent cost savings. This, we may surmise, is the main reason for the volume scale economies found in water distribution. In fact, the larger the settlement, the greater the scope to adopt such collective means of delivery (London's massive Ring Main providing a particularly striking example). Indeed, much the same is true of commuting itself. Larger (and denser) settlements should be able to provide collective means of transport for commuters (buses, metros) at a cost lower than if they travelled individually[91]. In these circumstances, reliance on the Arnott model will be misleading – suggesting that urban theorists should be cautious about assuming too readily that commuting costs are increasing in city size[92].

[91] However, provision of such collective transport systems may well involve large capital costs, itself requiring a large, dense market to be viable.

[92] For example, Fujita (1989) says: "Let us suppose as before that a city takes a monocentric form. Then due to the increase in commuting distance, the total transport cost of a city increases more than proportionally to its population."

We need next to consider how far the results obtained for water supply are applicable to infrastructure more generally. In the sections that follow, the discussion does not rest on new research but rather tries to make suggestive connections with the existing body of knowledge.

3. Application to other Area-type infrastructure

Referring back to the types of urban infrastructure identified in **Chapter II (Table 2.1)**, those classed as “Utilities” can be considered as Area-type infrastructure. Apart from water supply, the utilities identified there include sewerage, storm drainage, electricity and gas supply, and telecommunications[93]. The characteristics of sewerage and storm drainage (which often come together in combined drainage systems) can be expected to be similar to those found for water supply. As regards electricity and gas supply, when each town had its own gas works, and electricity generation was more local, the similarities with water supply were also substantial. However, since the 1960s, town gas works in Britain have been replaced by bulk supplies of natural gas from the North Sea and elsewhere, changing fundamentally the economics of gas production and distribution; similarly, electricity production has increasingly been concentrated in very large power stations, changing the character of the trade-off between economies of scale in production and diseconomies in distribution. In consequence, long distance bulk transmission plays an important role in electricity and gas distribution. This is less a feature of water distribution where treatment works tend to be located near the settlements they serve – although bulk supplies *to* treatment works are of some importance. In the case of telecommunications, the analogy with water supply is further strained: what exactly constitutes the production unit may be difficult to pin down and distribution costs are relatively much less important.

These similarities and differences are considered in more detail in sub-sections (a), (b) and (c) below. If there is a unifying theme in this section, it is how far taking into account distribution costs might favour a more decentralized pattern of infrastructure provision (e.g. small scale local power generation rather than centralized provision from large power stations through a national grid). To the extent that this is the case, agglomeration forces are weakened – or perhaps more accurately, the cost disadvantage of smaller settlements will be reduced.

[93] Other services with similar characteristics include fire and police services, postal services, and certain health services, where the operation requires a base station from which services are delivered to people or properties in a defined service area.

a. Sewerage and sewage treatment

The sewerage and sewage treatment activities of the WaSCs would be amenable to an analysis very similar to that which has been carried out for water acquisition and treatment, taking advantage of the information on this part of their functions provided in their June Returns to Ofwat. It can be anticipated that similar results would be obtained, and similar problems encountered due to the large number of settlements served by these companies. More detailed local information would then be required to get a clearer picture.

The likely similarities, and differences, between the two cases can be noted:

- The flow of sewage is in the reverse direction, from properties to the treatment works, and mainly by gravity, so that pumping costs are relatively unimportant;
- The pipes used for sewerage are generally of a larger gauge than those used for water supply, and the replacement cost of sewerage assets is about twice that for water supply (Water UK (2004, p.2))
- Casual observation suggests that the extent of economies of scale in sewage treatment works is likely to be less than in the case of water supply (it appears that enlargement consists mainly of increasing the number of filter beds);
- Volumes are larger and more unpredictable, as sewers often take rain water as well as other effluents (raising additional questions about how output for this part of the system should be measured).

Empirical investigation is needed to take this further but one can hazard that if it proves to be the case that economies of scale in sewage treatment are small and pipe costs relatively high, this would favour smaller, local works over large centralized ones. But there are also likely to be factors other than costs to consider, such as the capacity of the local environment to accept discharges. It has in the past been Environment Agency policy to encourage consolidation because the performance of large works is easier to monitor; however, recently there has been some softening of this position:

> "There are ways to treat sewage other than pumping it to a few large works. Small, local sewage works in a new development would help to maintain more natural water flows throughout a river catchment. But this has to be balanced against the efficiency that large STWs can provide. Different options will be appropriate for different places." Environment Agency (2007, p. 10)

In this case therefore the position of smaller settlements may be easing.

b. Electricity supply

Turning to electricity supply, there are some important differences in the characteristics of the industry to note. Distribution involves two stages: high tension bulk transmission and medium or low tension distribution. The former generally requires pylons and heavy duty wires; distribution to customers (after voltage reduction through transformers) is then usually through underground cabling. The post-WWII trend in production has been towards larger and larger power stations whose location is determined mostly by considerations such as proximity to fuel supplies (e.g. coal mines) or availability of cooling water (and public acceptance in the case of nuclear power) rather than distribution costs – the view being that economies of scale in production dominate other considerations. Although the capital costs of the transmission and distribution networks are substantial, operating costs in electricity distribution are low and the cost of power lost in transit (line losses) is about 6%[94] (compared with leakage rates of 20% or more in water distribution). The trade-off between economies of scale in production and the cost of distribution has not therefore played much part in decisions about the size and location of power stations. Nevertheless, the cost of extending the distribution network to small or remote settlements is relatively high.

Recently, the development of smaller scale types of electricity production, often using unconventional technologies, such as solar panels, wind turbines and CHP (combined heat and power) units have prompted some re-thinking on this score. An article in the New Scientist remarks[95]:

> "Almost all of us can trim our utility bills by generating our own electricity. Photovoltaic tiles or a small wind turbine on the roofs of houses or apartment blocks are no longer a rarity. If these and similar small-scale generators were installed in large numbers they could have a significant impact on energy policy, helping to slash carbon emissions *and taking the strain off overloaded distribution grids*. A growing enthusiasm for renewable energy has also stimulated development of new small scale energy generators that are reliable, simpler to install and, most importantly, *capable of exporting the power they create onto the grid*." [Emphasis added.]

Later in the same article, the possibility of a trade-off between production and distribution gets explicit mention:

> "On the plus side, microgeneration could avoid expensive upgrades to the distribution grid which would be needed if predicted growth in demand is met solely by centralized generators"

[94] "Losses are not insignificant in electricity infrastructure (roughly 6% of capacity cost) … the operational costs are only 2% of capacity costs." Furong Li (personal communication).

[95] Hamer M "Every home should have one" *New Scientist* 21 January 2006 (pp. 36-39).

Making due allowance for journalistic hype – the economics of small scale generators remain weak, and it will be a long time before they can make a quantitatively significant contribution to meeting demand – it does raise the interesting question whether savings in distribution costs could help to justify small scale local production.

As it happens, the question has attracted the attention of the industry regulator, Ofgem (Office of Gas and Electricity Markets). Its proposals following its 2004 Electricity Distribution Price Control Review included a new incentive framework for distributed generation to facilitate the connection to distribution networks of renewable generation – Ofgem (2004).

A study commissioned by Ofgem, Li *et al* (2005), examines some aspects of the question in greater depth. The method is to assess the change in future investment requirements on the high voltage network consequent on a change in the charging regime facing Distribution Network Organisations (DNOs). Different charging regimes would result in different patterns of growth in demand and distributed generation. Three types of charging regime are considered (the description below relies on Li *et al* (2005, pp. 11-14):

a. The existing DRM (Distribution Reinforcement Model) system, which "is essentially an allocation model that attributes the costs of the existing network to users depending upon the use they make of each voltage level of the distribution system, as inferred from their maximum demand and customer class characteristics";

b. Two economic pricing models, both of which start from an assessment of the marginal cost of adding an increment of demand or generation at each node of the system. Then:

i. In the ICRP (Investment Cost Related Pricing) model, it is assumed that incremental demand (or generation) is met by uniformly expanding the network. It is the same general approach as that employed by National Grid for transmission charging, modified to apply to the distribution network. A standard cost, known as the "expansion constant" is calculated for each circuit on the reference network, and applied to the "distance" power must flow to meet the increment in demand;

ii. In the LRIC (Long Run Incremental Cost) model, the marginal cost is assessed from the change in the present value of the anticipated costs of

> reinforcing the network as a consequence of adding the increment. The LRIC approach thus endeavours to recognize the existence of unused capacity on the network by assessing the additional cost that arises from the need to advance investment as a result of adding load or generation at any node on the system, or alternatively the reduction in cost that will result from delaying investment.

The pricing messages that emerge from these models are not simple, particularly as the authors add a scaling factor to ensure that the tariff delivers the revenue permitted under regulatory price control. It is beyond the scope of this short account to go into these implications in depth but the key point is that different pricing models convey different messages about the location of additional capacity. The following quotation gives a flavour:

> "For generation, where no scaling has been applied, the ICRP methodology mirrors the distance effect for the rural nodes by producing substantially negative charges (credits) for generation. This should attract generation to the more rural area. Under the LRIC approach, generation would find it most attractive to connect to nodes 3 and 5, which are in the urban area." (Li *et al* (2005), p.18)

What is interesting here is that there is very little discussion either in Li *et al* (2005), or in an Ofgem discussion document issued at about the same time[96], of the economics of different production systems or of scale effects in production. The aim is rather to present potential generators with a tariff for transmission and distribution that correctly reflects the costs imposed by connecting to the system. It is then for the generators to assess the costs and benefits of actually doing so – the "Where?" "When?" and "How much?" questions. Some hypothetical examples help to illustrate what may be involved:

> a. An electricity consumer wants to replace part or all of his standard supply with some form of self-generation (e.g. wind turbine or solar panels). This can be addressed by evaluating the investment using the relevant local electricity tariff to price the self-generated electricity. (Note that this tariff will include electricity production costs as well as transmission/distribution costs). The result is often expressed in statements of the form: "This investment will pay for itself in x years".
>
> b. An electricity consumer wants not only to replace part or all of his standard supply but also to be able to sell to the grid any surplus production. Additional

[96] Ofgem (2005) *Enduring transmission charging arrangements for distributed generation: A discussion document*, September 2005 (available on the Ofgem website).

elements in the cost/benefit calculation are now charges for connection to the grid and the price offered for any surplus produced. This is where the methodologies discussed above start to become relevant.

c. An entrepreneur, not necessarily an existing electricity customer or producer, wants to set up a production facility (e.g. a wind farm). Now the cost/benefit calculation will be driven entirely by the charges for connection to the grid and the price offered for the electricity generated.

In this set-up, the pattern of production that emerges will be the outcome of a series of incremental decisions by independent entities (customers and generators), guided by the charges and tariffs of the electricity distributors (in turn influenced by the price they pay to the existing large scale generators). It is evident that although the economics of electricity production are not emphasized in this system, they must powerfully influence the outcome: If the cost of large scale centralized generation is low, it will probably be difficult for small scale units (with relatively high unit costs) to achieve reasonable pay-back periods, even if transmission or distribution costs are saved. On the other hand, situations could possibly arise when it will be profitable for one or more small scale players to install their own production capacity even though a single large facility would be more economic. Overall, the effect, as regards agglomeration, of decentralized decision-making in this case is likely to tend towards making small settlements more viable.

c. Telecommunications

Our discussion of the economics of telecommunications is equally brief but it does provide an interesting contrast in one respect. Whereas in water supply (and sewerage), distribution costs are very significant, and in electricity supply they are still appreciable, in telecommunications they are negligible. This observation has excited futuristic speculation about “the death of distance” (Cairncross (1997)). However, the capital costs of establishing communications networks, including wireless networks, have not faded away. Companies have had to pay enormous licence fees for wave band access, transmission masts involve significant investment and many communications applications still require extensive cabling. Moreover there are network benefits in having large numbers of customers (the more customers who are connected, the more valuable the service to other customers).

The effect of these characteristics is an industry in which entry costs are very high but, once established, the marginal costs of supply are very low. It is the world of Dupuit (1844) with the associated conflict between economic pricing and cost recovery. We have no new insights to offer on these matters. However, we may note that the possibility of any noticeable trade-off between economies of scale in production and diseconomies in distribution in the case of telecommunications seems remote. At the same time, density economies may well be important. If there are large numbers of potential customers, or a large volume of traffic, within a relatively small area, it is more likely to be worthwhile making the substantial investments required; it is observable, for example, that high speed, high capacity optic fibre cables are only available in large cities where there is a high volume of traffic from business users, particularly in financial services and media.

A telling example from the developing world is mentioned in a recent article in the New Scientist (23 June, 2007, p.26):

> " 'Economists are head over heels in love with cellphones, but so far they have been a largely urban, big city phenomenon in the developing world' says Eric Brewer, a computer scientist at the University of California, Berkeley. 'The fact is that in rural areas, which by definition have a low population density, it's actually very difficult to deploy cellular base stations in an economically viable way.'
>
> Cellphone operators need enough users for each radio antenna tower to justify the cost of building and maintaining it. If a network cannot guarantee a threshold revenue per user, the tower will not be built. 'That tends to mean all cities will be covered for sure, and certain roads and railway lines. But not the rural areas.'"

Somewhat unexpectedly, it therefore seems that in this case, although very low operating costs for the distribution side of the business may seem to encourage dispersion, economies of scale in production and density economies are likely to mean that better services will be found where there is agglomeration.

4. Application to Point-type infrastructure

The "Buildings" section of **Table 2.1** in **Chapter II** includes a variety of facilities that can be identified as Point-type infrastructure. As a general proposition, it seems reasonable to suggest that such facilities (hospitals, schools, museums, etc) can be viewed as analogous to the production units of Area Type infrastructure, such as water treatment plants, power stations, etc. As such, it is likely that these facilities would be found, on investigation, to demonstrate economies of scale due to better utilization of large indivisible equipment (e.g. X-ray facilities, MRI equipment, operating theatres), more efficient use of skilled personnel (e.g. specialist teachers) or other factors (e.g. more complete collections of art or archaeology). However, it would not be right to conclude on these grounds that it would always be advantageous for such facilities to be made as large as possible so as to benefit from scale economies. The access costs of the users of the facilities should also be taken into account.

As regards access costs, reference has already been made to Arnott's (1979) basic result on commuting costs, and the assumptions on which that rests. On the same assumptions, the same basic algebra would apply to people traveling to access a facility such as a hospital or school. That is to say, access costs will be increasing in the size of the catchment area, unless these can be mitigated in some way. There is therefore a potential trade-off between economies of scale in the facility and higher access costs.

Of course, just as commuting costs are not generally a simple linear function of distance, and commuting trips are not all along radial links, the location of potential users of Point Type infrastructure and the access routes and modes of transport available to them will affect the trade-off in particular cases. So also will the nature of the facility: for example, fewer, larger units might be favoured for heart surgery, because they require expensive equipment and specialized personnel, notwithstanding that access costs incurred by users may on average be relatively high, as the catchment area will need to be large to ensure that the unit is fully utilized. However, the infrequency of the average user's need, and the high value of the procedure may mean that relatively high access costs are acceptable to users.

In fact, hospitals provide a particularly interesting example of the issues that arise with Point Type infrastructure. A leader in The Economist dated 11 March 2006, commenting on rising deficits in the UK National Health Service, observed (p.11):

> "The inefficient configuration of services is another reason why the red ink is appearing. Hospitals are doing things – such as diagnostics, some elective surgery and minor injuries – that might be done better in other places. And in some areas – in a ring around London, for instance – there are too many middling-sized hospitals offering treatment that could be provided more cheaply and safely at fewer, larger and more specialized hospitals."

The argument clearly indicates that The Economist believes there to be economies of scale in at least some kinds of hospital services – a belief which may well be correct, although The Economist cites no evidence for it. However, it is equally clear that The Economist's concept of an "inefficient configuration" takes no account of access costs for patients, quite an important consideration, one might think. Indeed, The Economist itself comes close to recognizing the issue later in the same article when it adds:

> "Ministers also need to grit their teeth and accept that, where the new market in health care reveals that hospitals are providing the wrong thing in the wrong place, some will have to close. There will inevitably be fierce local opposition, and the government will need to try to defuse that by providing more free-standing "A&E-lite" clinics to provide the emergency services that people reasonably expect to be available nearby when they need help quickly."

As this passage shows, a proper evaluation of the size and geographical disposition of health facilities ought to take into account access costs as well as scale economies in service provision. A *reductio ad absurdam* makes the point: Otherwise, why not just provide one gigantic hospital for the whole of England?

Just as the water companies have to take the existing pattern of settlement as given, so do the health authorities. It follows that the kind of results obtained from the analysis of water distribution costs are quite relevant to the determination of access costs to health facilities. That is to say, dispersed settlement patterns will imply high access costs; large, dense settlements, low access costs (as well as scale economies in production). The accessibility and relatively low cost of infrastructure services in the latter case can then contribute to the resurgence of urban areas as "consumer cities" (Glaeser & Gottlieb (2006)).

5. Application to Transport

While transport infrastructure features prominently in **Table 2.1** in **Chapter II**, it might not seem at first blush that the central theme of this thesis – the tension between economies of scale in production and diseconomies in distribution – is readily applicable to transport. After all, distribution costs are, for the most part, transport costs. Whether transport costs increase more or less than in proportion to the scale of operations, as measured by distance, say, or tonne-kilometres, depends on the particular case. Indeed, put in this way, there is probably a presumption in favour of economies of scale, as illustrated for example by bulk transport of commodities.

However, distribution does not typically involve just transport from fixed point A to fixed point B. It involves distribution to a number of destinations within a service area. The extra spatial dimension and the multiplicity of destinations undermines simple cost/distance relationships. Supermarkets, for example, try to optimize the number of depots in relation to retail outlets having regard to depot and holding costs as well as transport costs[97]. This problem can be seen as analogous to the water supply problem, with the depots taking the place of treatment plants and probably benefiting from economies of scale, which need to be traded off against diseconomies associated with expansion of the service area. At any rate, it requires an unusual combination of circumstances for a single depot serving a large market to be optimal. Moreover the size and location of supermarkets involves the same kind of trade-off between economies of scale and access costs for consumers as other Point Type infrastructure.

Other similarities appear if we consider transport services on a network. Railway stations, metro stations and bus stations perform a function similar to water treatment plants in bringing together and processing passengers who then travel, like water, along particular branches of the network to their destinations. A modified concept of scale economies could be developed to cover the functions of stations, and the unit cost of delivering passengers will depend on characteristics such as the size of the area to be served and population density – in particular, if the destinations are remote, widely separated or sparsely populated, unit costs will be higher, as attested by the difficulty of maintaining viable rural bus services, to take but one example.

A fully worked out treatment of these issues is beyond the scope of this research. However, we can draw attention to (a) parallels between the role of density in the provision of urban transport and its role in water distribution; (b) some recent work on the economics of aviation services which raises issues about the assessment of scale effects similar to those investigated in **Chapter V** on water distribution.

a. The role of density in urban transport

In the “new urbanism” literature, higher density is considered to be conducive to a more sustainable style of life, with respect to transport in particular. The idea, as Richardson & Bae (2004, p.255) express it, is that: “Higher densities may help to reduce automobile dependence by facilitating shifts to other modes (e.g. transit, bicycling or walking).” However, as the same authors point out: “To the extent that motorized modes dominate, higher densities mean more congestion and slower travel speeds.” Furthermore “There

[97] An additional consideration, not present in the water case, is the value of goods in transit. See e.g. McCann & Schefer (2004, p.184) “As the demand for delivery speed increases, the associated opportunity costs of lead-times also increase, and the average inventory levels maintained will fall.”

is a disconnect between the increasing emphasis on policies to make metropolitan areas denser and the overwhelming empirical evidence that most US metropolitan areas are becoming less dense … The experiences of Western Europe and many other parts of the world are similar." Richardson & Bae are therefore sceptical about the supposed sustainability benefits of urban compactness[98]. Commenting on the well-known work of Newman *et al* (1999) correlating automobile dependency negatively with density across 46 cities around the world[99], they observe "The negative relationship between automobile use and compactness is much more convincing in cross-sectional terms. But the rate of growth in automobile ownership in Europe and Asia is much faster than in the USA … The differential is much higher than can be explained by the acceleration of decentralization trends in these countries, so clearly there are other forces at work besides urban form."

Richardson & Bae's conclusion is perhaps too negative. They rather downplay the potential role of public transport arising from its ability to provide low cost, high capacity transit where the density of demand is sufficiently high (despite noting (p.257) that "in Central and Inner London in the UK, 60.3% of commuting trips are by public transit.") The obverse is equally compelling: low density settlement renders public transport less and less viable, for reasons entirely analogous to those applying to water distribution, as identified in **Chapter V**, i.e. higher infrastructure requirements and longer distances per unit of output. Looked at from the supply side, we find much to support a more positive view of higher densities. Admittedly, generalizing from our water supply findings omits some factors important in the transport context. First there is congestion: as cities expand, whether by densification or suburbanization, traffic on the existing transport infrastructure intensifies, leading to deterioration in service. This leads to a second issue: To ease congestion then requires more capacity or a step change in transport technology (e.g. rail or metro). Either way, the investment required is large and there is no automatic mechanism to provide it. Demand considerations might also qualify a conclusion in favour of high density, at least for some types of infrastructure – for example, a high income elasticity of demand for personal space could favour a more

[98] Richardson & Bae also note that although urban compactness is usually measured in terms of population densities, other measures exist such as radius of the urbanised area, median radial distance (the distance beyond which one half of the metropolitan population lives) and the compactness index of Bertaud & Malpezzi (1998).

[99] Who find (p.628) a coefficient of -0.744 when "Public transportation cost per passenger kilometer" is regressed against "Urban density (persons/ha)".

dispersed pattern of settlement even if this entails higher costs for infrastructure services[100].

b. Assessing economies of scale in aviation services

The work taken up for comment here does not concern itself with the economics of airports, only with the aviation services operated between airports, the number and location of which is exogenous. Thus it is concerned with the assessment of economies of scale in that part of the activity which is analogous to water distribution. As with much work in the transport field, output in this case is measured as a composite of quantity and distance, as we have done with water distribution.

The literature starts with an article by Caves *et al* (1984) entitled “Economies of density versus Economies of scale”. The approach taken is succinctly summarized by the authors (p.472):

> “The purpose of this article is to explore the apparent paradox of small air carriers with a purported unit cost disadvantage competing successfully against the large trunk carriers. We do this by developing a model of costs for airline services … Our model of airline costs is novel in that it includes two dimensions of airline size – the size of each carrier’s service network and the magnitude of passengers and freight transportation services provided. This allows us to make the crucial distinction between *returns to density* (the variation in unit costs caused by increasing transportation services within a network of given size) and *returns to scale* (the variation in unit costs with respect to proportional changes in both network size and the provision of transportation services).”

Using panel data from the US for 1970-1981, Caves *et al* find substantial returns to density for air carriers of all sizes but constant returns to scale (in the sense defined above) for both trunk and local airlines.

Since then, as Basso & Jara-Diaz (2006), observe (p.1):

> “For more than 20 years, the cost structure of transport industries in general, and the airline industry in particular, has been analyzed through the calculation of two indices: returns to density (RTD) and returns to scale (RTS), which were originally proposed by Caves *et al* (1984).”

However, Basso & Jara-Diaz go on to argue that “RTD and RTS should be replaced with three concepts: a corrected version of economies of density …; the multioutput degree of economies of scale …; and the degree of economies of spatial scope …”

[100] Although in practice this trade-off may be obscured by average cost pricing and/or the free or subsidised provision of some infrastructure.

In developing their critique of the standard approach, Basso & Jara-Diaz note (p.1) that the output of a transport firm is a vector of flows between many origin-destination pairs of the form $Y = \{y_{ij}\}$. They go on to note (p.5) that in practice the large size of Y precludes its direct use in empirical work. Instead it is necessary "to estimate cost functions using aggregate output descriptions $\tilde{Y} = \{ \tilde{y}_h \}$, which represents outputs and *attributes* such as ton-kilometers, seat-kilometers, average distance or load factor." Now, "when a network size variable, *N*, is included in the estimation, empirical studies of transport industries distinguish between two concepts of *scale*: returns to density (RTD) and returns to scale (RTS). In the former, it is assumed that the network is fixed when output increases; it is said that traffic density increases. In the latter, though, both output and network size increase, keeping traffic density unchanged."

As regards the measurement of RTD in this way, Basso & Jara-Diaz point out that there is an implicit assumption not only that the network size does not change but also that the route structure is unchanged (i.e. that the origin-destination pairs served remain the same). They propose that the term RTD be reserved for this case. It then measures whether the average cost increases more or less than in proportion to an increase in flows along existing links. For the case where the network size remains the same but route structure changes (because, for example, it becomes more economic to operate different links when flows increase), they suggest the term "multioutput degree of economies of scale (S)".

Turning to RTS (as defined by Caves), Basso & Jara-Diaz argue that an increase in network size necessarily implies a change in the number of underlying origin-destination pairs, which should be examined through a scope analysis. For this purpose, they therefore propose a different measure which they call "economies of spatial scope (SC)".

It is time to relate this work back to the assessment of scale effects in water distribution. First, the airline case confirms that such assessment is not straightforward when a product is distributed through a network and there is more than one scale attribute to consider (e.g. passengers carried, number of points served and trip length in the case of airlines; volume of water distributed, number of properties and size of service area in the case of water distribution). The answer will differ according to which attribute is the focus of attention. Next, we note the similarity between ε_w, the elasticity of distribution cost with respect to average water consumption, and RTD in Basso & Jara-Diaz 's definition. Both assess the effect of more intensive use of an existing network. There is less similarity in their other measures. Their *S* has no obvious parallel in water distribution, although it is possible to imagine a situation in which a water company might re-organise its distribution network in response to higher demand, by (for example) investing in a ring main for a large settlement, which would be somewhat analogous to an airline changing its route structure. Their *SC* does not quite correspond either to ε_N, the elasticity of distribution cost with respect to number of properties, or to ε_A, the elasticity of distribution cost with respect to size of service area; it could however be seen as a composite of the two since it aims to assess the effect of an increase in network size (as measured for example by number of points served).

In fact, one could go further and argue that although Basso & Jara-Diaz have dealt with one source of ambiguity, there remains ambiguity regarding network size. A case can be made for separating scale economies (or diseconomies) associated with number of points served (cf. number of connections in water supply) from scale economies associated with stage length or network miles (cf. service area or length of mains in water supply). The general lesson to emerge from this work is that measuring transport output as a composite of quantity and distance enables a wider range of scale effects to be investigated, a lesson relevant to the case of water distribution and to other activities where there is a spatial aspect to the delivery of services.

6. Implications for urban economics

The central question to be addressed here, in the light of the results for urban water supply, is how far the urban infrastructure can be considered to be a factor favouring agglomeration. There is an ambiguity in the term "agglomeration": it can mean either the *process* by which activity becomes more concentrated in space, or it can mean the *outcome* of that process. Taking the latter meaning first, it is not difficult to suppose that if a large settlement is well-endowed with infrastructure, new residents will be able to make use of it at rather low marginal cost. In this setting, infrastructure is indeed a "shareable input" (Eberts & McMillen (1999), p.1457) favouring agglomeration, although sharing is not costless, because of distribution or access costs.

It is less easy to show that infrastructure contributes positively to the *process* of agglomeration, creating a cumulative interaction whereby city growth leads to more infrastructure which in turn attracts more growth – a mechanism, in the words of Duranton & Puga (2004) "whereby different activities subject to small non-convexities gather in the same location to form a city". The problem lies in the weakness, or lack of automaticity, in the response of infrastructure to growth, due largely to the fact that infrastructure investments tend to be large and indivisible. There are plenty of examples of cities attracting incomers, at least partly on the strength of the various public facilities on offer. Often however this results in the existing infrastructure being overwhelmed, leading to deterioration rather than expansion and improvement[101]. In consequence, it is not uncommon to observe a cycle whereby city growth continues well beyond the capacity of the existing infrastructure (driven no doubt by the other agglomeration forces considered by Duranton & Puga), leading perhaps to a period when growth is checked. It then takes a massive effort to renew the infrastructure (which may be accompanied by, or spurred by, a step change in technology), paving the way for a new phase of growth. The mechanisms involved are as much political as economic. It may perhaps be the difficulty of modelling this kind of process in a satisfactory way that explains the limited attention given by urban theorists to infrastructure.

What is clear is that economies of scale in the production of infrastructure services are not the only factor at work. The spatial aspect, with its impact on distribution and access costs, is also important. In this research, we have tried to bring this aspect into focus by considering four contrasting urban growth scenarios, characterised as (a) *densification*, (b) *dispersion*, (c) *suburbanisation*, and (d) *constant density*. The results have been documented elsewhere in this thesis – particularly in **Chapter VI** – and need not be repeated here. The general conclusion from this work is that scale effects in infrastructure may depend as much on density as on size *per se*. While high density settlement has the potential to permit both large scale production and low cost distribution, more dispersed settlement patterns lead to higher (per capita) costs of distribution and access. It follows that the general presumption in urban economics that infrastructure services are always characterised by economies of scale and therefore conducive to agglomeration, may not be correct. This suggests that there should be more direct consideration of density effects in studies of urbanisation economies (by

[101] "Even leading cities have been brought to the brink of non-functionality in the not too distant past by failure to address problems as basic as waste disposal, air quality or security. The stench from the Thames had to become so bad that Parliament was suspended before a plan to improve London's sewerage was adopted in 1858; and it took the death toll of the smogs in the 1940s to stimulate action in the form of the Clean Air Act." Wenban-Smith (2000)

including density as an independent variable, or both population and area, or by using some measure of sprawl as a proxy for density).

7. Conclusions

Much of the man-made urban infrastructure can be seen as belonging to one of two broad types:

- **Area-type**: Provides services within a defined area (e.g. utilities, transport systems). In such cases, getting the service to users involves distribution costs;
- **Point-type**: Provides services at a specific point (e.g. hospitals, schools, offices, shops, museums, theatres, etc). In such cases, the equivalent consideration is the cost to users of accessing the facility.

Either way, there is potential for there to be a trade-off between economies of scale in production and diseconomies in distribution (or access); and whereas economies of scale in production and density economies would be conducive to agglomeration, diseconomies in distribution would act in the opposite direction. It is indeed precisely the interaction between these effects, i.e. economies of scale, distribution costs and density effects, that this research has aimed to elucidate, using water supply to provide illustration and quantification.

A number of conclusions have emerged about the appropriate methodologies to use when the aim, as here, is to estimate scale effects in water supply at settlement level.

- First, the quasi-fixed character of much of the capital invested in the water industry justifies the use of variable cost models, with capital treated in effect as a control variable. Indeed, in the case of water distribution, the lack of much choice of technology justifies the adoption of a Leontief-type production function, when no capital term is required.
- Secondly, the non-separability of water production and water distribution means that treating water supply as a single activity risks obscuring the distinctive characteristics of water distribution. Equally, it may not be valid to assume cost minimization at the production stage if (as is likely) there is interaction with distribution costs. There is merit therefore in examining water production and water supply separately, even if this means a somewhat clumsy procedure to analyse their interaction.
- Thirdly, the measurement of distribution output needs to capture in some way the spatial aspect of distribution. In **Chapter III**, a measure of distribution

output (*DO*) as the product of quantity consumed ($QC = w.N$) and the average distance to properties (φ) is developed. Conceptually, this is similar to the use of tonne-kms or passenger-miles in transport studies. The useful insights offered by this measure have been explored in **Chapter V**. In fact, an output measure of this kind may be preferable to a simple quantity measure in other studies of utilities, when distribution as well as production are under consideration.

- It follows that assessment of scale effects requires more than one elasticity to be considered. We draw particular attention to:
  a. ε_W – the elasticity of distribution cost with respect to consumption per property;
  b. ε_N – the elasticity of distribution cost with respect to numbers of properties;
  c. ε_A – the elasticity of distribution cost with respect to size of distribution area.

In this research, water supply has been seen as of interest, not just for its own sake, but also as a model for a wider range of types of urban infrastructure which, while characterized by economies of scale in production, also involve distributing additional output over wider areas (or enabling additional consumers to access the production facility – as with hospitals). The focus has therefore been on effects at settlement level, giving particular attention to the spatial aspects.

The objective then of the empirical work on water production and distribution brought together in **Chapter VI** has been to throw light on the determinants of costs at settlement level. This focus is different from that found in the mainstream utilities literature, where the objective usually is to study relative efficiency, although there are similarities in the methods used. In fact, it has been found that settlement level effects are hard to discern where, as in England & Wales, water companies are mostly rather large, serving areas comprising numerous cities, towns and villages. To overcome this problem, use has been made of a considerable amount of more detailed information provided in confidence for the purposes of this research by one large company (‘BWC’), which has enabled a much clearer picture of local effects to emerge. Further evidence has come from a 1996 survey of US water utilities by the AWWA. Water utilities in the USA are generally quite small, often serving a single community, and so rather suitable for studying settlement level effects. Some insight has also come from information in the June Returns to Ofwat by the smaller water only companies (WOCs) in England & Wales. Although quite large by international standards, these companies mostly serve areas centred on a single large town (e.g. Bristol, Cambridge, Folkestone, Portsmouth) and so also have some potential to yield results relevant to this research.

At the outset, it had been anticipated that while economies of scale in water production would be confirmed, diseconomies would be found in water distribution. It would follow that in urban water supply systems, a trade-off between these effects would be at work, qualifying the popular view that infrastructure services, such as water supply, are characterized only by economies of scale. In fact, a more complicated story has emerged, in which density plays as important a role as size.

It is conventional wisdom that there are economies of scale in water production. The evidence in **Chapter IV** confirms that this is indeed the case for water treatment works (WTWs), even when water acquisition is included, with returns to scale of the order of 1.25 on a full cost basis (and probably higher for operating costs). However, it is important to recognize that this finding applies at plant level. When two or more works are operated (for example, because the size of works is limited by the capacity of the water sources; or because the communities it serves are widely separated), these scale economies will no longer be very evident. Moreover, for borehole supplies, economies of scale are hard to discern, even at plant level. The benefits of large scale production can therefore only be reaped where circumstances permit the operation of large WTWs, typically where there is a large population and access to high capacity water resources. Birmingham, for example, which has a population of over 1 million and access to water from the Elan Valley is mostly supplied by a single large WTW (the Frankley works) leading to relatively low water supply costs for that city.

For water distribution, as has been widely recognized (e.g. Stone & Webster Consultants (2004)), the concept of scale has more than one dimension. The modeling and empirical estimation in **Chapter V** indicates that two aspects are particularly important: the volume of water distributed and some measure of the size of the service area. The volume of water is the product of numbers of properties and usage per property (and leakage) but if usage per property does not vary much from place to place, estimation of the volume scale effect will be much the same whether volume or numbers of properties is used. For the service area measure, the more obvious possibilities include the actual area and length of mains. As the former will often include areas of unserviced land, the latter is preferable. However, better still would be a measure which can capture the spatial distribution of properties and this is what our measure φ aims to do. This measure is derived by treating service areas as monocentric settlements of a size determined by the observed length of mains and property density for each area. This produces a measure of the average distance to properties, which can be applied flexibly to a wide range of actual situations. Although an approximation, it provides a versatile tool with which to represent the spatial aspect of distribution. Indeed it would appear to have the potential to be used more widely in urban studies.

Armed with this tool, the results summarized in **Table 5.7** are obtained. They indicate economies of scale in distribution with respect to volume but diseconomies with respect to average distance to properties. The implications for distribution costs then depend on how these influences balance out. While there are cases when there will be a trade-off between economies of scale in production and diseconomies in distribution, there will be other cases when volume economies in distribution as well as production will predominate. To explore these effects, comparisons can be made between contrasting hypothetical cases.

The results of this kind of exercise can be clearly seen in **Table 6.4**. Here four types of comparison between settlements are set up, labeled (a) *densification*; (b) *dispersion*; (c) *suburbanization*; and (d) *constant density*, in each case assuming a single WTW of the appropriate size. Now, as numbers of properties are increased in each scenario (leading to higher volumes, given constant usage per property), the key difference is how density is affected.

- With (a) *densification*, because the urban boundary does not change as property numbers increase, density increases in parallel, so that volume economies predominate in distribution as well as production. For example, unit water supply costs for a town doubled in size to 50,000 properties occupying 2,250 Ha (density 22.2 properties/Ha) will, according to these calculations, be 16.2% lower than for a town of 25,000 properties occupying the same area (density 11.1 properties/Ha), about half of the reduction coming from lower unit water production costs and half from lower unit distribution costs.
- With (b) *dispersion*, the number of properties does not increase, so that there is no volume effect, but the more dispersed pattern of settlement means lower density and an increasing average distance to properties, and hence higher distribution costs. For example, unit water supply costs for a town of 18,000

properties spread out over 2,090 Ha (density 8.6 properties/Ha) will be 10.8% higher than for a town of 18,000 properties occupying only 735 Ha (density 24.5 properties/Ha), all due to a 23.4% increase in unit distribution costs.

- With (c) *suburbanization*, the number of properties increases but because the increase is into less dense peripheral areas, average density falls and average distance to properties increases, albeit to a lesser extent than with (b). In this case, volume economies (in both production and distribution) are more or less balanced by average distance diseconomies. For example, unit supply costs for a town which has grown to 50,000 properties occupying over 20,000 Ha (density 2.4 properties/Ha) will be much the same as for the same town when it was only 15,000 properties occupying 985 Ha (density 15.2 properties/Ha) with the 25% reduction in unit production cost due to higher volume largely offset by a similar increase in unit distribution cost (the distance effect outweighing the volume effect in distribution here).
- With (d) *constant density*, the number of properties increases in line with the increase in area so that density is unchanged although the average distance to properties does increase. In this case, volume economies (in both production and distribution) outweigh the average distance effect. For example, unit supply costs for a town of 50,000 properties occupying 5,000 Ha (density 10 properties/Ha) will be 16.7% lower than for a town of 15,000 properties occupying 1,500 Ha (also 10 properties/Ha),about three-quarters of the reduction coming from lower unit production costs and one quarter from lower unit distribution costs.

These examples are enough to illustrate the range of effects that might be observed, but, it might be asked, which are particularly relevant when thinking about urban infrastructure? In studies of agglomeration, it is common to use population as the measure of size. One lesson from these examples is that it may not be sufficient to look at numbers alone. Whereas increase in size through densification would, it seems, bring economies of scale (in water supply at least), with a positive influence on agglomeration, as would (to a lesser extent) constant density expansion, increase in size through suburbanization would be roughly neutral in cost terms. To get the full picture, it would appear necessary to take density explicitly into account, not just size. Moreover, it would be misleading to regard urban areas of similar size, as measured by population, as equivalent from an agglomeration perspective, if they have very different densities. As the ‘dispersion’ example suggests, lower density towns or cities are likely to have higher distribution (and access) costs. Put differently, agglomeration by densification would have real cost advantages (at least up to the point where congestion costs become appreciable) whereas suburbanization would not.

Another way to look at the matter is to compare water supply costs as between a small town and a large one. Even if they have the same density, the ‘constant density’ calculations point to lower costs in the latter. If this effect generalizes to other types of infrastructure, it suggests an important reason why large settlements might over time prosper more than small ones; and if the larger one is also denser, the advantage becomes greater still. Of course, infrastructure costs are not the only consideration but if, for example, people have a preference for suburban living, these calculations indicate that there is likely to be a cost penalty (whether or not this is visited on suburbanites through tariffs and connection charges).

It has not been possible in this chapter to go beyond some pointers to the application of our water supply findings to a wider range of urban infrastructure. It is likely that distribution costs are less significant in the case of other utilities, although capital investment in distribution systems is important. While in general lower distribution costs can be expected to favour agglomeration by extending the area that can be economically served, high capital costs will still require that settlements be dense as well as relatively large if the necessary investments are to be viable. At the same time, we have pointed in **section 3** above to some developments, such as small sewage treatment works and local power generation, which may help small settlements. The scope for application to Point-type infrastructure, such as hospitals, appears good. There has been a tendency to disregard access costs in these cases but the methods we have developed for water distribution costs could readily be applied – the effect, it appears, given that health authorities (like water companies) have to take the existing pattern of settlement as given, would be to moderate enthusiasm for over-large facilities.

Application to transport is less obvious. While there are some suggestive similarities, notably when the spatial aspect of transport networks is under consideration, transport also raises issues which go beyond those examined in this thesis. An important instance is congestion, which hardly arises in the case of water supply[102] but is of considerable importance in transport. At the same time, the role of density in facilitating the provision of low cost, high capacity transit has parallels in water supply, as does the difficulty of maintaining viable public transport where density is low, for reasons entirely analogous to those applying to water distribution, as identified in **Chapter V**, i.e. higher infrastructure requirements and longer distances per unit of output.

Demand considerations might qualify these conclusions, at least for some types of infrastructure – for example, a high income elasticity of demand for personal space could favour a more dispersed pattern of settlement even if this entails higher costs for infrastructure services[103].

What is clear is that economies of scale in production are not the only factor at work. The spatial aspect with its impact on distribution and access costs is also important. In this research, we have tried to bring this aspect into focus by considering four contrasting urban growth scenarios, characterised as (a) *densification*, (b) *dispersion*, (c) *suburbanisation*, and (d) *constant density*. The results have been discussed in **Chapter VI**. The general conclusion emerging from this work is that scale effects in infrastructure may depend as much on density as on size *per se*. High density settlement has the potential to permit both large scale production and low cost distribution but more dispersed settlement patterns lead to higher (per capita) costs of distribution and access. This suggests that there should be more direct consideration of density effects in studies of urbanisation economies (by including density as an independent variable, or both population and area, or by using some measure of sprawl as a proxy for density).

[102] The drop in pressure which can occur at times of peak demand for water is perhaps the nearest equivalent.

[103] Although in practice this trade-off may be obscured by average cost pricing and/or the free or subsidised provision of some infrastructure.

Appendix A

USING THE OFWAT DATA FOR WATER SUPPLY (ENGLAND & WALES)

Each year, the water companies submit 43 tables of information to Ofwat to assist in the discharge of its regulatory duties. About half these tables relate to the companies' sewerage and sewage treatment functions, while others summarise information from later more detailed tables. Of the tables relating to water supply, those that have been drawn on in this research are briefly described below:

Table 7: Non-financial measures – Water properties and population

This table gives information for the reporting year, together with 5 previous years and projections for the next 2 years. The first line reports the number of **new properties** connected during the year. The next 10 lines give an analysis of **connected properties** by billing status (measured or unmeasured) and whether household or non-household. Finally, 5 lines analyse the **connected population** by billing status.

Table 8: Non-financial measures – Water metering and large users

This table gives information for the reporting year, together with 5 previous years. The first 9 lines give information about **household meter installations**; the next 3 lines give additional information about meter optants. The remaining 6 lines analyse **non-household consumption** by the amount of water taken, using 3 size categories: <100Ml/year; 100-250Ml/year; >250Ml/year.

Table 10: Non-financial measures – Water delivered

This table gives information for the reporting year, together with 5 previous years and projections for the next 2 years. The first 6 lines give volume of **water delivered** by billing status (household/non-household; measured/unmeasured). The next 26 lines give a detailed analysis of the components of water delivered (see **Appendix D** for a diagrammatic summary of how these components relate to each other). The key figures for this research are:

Line 29: **Total leakage**;
Line 30: **Distribution input**;
Line 31: Bulk supply imports;
Line 32: Bulk supply exports;

Line 33: Water treated at own works to own customers.

Table 11: Non-financial measures – Water mains activity

This table gives information for the reporting year, together with 5 previous years. The first line gives **total length of mains** at the start of the year; the next 8 lines report changes during the year (including information about renewals, relining, pipes replaced and bursts per 1000km) leading to total length of mains at end-year (line 21). The next 5 lines report information about **distribution zone studies**, while the last 8 lines report **other water service activities**, such as refurbishment work on aqueducts, reservoirs and water treatment works.

Table 12: Non-financial measures – Water explanatory factors

This table gives information for the reporting year only. The first 4 lines cover **source types** (impounding reservoirs, river abstractions, boreholes and total) giving in cols 1 and 2 the number of each type of source and the proportion of distribution input from each. Bulk imports are indicated in col. 3. The next 3 lines report **average pumping head** (resource, distribution and total). The next 7 lines indicate the number of plants and the proportion of distribution input by five **treatment types** (simple disinfection, W1 – W4); the 7th line gives the **total number of treatment works**. The next 11 lines indicate the number of plants and the proportion of distribution input by nine **capacity size bands** (< 1Ml/day; 1 – 2.5Ml/day; 2.5 - 5Ml/day; 5 - 10Ml/day; 10 - 25Ml/day; 25 - 50Ml/day; 50 - 100Ml/day; 100 - 175Ml/day; > 175Ml/day).

Table 21: Regulatory accounts (CCA) – Water service, Activity costing analysis

This table gives information for the reporting year only. It gives an analysis of **total operating costs**. The information given in this table can be summarized as:

Type of cost	**Water resources and treatment**	**Water distribution**	**Total**
I. Direct costs	√	√	√
II. Other operating expenditure			√
III. Reactive and planned maintenance (incl in I and II above)	(√)	(√)	(√)
IV. Capital maintenance	√	√	√
V. Other expenditure and adjustments			√
VI. Total operating costs			√

Thus, “Direct costs” and “Capital maintenance” are allocated between “Water resources and treatment” and “Water distribution” but other elements of total operating costs are not. As an indication of the magnitudes, the figures for Bournemouth & West Hants for 2003 are shown below. This shows that for this company around 75% of total operating costs could be allocated. The main items not allocated are in "Other operating expenditure" and include customer services (incl. billing), scientific services and local authority rates.

Type of cost	**Water resources and treatment**	**Water distribution**	**Total**	**%**
I. Direct costs	4.082	4.402	8.484	**41.7**
II. Other operating expenditure			4.478	**22.0**
IV. Capital maintenance	3.889	2.868	6.925	**34.1**
V. Other expenditure and adjustments			0.446	**2.2**
VI. Total operating costs			20.333	**100**

The total for "**Direct costs**" is made up of the following items:

1. Employment costs;
2. Power;
3. Agencies;
4. Hired & contracted services;
5. Associated companies;
6. Materials & consumables;
7. Environment Agency charges (Water resources & treatment only);
8. Bulk supply imports (Water resources & treatment only);
9. Other direct costs;
10. General & support expenditure.

Some of these items are potentially of interest in their own right.

There is a complication with "**Capital maintenance**" linked to the lumpiness of this type of expenditure and the depreciation policy for infrastructure assets adopted by water companies (due to change after 2005 as a result of a change in accounting standards). In consequence, capital maintenance is made up of three components:

A. Infrastructure renewals expenditure;
B. Movement in infrastructure accruals/pre-payment (which can be + or -);
C. CCA depreciation allocated.

As regards A and B, “The depreciation charge for infrastructure assets is the estimated level of annual expenditure required to maintain the operating capability of the network, which is based on the group’s independently certified asset management plan.”[104] Thus A + or – B is intended to provide a measure of the cost of maintaining the operating capability of the existing network and enters the accounts in lieu of a depreciation charge. The issue mainly affects distribution assets. The third component (C) is normal depreciation on non-infrastructure assets.

Table 25: Regulatory accounts (CCA) – Analysis of fixed assets by asset type

This table gives information for the reporting year only. The first 4 columns of the table (further columns report similar information for sewerage service assets) divide **water service assets** into:

- Water service infrastructure assets;
- Water service operational assets;
- Water service other tangible assets;
- Water service total

The first 6 lines then provide a reconciliation between the gross replacement cost value of assets at the year-end with the start-year value, involving an RPI adjustment, disposals and additions. The remainder of the table gives an analysis of the depreciation charge for the year (except for infrastructure assets – see note on Table 21, “Capital maintenance” above), leading to the net book value for non-infrastructure assets at year end.

It is worth quoting here the definitions of “infrastructure assets” and “operational assets”, which indicate that the former include assets related to water acquisition (e.g. dams and reservoirs) as well as water mains, while the latter also include some water acquisition assets (e.g. boreholes):

“**Infrastructure assets** cover the following: underground systems of mains and sewers, impounding and pumped raw storage reservoirs, dams, sludge pipelines and sea outfalls.”

“**Operational assets** cover the following: intake works, pumping stations, treatment works, boreholes, operational land, offices, depots, workshops, etc …”

[104] United Utilities Annual Report ’05, p.60 (Note 1(g) to financial statements). Other companies’ depreciation policies for infrastructure assets are substantively the same.

In consequence, the information on operating costs from Table 21 cannot consistently be linked with the information on fixed assets in this table (although an approximate re-allocation can be made – See **Appendix C**).

Annex to Appendix A

Data used in Chapters IV and V, and Appendix E

For ease of reference variables are listed here in alphabetic order, rather than in order of appearance in the text or grouped by equation, but separately for water treatment and water distribution (which results in some items being duplicated).

a. Water treatment

AQP = Average quantity treated in a works (Ml/day)
AQP = QP/TN (see below) – may be for boreholes and WTWs separately.

BH_{prop} = Proportion of borehole water input to works
From JR Table 12, col 2.

FCP = Financing costs, water resources and treatment (£m/year)
Obtained by multiplying regulatory capital value $\overline{K_P}$ by allowed rate of return.

CMP = Capital maintenance, water resources and treatment (£m/year)
From JR Table 21, col 1, lines 25-27.

$\overline{K_P}$ *= Regulatory capital value attributable to water resources and treatment (£m)*
For derivation see **Appendix C**.

PHR = Resource pumping head (metres)
From JR table 12, col 4, line 6.

p_i = Proportion of output from works in the ith size band
From JR table 12, col 1, lines 21-30.

QDI = Water put into distribution (Ml/day)
From JR Table 10, line 30 ("Distribution input").

QP = Total water produced (Ml/day)
This is generally taken to be equal to "Distribution input" (JR table 10, line 30), as it appears that bulk imports or exports in England & Wales are generally of untreated water.

SP = Proportion of surface water treated
From JR table 12, cols 2 and 3, lines 1 and 2 (i.e. "Impounding reservoirs" and "River abstractions", including any bulk imports from either source).

TCP = Total water resource and treatment costs (£m)
This total is made up of "Direct costs" (JR table 21, col 1, line 12) and "Capital maintenance" (JR table 21, col 1, lines 25-27).

TN = number of treatment works (No)
From JR table 12, col 2, line 31.

UVCP = Unit variable cost, water resources and treatment (£/Ml)
UVCP = VCP/QT (see below/above)

UTCP = Unit total cost, water resources and treatment (£/Ml)
UTCP = TCP/QT (see below/above)

VCP = Water resource and treatment variable costs (£m/year)
From JR Table 21, col 1, line 12 ("Direct costs – functional expenditure").

W4D = Proportion of water receiving level 4 treatment
From JR Table 12, col 1, line 13.

b. Water distribution
CCD = Capital costs, water distribution (£m)
Total of CMD (see below) and FCD (regulatory capital value of assets attributable to distribution – see **Appendix C** – times allowed rate of return).

CMD = Capital maintenance, water distribution (£m/year)
From JR Table 21, col 2, lines 25-27.

$\overline{K_D}$ *= Regulatory capital value of assets attributable to distribution (£m)*
For derivation see **Appendix C**.

L = Leakage (Ml/day)
From JR Table 10, line 29 ("Total leakage")

M = Length of water mains (km)
From JR Table 11, line 1.

PHD = Distribution pumping head (metres)
From JR Table 12, col 4, line 6 ("Average pumping head – distribution").
And $APHD = \sum_k PHD_k / k$, where *k* is the number of companies.

N = Number of properties (`000)
This is the sum of "Household properties" (JR Table 7, line 15) and "Non-household properties" (JR Table 7, line 19). It thus omits "Void properties".

QC = Volume of water reaching customers (Ml/day)
Approximated as Distribution Input (JR Table 10, line 30) *less* Total Leakage (JR Table 10, line 29).

QDI = Water put into distribution (Ml/day)
From JR Table 10, line 30 ("Distribution input").

UCCD = Unit capital cost, water distribution (£/Ml)
UCCD = CCD/QC (see above)

UTCD = Unit total cost, water distribution (£/Ml)
UTCD = TCD/QC (see above)

UVCD = Unit variable cost, water distribution (£/Ml)
UVCD = VCD/QC (see below/above)

VCD = Water distribution variable costs (£m/year)
From JR Table 21, col 2, line 12.

w = Water consumed per property (litres/property/day)
w = QC x 1,000,000/(N x 1,000)

Appendix B

ESTIMATING SCALE EFFECTS FOR UTILITIES: A SELECTIVE REVIEW OF THE LITERATURE

1. 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[105]. Schmalensee (1978) seems to have been among the first to give systematic consideration to the economics of distribution through a network. We therefore start in **Section 2** with a summary of his contribution. Unfortunately, it does not lend itself readily to empirical testing. More useful is the approach first developed in the context of electricity supply, using cost functions. The contributions of Nerlove (1963), Roberts (1986) and Thompson (1997) to this literature are therefore reviewed next in **Section 3**. We note that despite making important advances, the characterization of the distribution stage of electricity supply remains weak.

The economics of water supply has not generated a large academic literature. However, the trickle of articles has increased somewhat in recent years, particularly in the context of the regulation of privatised utilities. In this more recent work (from the mid-1980s), much of which closely follows the methodologies developed for electricity, the use of flexible form cost functions to investigate economies of scale has become standard. **Section 4** reviews the work of Kim & Clark (1988), Garcia & Thomas (2001), Stone & Webster Consultants (2004), Saal & Parker (2005) and Torres & Morrison Paul (2006), which is in this tradition. **Section 5** then reviews some other contributions to the literature which appear relevant to our concerns. These include Clark & Stevie (1981), which is an attempt to apply Schmalensee's approach; the econometric models developed by the UK water regulator, the Office of Water Services (Ofwat) to assess the relative efficiency of the water companies in England & Wales; Duncombe & Yinger's (1993) analysis of returns to scale in fire protection services; and the "Public Facilities

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

Location" literature. In our judgement, the work reviewed in Sections 4 and 5 is little more successful in elucidating the economics of water distribution than was the work in Section 3 in respect of electricity distribution. Finally, **Section 6** seeks to draw out lessons from the reviewed literature to inform the empirical work which constitutes the core of this thesis and which is reported in **Chapters III** to **VI**. In particular, it concludes that to get a clearer picture of the economics of urban water supply, it is necessary to examine water production and water distribution separately even though this means adopting somewhat *ad hoc* methodologies to bring the two stages of water supply together. It is important also to keep in mind that the objective of this research are different from those in the mainstream utilities literature, in that the focus is on settlement (not company) level effects.

2. Schmalensee's approach

Schmalensee (1978) expressed concern (p.271) 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 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:

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

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:

$$TC = \int_0^R c[r, Q(r)] dr \qquad \ldots\ldots\ldots \qquad \text{(B.2)}$$

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. Thus, in practice, other approaches have been used in empirical work on utilities. A number of relevant contributions are reviewed below, starting with electricity as this is where the dominant approach using cost functions was pioneered

3. 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)[106], 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:

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

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:

$$\ln C = \beta_0 + \beta_q \ln Q + \beta_K \ln P_K + \beta_L \ln P_L + \beta_F \ln P_F + u_i \qquad \text{…………} \qquad \text{(B.4)}$$

[106] The exposition here follows Greene closely.

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

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 of consideration 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:

$$T(K_G, M_G, E_P, K_D, M_D, Q) = 0 \qquad \ldots\ldots\ldots \qquad \text{(B.5)}$$

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[107].

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:

$$T\{E_I(K_G, M_G, E_P), K_D, M_D, Q\} = 0 \qquad \ldots\ldots\ldots \qquad \text{(B.6)}$$

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 …"

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

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

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

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:

[107] 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.

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 (B.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:

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

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:

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

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*[108] 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." At first sight, this might appear difficult to reconciled with the assumption, noted above, of constant unit costs in power generation. However, it is quite plausible that the marginal cost of supplying additional electricity to existing customers through the existing network is little more than the cost of generation, implying economies of scale in distribution in this case. 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.

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

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

> "… 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.

d. 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)[109] 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[110]. However, if there are scale economies in the production stage but diseconomies of scale in distribution, this assumption is inappropriate. Transferring attention from electricity to water supply, the point can be simply illustrated by reference to the diagrams in **Figure B.1** below:

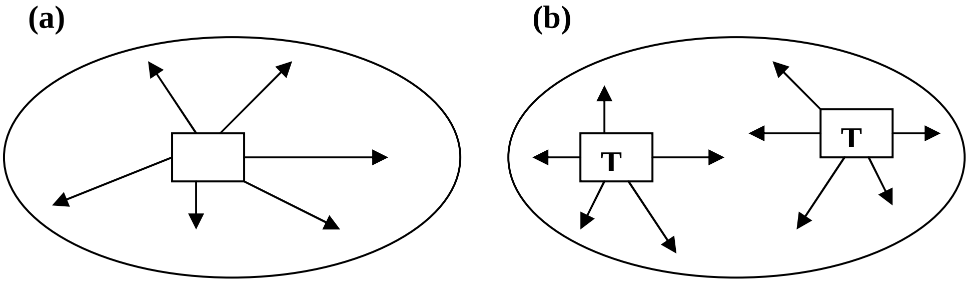

Figure B.1: Water supply: Should this area be served by (a) one treatment works or (b) two (or more) treatment works?

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.

[109] See Chambers (1988) pp.41-48 on separability in production functions and pp.110-119 on separability in cost functions.

[110] A similar assumption is made by Duncombe & Yinger (1993) in their two stage specification of a cost function for fire protection.

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 as it is a key part of the hypothesis being investigated in this research, this potentially important element of the situation will be missed if one proceeds to try to estimate scale economies in water supply with a cost function specification incorporating Roberts' assumption of separability.

With this observation in mind, work specifically concerned with water supply can now be reviewed.

4. Use of cost functions in analysis of water supply

a. Kim & Clark (1988)

Kim & Clark's initial characterisation of water supply seemed to provide welcome confirmation that there is much to be said for examining water distribution separately from water acquisition and treatment (although their article gives little attention to water acquisition). They state:

> "Many engineering cost studies have pointed to practically unexhausted economies in water treatment. In this regard, as far as plant size is concerned, the dominant view has been that 'biggest is also best'. However, an important factor limiting the growth of plant size in water supply is market size or distribution system which might offset economies of plant size. The problem then involves a trade-off of scale economies in production versus diseconomies in distribution, which affects the choice of the optimal size, location and distribution patterns of one or more plants. The trade-off between plant size and distribution diseconomies also goes to the heart of the matter of determining an optimal service area."

However, instead of then pursuing this perception in their specification of the cost function for water supply, they put their main emphasis on viewing water supply as a multi-product activity with two outputs: residential supply and non-residential supply. As both involve treatment *and* distribution, the distinctive economics of distribution are obscured.

Kim & Clark's postulated translog multi-product cost function includes on the right hand side variables to represent the two outputs, input prices for labour, energy and capital and two "operating variables" service distance and capacity utilization. Their introduction of the service distance variable is “to take into account the spatial variation in demand” (p.481). Service distance is defined in an earlier paper by Kim (1985) as “the total number of miles of pipe in the utility service area”[111]. It is therefore through this variable that distribution costs must get picked up. There is no attempt to account for leakage and the output variables are quantity treated rather than quantity delivered.

In their empirical implementation, using 1973 EPA data for 60 US water utilities, Kim & Clark find a marked difference in scale effects as between residential and non-residential supply, with rising average and marginal costs for residential supply but falling costs (over most of the range) for non-residential. They comment “This implies diseconomies of scale associated with supplying water to residential customers … [but] … substantial economies of scale for non-residential water supply.” (p.492). They also note that “marginal and average incremental costs for residential customers are uniformly greater than those for non-residential customers throughout the range of output. This is in accord with prior expectations, due to a larger distribution network (service lines and connections), higher system losses, and the large number of smaller customer accounts associated with supplying water to residential customers” (p.493). They further note that the degree of overall scale economies varies with the level of output, with small utilities showing rather marked economies of scale, large utilities moderate diseconomies of scale and the average utility more or less constant returns to scale.

They then turn their attention to the question whether economies in the treatment of water may be offset by diseconomies in the distribution of water. This is assessed by examining the effect of the service distance variable on scale economies, which is shown in **Table B.1** below:

	Utility size		
Returns to Scale (1/ε)	**Small**	**Average**	**Large**
With distance fixed	1.99609	1.26939	1.15210
With distance varying	1.33296	0.99226	0.87503

Table B.1: Effect of distance on overall scale economies
(Kim & Clark (1988), Table 4, p.499)

[111] In Kim & Clark (1988) service distance is stated to be “the distance from the treatment plant to the service area” but as the mean value of this variable is 539.5 miles, it seems clear that this is a mis-statement.

Kim & Clark comment: “As is clear from the table, utilities experience economies of scale in the treatment of delivered water, as exemplified by their values greater than one with fixed distance. However, we can immediately see the pronounced effects of distance in the determination of overall scale economies. The scale economies achieved in water treatment are by and large lost in the distribution of water” (p.499). They go on to note that service area and distance are loosely related to size of utility, and show that although large utilities enjoy considerable economies of scale in the treatment of water relative to small utilities, they also suffer from substantial diseconomies due to the size of the area being served. Kim & Clark’s work thus lends support to the idea that there is a trade-off between economies of scale in production and diseconomies of scale in distribution.

b. Garcia & Thomas (2001)

Like Kim & Clark, Garcia & Thomas (2001) view water supply as a multi-product activity but their two products are water delivered to customers (whether residential or non-residential) and water lost through leakage in the distribution system. This unusual approach is justified on the grounds that water managers, in responding to increases in demand can choose between increasing production or cutting back on leakage. It can also be seen as recognising to some extent the distinctive economics of distribution. It would seem more natural however to regard leakage as part of distribution costs. Garcia & Thomas are nevertheless correct that leakage costs can be optimised having regard to supply costs on the one hand and repair costs on the other[112].

Garcia & Thomas suggest that municipal water supply in France has five main functions: Production and treatment; transfer; stocking; pressurisation; and distribution. They comment (p.9) that: “It is difficult to provide an adequate representation of the water supply technology by means of a representative utility cost function, as the technical environment within which utilities operate is very different. Water utilities have first to be distinguished depending on the origin of the resource: groundwater or surface water. Groundwater use implies higher drilling and pumping costs, whereas treatment costs are usually higher with surface water. Differences in average costs are also found depending on the distribution process, on the size of the utility area, population per mile of water pipeline, and so on. Therefore, it is necessary to deal with

[112] In England & Wales, this is made explicit in guidance issued by Ofwat on the “Economic Level of Leakage” (ELL) – Ofwat (2002, revised 2003).

such heterogeneity by incorporating in the cost function, along with prices and outputs, variables that represent capital stock (production and treatment stations, storage facilities, pumping stations and pipelines) and technical environment (number of municipalities and customers served by the utility)."

In developing a specification, Garcia & Thomas argue that in this case capital stock is a quasi-fixed input in the sense that its modification in the short-run is either not feasible or is prohibitively costly. Their cost function is therefore a short-run one of the general form:

$$C_{SR}(y, w_v, w_K; K, Z) \qquad \ldots\ldots\ldots\ldots.. \qquad \text{(B.11)}$$

Where y is a vector of outputs (water delivered to customers, water lost in distribution), w_v is a vector of variable input prices (labour, electricity, materials), w_K is the price of capital, K is a vector of the elements of the (fixed) capital stock (network length, production capacity, storage and pumping capacity) and Z is a vector of "technical variables" (number of metered connections, number of communities served). Garcia & Thomas then put some emphasis on analysing "returns to network density", arguing that "the inclusion of variables such as the number of customers and the number of municipalities serviced by a single utility in the case of a district allows a distinction to be made between economies of density and economies of scale, that reflect the different ways production may increase" (p.12). They then define *economies of production density* as arising when average variable costs decrease when production increases, for a given network size and a given number of customers (i.e. demand per customer increases). In contrast, *economies of customer density* arise when new customers are connected to the existing network, demand per customer remaining constant; and they develop in addition a long run version of this measure to address the situation when adding new customers to the existing network requires consequential adjustments, such as additional pumping and storage capacity. They do not consider the effect on costs of an expansion of the service area although they do try to assess the effect of combining existing operations into larger water districts.

For estimation purposes, (B.11) is specified as a translog cost function, with cost share equations added to complete the system to be estimated. Implementation then follows using a 2 stage GMM procedure, leading to the estimation of 66 parameters. Among the conclusions reported are:

1. There are cost advantages not to minimise network water losses, because the labour and material cost involved in repairs is significantly higher than the energy cost associated with increasing production;
2. Elasticity of production density (short run) declines from about 1.4 for utilities with a low volume delivered per customer to about 1.0 for utilities with a high volume delivered per customer. That is to say, there are, on this measure, scale economies at low volumes but constant returns to scale at high volumes;
3. Elasticity of customer density (short run) declines from about 1.2 for utilities with a low density of customers per km of network to about 0.9 for utilities with a high density of customers per km of network. That is, there are economies of scale on this measure at low customer densities but not at high customer densities;
4. On returns to scale, "merging local communities in a water district of up to 5 communities seems to be profitable, whereas the gain in merging a higher number of communities is not clear. Moreover, creating a water district seems to be less profitable for local communities with low population density" (p.27).

While these conclusions seem broadly plausible, it is possible to wonder whether the sophistication of the econometric methods deployed is really justified. In particular, the variation between the years 1995, 1996 and 1997 in the French data is likely to be small so that it may come close to repeating essentially the same data for the 55 utilities three times, casting some doubt on the reported results (with only 55 data points, it would not be possible to estimate 66 parameters). No attempt has been made to distinguish between utilities using different proportions of ground and surface water; and the capture of distribution effects through numbers of connections and numbers of communities served seems inadequate. Nevertheless, with its innovative treatment of distribution losses, this is a useful addition to the literature on the economics of water supply.

c. Stone & Webster Consultants (2004)

In a report for Ofwat on economies of scale in the water industry in England and Wales, Stone & Webster Consultants (2004) – hereinafter S&W – also adopt a multi-product framework for their analysis, specifying a translog cost function for their main results. Their task was complicated by the need to apply their analysis to combined water and sewerage companies (WaSCs) as well as companies involved in water supply only (water only companies, or WoCs) – the work of Kim & Clark and Garcia & Thomas reported above related to water only companies, in USA and France respectively. Evidently, the case for viewing WaSCs as multi-product operations is strong. Water supply and sewerage each have their own operating systems, only coming together in the premises of customers where clean water becomes dirty after use; sewers also often carry rainfall run-off and fulfil other drainage functions. There may however be some commonality in functions such as the repair and maintenance of underground pipes and in overhead functions such as billing and research. The question whether or how strong are economies of scope is thus highly pertinent from a regulatory point of view and justifies the use of a multi-product specification in analysing WaSCs.

Apart from the issue of economies of scope, S&W note that (p.10): "The concept of scale in the context of water service provision has a number of dimensions. Production may be measured in terms of the volumes of water and wastewater delivered and collected, in terms of the number of connections or population served or in terms of the supply area covered. Water companies with a similar scale, as measured by some physical measure such as the number of connected properties, may have very different cost characteristics because of differences in the density of those connections. This means that *economies of density* must be considered simultaneously with economies of scale …"

S&W also note that the standard cost function assumes that companies are free to adjust in the long run the level of all factor inputs to ensure that costs are minimised but that this is not very appropriate in the water company context because the technology used in water services can be indivisible, with very long service life, and there are legal obligations to meet quality standards or connect customers to network systems. They therefore argue (following Garcia & Thomas and others) that it may be more appropriate to treat capital as a quasi-fixed input. Their cost function is therefore basically the same as (B.11), with the justification (pp.14-15): " This variable cost function satisfies the same properties as the long run function, without imposing the assumption that quasi-fixed inputs such as capital have been optimally chosen by the firm. Hence, from an empirical viewpoint, estimation of the variable cost function will yield the same economically relevant information contained in the underlying production technology, but without the risk of mis-specification because the level of observed capital inputs have not been optimally determined … The variable cost function will reflect the same information underlying technological relationships that govern the relationship between costs and outputs … The modelling of variable costs therefore provides a way of distinguishing between *short-run* and *long-run* economies of scale.” S&W also follow Garcia & Thomas in their definition of *economies of production density* and *economies of customer density*.

S&W’s system of equations to be estimated then is a translog cost function of the general form:

$$\ln C = \mathrm{C}(\ln W, \ln K, \ln Y, \ln Z, D, t) \quad \ldots\ldots\ldots\ldots \quad (\mathrm{B.12})$$

together with input cost share equations:

$$S = \mathrm{S}(\ln W, \ln K, \ln Y, \ln Z, t) \quad \ldots\ldots\ldots\ldots. \quad (\mathrm{B.13})$$

Where C is variable costs, W is a vector of variable input prices, K is a vector of quasi-fixed capital inputs, Y is a vector of outputs, Z is a vector of “hedonic variables” (environmental and operating characteristics), D is a company dummy and t is a time trend.

In their analysis of water supply, S&W take the principal outputs to be volumes of water delivered and number of properties for water supply. However, they felt that additional aspects needed to be considered. S&W addressed this by adopting a graduated approach, starting with a simple output model and then testing for improvements in model significance (using a Chi test) as additional variables were introduced. The results are summarised below:

Outputs specified	**Chi test (short run model)**	**Chi test (long run model)**
I. Base model – water delivered only	27.88	8.61
II. Water delivered + water connections	1.61	4.40
III. As above + distribution losses	1.39	2.80
IVa. As above + water quality hedonics	0.28	0.53
IVb. As above + metering hedonics	0.19	0.30

[Source: Adapted from Stone & Webster Consultants (2004), Tables 9 and 11.]

Table B.2: Model significance of different versions of S & W’s model

S&W conclude that the model specification is improved by adopting a multi-product approach. The base model (I), they suggest, provides estimates of scale economies which are comparable to the estimates of *economies of production density* in Garcia &

Thomas. Model II in which connected properties feature as an additional output provides estimates of scale economies based on changes in both production and customers served. (Using numbers of connected properties can be seen as a move towards recognising the different characteristics of water distribution.) In model III they follow Garcia & Thomas in treating distribution losses as another output. Finally, a number of “hedonic” variables are introduced to control for “differences in service quality and characteristics of the operating environment for companies”. These hedonic variables cover compliance with drinking water standards, water pressure, supply interruptions, % of properties metered, average pumping head and % of water from river sources. Generally, S&W conclude that it is appropriate and necessary to include hedonic variables in the estimated cost functions.

S&W’s results are summarized in **Table B.3** (for WaSCs, covering both water and sewerage operations) and **Table B.4** (for WOCs, covering just water supply). S&W’s scale parameter is defined so that a value greater than one indicates economies of scale; a value less than one indicates diseconomies of scale.

	Short run Economies of scale		**Long run Economies of scale**	
	Parameter	**S.E.**	**Parameter**	**S.E.**
I. Base model (outputs only)	1.01	0.06	0.93	0.18
II. Base model + operating hedonics	0.91	0.05	0.80	0.19
III. Base model + connections	0.76	0.10	0.77	0.17
IV. Base model + connections + operating hedonics	0.67	0.07	0.62	0.16

Table B.3: S&W’s estimates of short and long run economies of scale for water supply and sewerage operations of WaSCs (Adapted from Stone & Webster Consultants (2004), Tables 8 and 10, pp. 40-41)

	Short run Economies of scale		**Long run Economies of scale**	
	Parameter	**S.E.**	**Parameter**	**S.E.**
I. Base model (water delivered only)	1.42	0.08	1.25	0.09
II. Base model + connections	1.10	0.08	1.13	0.06
III. Base model + connections + distribution losses	1.09	0.08	1.11	0.07
IV. As III + water quality hedonics	1.04	0.08	1.05	0.07
V. As IV + metering hedonics	1.04	0.10	1.06	0.11

Table B.4: S&W’s estimates of short and long run economies of scale for water supply operations of WOCs (Adapted from Stone & Webster Consultants (2004), Tables 9 and 11, pp. 40-41)

The most striking feature of these results is the finding of significant diseconomies of scale for the combined water and sewerage operations of WaSCs in the preferred model – Model IV in **Table B.3**. For WOCs, the preferred model – Model V in **Table B.4 –** produces a result (for water supply only) which is not significantly different from constant returns to scale. It is also noticeable that adding “connections” as an explanatory variable leads to a sharp drop in the estimated scale parameter for both WaSCs and WOCs, which can perhaps be interpreted as some kind of *dis*economy associated with numbers of connections (and/or density?).

At the same time, it does not appear that S&W’s specification allows directly for the possible effects on distribution costs of differences in population densities, since the size of the supply area (or some proxy for it, such as length of mains) does not feature in the analysis – indeed, the inclusion of a company specific dummy variable in S&W’s specification has probably removed any such effect. Also the large size of the typical water company means that scale effects at a more local level, such as urban settlements (which are the focus of interest in this research) are not apparent. Nor does it appear that analysis at this level of aggregation allows any interaction between economies of scale in production and possible diseconomies in distribution to be identified. Nevertheless, this study is by far the most rigorous investigation of scale economies in the water industry of England & Wales yet to appear.

d. Saal & Parker (2005)

David Saal has written extensively about performance assessment in the water industry in England & Wales post-privatisation and he is a leading authority on the subject. He was in fact also involved in the Stone & Webster report cited above. The reason for including this particular reference here is to draw attention to a couple of additional issues relevant to the research reported in this thesis.

Saal & Parker note that Ofwat’s own assessment of the relative performance of companies[113] employs a set of cross sectional models at the function level (water distribution, water treatment, etc.) which are then aggregated in order to generate separate assessments of a company’s performance in water operations and sewerage operations. This means that Ofwat has implicitly assumed that the water operations of a

[113] Ofwat’s econometric models are discussed below in **section 5 (b)** of this Appendix.

WaSC are fully separable from its sewerage operations because it has assessed the water operations of both WaSCs and WoCs against jointly estimated common frontiers.

At the same time, Ofwat and other regulators have shown interest in alternative approaches, such as the panel based assessments done by Stone & Webster. In this paper, Saal & Parker explore the potential for using a panel input distance function stochastic frontier model[114] to assess the overall water operations performance of both WaSCs and WoCs in a single model.

It is assumed that technology can be represented by a translog input distance function. Output is modelled as two multiple outputs, water supplied and number of connections. "Such a specification is appropriate if we consider that there are distinct output characteristics associated with the physical volume of water supply, as opposed to the provision of connections to the water network. Moreover, the input requirements of providing a new network connection are substantially different from the input requirements of delivering additional water to an existing customer" argue Saal & Parker. This specification also goes some way towards distinguishing between water treatment and distribution. The inputs are specified as: (i) "fixed physical capital stock based on the modern equivalent asset (MEA) estimation of the replacement cost of water operations net tangible fixed assets, as provided in each water company's annual regulatory accounts"; and (ii) "variable input usage … measured as a company's total water service opex costs as reported in the regulatory accounts, … deflated using the ONS producer price index for materials and fuel purchased in the collection, purification and distribution of water industry." In addition, three variables are included to account for the potential impact of exogenous operating characteristics: (i) network density measured as the total water population served per kilometre of water mains; (ii) average pumping head as reported in Ofwat's regulatory returns; and (iii) a water quality index, defined as the average percentage of each company's water supply zones that are compliant with nine key water quality parameters drawn from the Drinking Water Inspectorate's annual reports on water quality[115]. These exogenous variables were tested in logged and squared logged form in the specification.

[114] See Knox Lovell & Schmidt (1988) for a general introduction to stochastic frontier methods. Saal, Parker & Weyman-Jones (2004) demonstrate how such methods can be adapted to assess the company level performance of English & Welsh WaSCs.

[115] This index was developed by Saal & Parker in earlier publications.

Saal & Parker's data covers the 11 years 1993 to 2003, during which time the number of WoCs declined from 20 to 12, while the number of WaSCs remained stable at 10. They note that "WaSCs have statistically lower density and quality compliance than the WoCs". They suggest that the difference in density can be explained by the fact that WoCs tend to be concentrated in relatively urban areas while the WaSCs have responsibility for many rural parts of England and Wales as well as urban areas; the difference in quality compliance may relate to the legacy of public ownership.

The results of this study are of considerable interest. Here, we focus on two aspects. First, a dummy for WaSCs carried a significant negative coefficient "suggesting that *ceteris paribus*, the input requirements for a WaSC are substantially higher than for a WoC, thereby suggesting a systematic difference between these types of companies." Saal & Parker conclude that "it is in fact inappropriate to assume that the underlying frontier for WoCs and WaSCs is the same." "Therefore, while this model clearly demonstrates the potential for employing panel stochastic frontier techniques in assessing water operations performance, it also suggests that it is inappropriate to jointly assess the performance of both WaSC and WoC operations within this framework … Moreover, as previous research … has demonstrated substantial cost interactions between water and sewerage operations, the inappropriate assumption of separability between WaSC water and sewerage operations in our model may at least partially explain why the WaSC and WoC frontiers for water operations are different from one another."

The second aspect of interest is the influence of density. Saal & Parker find a positive coefficient on *ln(density)* implying that as density increases, input requirements decline. And this is the case for WaSCs and WoCs separately as well. However, the coefficient on the square of *ln(density)* is negative so that the overall elasticity of input requirements with respect to density declines in magnitude and becomes positive for the 7 observations in the sample with density more than 49% higher than the average sample firm. Saal & Parker comment: "These results therefore suggest that increased density reduces input requirements, but the benefits of reduced customer dispersion are eventually offset by higher input requirements, perhaps associated with greater input requirements in heavily urbanised areas." But there is a difference between WoCs and WaSCs in this respect. The coefficient on the square of *ln(density)* is negative for the latter but positive for the former. "This suggests that for WoCs, increases in density always result in reduced input requirements, which may relate to the WoCs higher average density, or alternatively may suggest that costs associated with distribution are significantly more important for the WoCs.” Representations by some of the larger WaSCs to Ofwat about the need to take into account the higher costs of working in large conurbations (e.g. higher labour costs and congestion) may point to another explanation.

e. Torres & Morrison Paul (2006)

In this recent study, the authors focus on the significance of “output density” picking up the earlier remark by Schmalensee (1978)[116] that for network utilities “cost must in general depend on the entire *distribution* of demands over space”. Commenting on the scope for consolidation of the very numerous small community water systems in the US, they observe (p. 105):

> “ … any consolidation policies … must recognise that the resulting firms will [not only] produce larger water volumes but will also have to deliver water to more customers through larger service areas. That is, they must take into account potentially significant cost trade-offs involving water production and the network size, which depend on output density relative to customer numbers and service area size.”

They then proceed to develop and empirically implement a cost structure model of the US water utility industry.

They propose a short run transformation (production) function of the general form:

$$t(Y,V,\overline{X},Z) \qquad \ldots\ldots\ldots\ldots.. \qquad \text{(B.14)}$$

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), $\overline{X}$ is a vector of quasi-fixed inputs (e.g. storage and treatment capacity – this is what makes the approach short run) and *Z* is a vector of technical/environmental characteristics. This leads to the short run cost function:

$$VC(Y,P,\overline{X},Z) \qquad \ldots\ldots\ldots\ldots \qquad \text{(B.15)}$$

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*”. They thus avoid assuming that the number or scale of existing works is optimised. From this short run cost function, the vector of cost-

[116] See **section 2** above.

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 exogenous, i.e. water companies are obliged to supply whatever is demanded [117]) by adding to the system of equations, the identity:

$$Y_f = GS + X_{Pw} - Y_w - loss \qquad \text{…………..} \qquad \text{(B.16)}$$

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. 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 :

$$\varepsilon_{CY} = \frac{\partial VC}{\partial Y_w}\frac{Y_w}{VC} + \frac{\partial VC}{\partial Y_f}\frac{Y_f}{VC} \qquad \text{…………} \qquad \text{(B.17)}$$

The double term is necessitated by the decision to treat retail and wholesale water as multiple products – presumably because it is anticipated that although the two kinds of water are indistinguishable in the treatment plant, there may be a systematic difference in distribution costs. 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:

[117] In contrast, Saal & Parker (2005) 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)

$$\varepsilon_{CYN} = \varepsilon_{CY} + \varepsilon_{CN} \qquad \text{..........} \qquad \text{(B.18)}$$

where

$$\varepsilon_{CN} = \frac{\partial VC(Y,P,\overline{X},Z)}{\partial N}\frac{N}{VC} \qquad \text{..........} \qquad \text{(B.19)}$$

Here *N* is number of customer connections, which is a component of *Z*.

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:

$$\varepsilon_{CYS} = \varepsilon_{CY} + \varepsilon_{CS} \qquad \text{.........} \qquad \text{(B.20)}$$

where

$$\varepsilon_{CS} = \frac{\partial VC(Y,P,\overline{X},Z)}{\partial Sa}\frac{Sa}{VC} \qquad \text{..........} \qquad \text{(B.21)}$$

and here *Sa* is service area size, also a component of *Z*.

Finally, *economies of size* “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.(p. 111)” That is, if

$$\varepsilon_{Size} = \varepsilon_{CY} + \varepsilon_{CN} + \varepsilon_{CS} < 1 \qquad \text{...........} \qquad \text{(B.22)}$$

Before examining Torres & Morrison Paul’s empirical results, some comments can be made on these measures. Although water treatment and water distribution have not been analysed separately in their model, volume economies seem likely to arise mainly at the treatment stage while economies (or diseconomies) linked to customer numbers or service area are more likely to relate 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, bearing in mind that their results relate to a cross-section of US water systems, we need to ask what kinds of comparisons are meaningful and interesting. To be sure, one useful comparison is between systems which differ only in size, that is, volume, number of customers and service area vary in the same proportion so that demand density is constant: ε_{Size} will help with that kind of comparison. A second comparison might be between service areas of similar size but different demand density: here ε_{CY} (if the difference is entirely in consumption per head) or ε_{CYN} (if the difference is in numbers of similar customers) will be useful. Another interesting comparison would be between systems which differ in size of service area but have the same number of customers and consumption per customer: for that ε_{CS} would seem more relevant than ε_{CYS}, as the latter implies that consumption per customer rises as the size of the service area increases. There is also an important conceptual difference between the measures, which is evident if we think in terms of the unit cost of water supply. With ε_{CY} a value less than 1 implies economies of scale in the sense of a unit cost that falls with increase in volume; however, with ε_{CN} and ε_{CS} a value less than 0 is needed if unit cost (per gallon of water consumed) is to fall with increases in numbers of customers or size of service area – positive values imply diseconomies.

Bearing these points in mind, we can turn to Torres & Morrison Paul's results. The data consist of 255 observations from a 1996 survey conducted by the American Water Works Association (AWWA). A generalized Leontief quadratic form is specified for the cost function, which is estimated using full information maximum likelihood methods, achieving an overall R^2 of 0.96. The values obtained for scale and density economies are summarized in the **Table B.5** below.

Measure	**Sample mean (8778 Mgal)**	**Small (675 Mgal)**	**Medium (1794 Mgal)**	**Medium-large (5962 Mgal)**	**Large (29590 Mgal)**
Volume (ε_{CY})	0.58 (*)	0.33 (*)	0.46 (*)	0.53 (*)	0.61 (*)
Service area (ε_{CS})	0.16 *	0.16 *	0.17 *	0.15 *	0.30 *
Customer Nos (ε_{CN})	0.49 *	0.49 *	0.53 *	0.51 *	0.54 *
Spatial density (ε_{CYS})	0.74 (*)	0.49 (*)	0.63 (*)	0.68 (*)	0.91
Customer density (ε_{CYN})	1.07	0.82 (*)	0.99	1.04	1.15
Size (ε_{Size})	1.23 (*)	0.98	1.16	1.20 (*)	1.45 (*)

Table B5: Estimates of scale and density economies for 255 US water systems (adapted from Torres & Morrison Paul (2006, p.115)

These results are well summarized by Torres & Morrison Paul p. 116):

> "At the sample mean of the data, estimated ε_{CY} is 0.58, indicating that the average water utility is realizing increasing returns to volume. When divided into small, medium, medium-large and large utilities, the estimates indicate significant (statistically and in magnitude) increasing returns to volume for all firms that rise according to size (ε_{CY} = 0.33, 0.46, 0.53, and 0.61 respectively). These estimates suggest a flattening of the average cost curve for larger firms, broadly consistent with an L-shaped cost curve in terms of volume. … The estimated economics of vertical and horizontal expansion, or customer and spatial density, show that both counteract economies of volume by shifting the cost curve up. ε_{CN} and ε_{CS} are positive and statistically significant overall and for all size firms, but the former is much higher than the latter. The number of customers thus, *ceteris paribus*, has a stronger influence on costs than breadth of the service area over all sizes of utilities. In other words, the costs from additional customers, holding constant service area size and volume (e.g. connection costs, electricity costs to pump water, and cost associated with more complicated networks), seem relatively higher than the costs associated with marginal increases in service area size (e.g. costs associated with longer pipelines), given the number of customers and production volume. This difference in costs is, however, less prominent in the larger systems."

While this article is undoubtedly a useful contribution to the literature, bringing out more clearly than before the effect of demand density on costs, there is a possible qualification as regards 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.111, 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.

5. Other approaches

a. Clark & Stevie (1981)

There is an earlier thread in the literature in which the distribution cost function is estimated directly from a consideration of the physical lay-out of the network rather than indirectly from an assumed production function. Kim & Clark (1988, p.479) argue that such approaches suffer from severe shortcomings:

> "These studies excluded input prices completely from their cost function specification, a restriction which implicitly assumes either that input prices are identical for all water supply firms or else that water supply technologies are characterised by zero input substitution."

However, if in fact there is very little substitutability available in practice[118], then the criticism loses force and there may be merit in re-examining these earlier approaches. Indeed, the same Clark in Clark & Stevie (1981) builds on Schmalensee's approach to develop "an analytical model representing the cost of distributing water supply services in a single urban area", which explicitly includes "the relationship of transmission costs to the problem of serving spatially distributed demand." Their model includes production as well as distribution. Clark & Stevie set out their approach as follows (p. 18):

> "Physically, it is possible to separate the water supply system into two components: (1) the treatment plant, and (2) the delivery (transmission and distribution) systems … Each of these components has a different cost function. The unit costs associated with treatment facilities are usually assumed to decrease as the quantity of service provided increases. However, the delivery system is more directly affected by the characteristics of the area being served. The cost trade-offs between the two components will determine the least-cost service area … The purpose of this paper is to examine some of the trade-offs that may exist between the economies of scale for producing water and the diseconomies of transporting it to a point of use."

The key relationships of Clark & Stevie's model are:

$$C_{TOT} = C_T + C_D \quad \ldots\ldots\ldots \quad \text{(B.23)}$$

where: C_{TOT} = annual cost of water supply;

C_T = annual cost of water treatment;

C_D = annual cost of water transmission and distribution.

Normally, C_T and C_D will be a function of the volume of water produced, Q. Clark and Stevie next develop a relationship between Q and d (the radius of the service area, assumed at this stage to be circular).

$$Q = c.p.A \quad \ldots\ldots\ldots\ldots \quad \text{(B.24)}$$

Where: Q = total annual water use for the area A;

c = annual per capita water use;

p = population density.

[118] It has been suggested, only half in jest, that the only alternative to distributing water through a network of pipes would be to form a chain of men with buckets! (Joking aside, it is of course the case that in many developing countries, people must walk to fetch water, sometimes from a considerable distance.)

Population density is taken to decline exponentially with distance from the centre, so that $p = Ke^{-\lambda r}$ where K is density at the centre and λ is a measure of the rate at which density declines with distance - if λ is small, density falls off slowly; if large, density falls off rapidly; if zero, density is uniform over the service area[119]. It then follows that:

$$Q = 2\pi cK \int_0^d re^{-\lambda r} dr = \frac{2\pi cK}{\lambda^2}\left[1 - e^{-\lambda d}\left(1 + \lambda d\right)\right] \qquad \ldots\ldots\ldots\ldots \qquad (B.25)$$

In cases where the service area is not a full circle, 2π can be replaced by θ (angle in radians) to obtain a better approximation to the actual geography. Next, Clark & Stevie adopt a simple form for the relationship between water acquisition/treatment costs and volume, reflecting their empirical evidence on scale economies in water treatment:

$$C_T = A.Q^{\alpha} \qquad (\alpha < 1) \qquad \ldots\ldots\ldots\ldots\ldots \qquad (B.26)$$

Whence, using (B.23): $C_T = A.\left(\frac{\theta cK}{\lambda^2}\left[1 - e^{-\lambda d}\left(1 + \lambda d\right)\right]\right)^{\alpha}$ …… (B.27)

For the transmission/distribution component of costs, Clark & Stevie propose (based on an analysis of data from the pipe network of one water utility) an expression for the cost of supplying a quantity Q_p to a point in the service area of the form:

$$C = B.(\, Q_p)^{\beta} \qquad \ldots\ldots\ldots\ldots\ldots \qquad (B.28)$$

To obtain the total cost of distribution, this expression is then integrated over the whole service area[120], yielding:

$$C_D = \frac{B\theta\left(cK\right)^{\beta}}{\beta^2\lambda^2}\left[1 - e^{-\beta\lambda d}\left(1 + \beta\lambda d\right)\right] \qquad \ldots\ldots\ldots\ldots\ldots \qquad (B.29)$$

Clark & Stevie now have all the elements in place to compute C_{TOT} by adding (B.27) and (B.29), and to obtain an expression for unit cost by dividing the resulting sum by Q from (B.24). The resulting expressions are quite complicated, and not easy to interpret, and it does not seem necessary to set them out here. However, the simulations run by Clark & Stevie using their results are of considerable interest.

Reproduced below is their table showing calculations of cost/distance relationships using the expressions for treatment and distribution unit costs derived above, and parameters based on data from one US water utility.

[119] This formulation is not unusual – see, for example, DiPasquale & Wheaton (1996), p.63 “The variation in population density with distance from the central city is often summarised through the estimation of a population density gradient … The standard specification is the negative exponential …” They estimate values of –0.09 and –0.11 for the Boston metropolitan area in 1990 and 1970 respectively.

[120] This step, although mathematically correct, is open to question as not satisfactorily aggregating the relevant distribution costs because it implies, as Clark & Stevie themselves note, that “all transmission of services is strictly outward from the center and that lateral flows are never necessary”.

Q m gal/yr	d miles	Treatment unit cost ($/m gal)	Distribution unit cost ($/m gal)	Total unit cost ($/m gal)	Marginal Cost ($/m gal)
8808.47	2	63.50	125.45	188.95	188.95
30172.25	4	47.84	128.91	176.75	171.72
58332.44	6	41.11	132.31	173.42	169.80
89416.04	8	37.26	135.64	172.90	171.94
120893.44	10	34.77	138.89	173.66	175.80
151181.43	12	33.02	142.05	175.07	180.71
179354.48	14	31.75	145.10	176.85	186.30
204936.20	16	30.79	148.03	178.82	192.64
227749.76	18	30.05	150.83	180.88	199.42
247811.79	20	29.48	153.50	182.98	206.88

Assumptions**:** per capita water use, c = 0.054750 m gal/person; coefficient of dispersion, λ = 0.12; population density at centre, K = 15,000 persons/sq mile.The process generating this table starts with the settlement radius, d. Given the central density, K, and the coefficient of dispersion, λ, this implies a certain population. Multiplying population by per capita water use, c, then gives total water use, Q. Calculation of treatment and distribution unit costs then follows.

Table B.6: Water treatment and distribution costs as modelled by Clark & Stevie (1981, p. 28)

As can be seen, treatment unit cost is decreasing – not surprisingly given the functional form of (B.24). Distribution unit cost however is increasing, something which is not obvious from (B.27) – it turns on the sign of $\left[1 - e^{-\beta\lambda d}(1 + \beta\lambda d)\right]$. The effect is due to the increasing distances over which water is being distributed as settlement size increases and the greater dispersion (lower density) of the further out (suburban) population. In the example, the distribution cost effect dominates beyond d = 6 miles, leading to overall increasing unit costs beyond this size of settlement. Clark & Stevie comment (p.25) “A problem that plagues most utility managers is determining the most efficient size for a utility service area … Water utility managers tend to assume unlimited economies of scale for water treatment and delivery systems. This often leads to water utilities that have service areas larger than the most efficient size.”

They go on to use their relationships to analyse a case where it is preferable for an outlying community to provide its own works rather than connect to an existing system. This example, while demonstrating the possibility of such an effect, falls short of establishing it as a general result, despite the use of data from a real utility. This is because it rests on a number of rather specific assumptions, which may not be generally valid. In addition, although Clark & Stevie refer to “total costs”, it is not clear how far capital costs have been taken into account (certainly no distinction is made between capital and operating costs). There is also the question whether the distribution cost function has been correctly specified (see **Footnote 120**). A further question relates to the treatment of leakage: by taking the quantity of water produced to be the same as the quantity used (*Q* in both cases), there is an implicit assumption that distribution losses are zero. In practice leakage is often large – typically 20% or more of water put into distribution – and therefore a significant factor which ought if possible to be allowed for in the distribution cost function.

Notwithstanding these reservations, there is much to be said for Clark & Stevie's approach, particularly in analysing distribution costs separately from acquisition/treatment costs. Nor does the failure (as compared with Kim & Clark (1988) or Garcia & Thomas (2001), for example) to allow for input substitutability appear a serious drawback, given the very limited opportunities for making different technology choices in most urban water supply situations. In fact, it is arguable that the bringing into play of flexible functional forms may be using a sledgehammer to crack a nut. At the least it would seem worth checking whether simpler approaches would not adequately capture the essentials of the situation.

b. Ofwat's econometric models

Interestingly, in its work to compare the efficiency of water companies, Ofwat (2004, Appendix, pp.44-52) does not use a single comprehensive model. Instead, it subdivides water and sewerage functions into a number of components and then derives relationships between each element of expenditure and various explanatory factors. In the case of water supply, this process results in 8 econometric models:

> Operating expenditure – water distribution
> Operating expenditure – water resources and treatment
> Operating expenditure – water power
> Operating expenditure – business activities
> Capital maintenance – water distribution infrastructure
> Capital maintenance – water distribution non-infrastructure
> Capital maintenance – water management and general
> Capital maintenance – water resources and treatment.

The stages of the process used to derive these models is explained as follows:

Step 1: Expert review of potential drivers

Step 2: Data collection and validation

Step 3: Identification of atypical expenditure and exceptional items[121]

Step 4: Produce revised data for statistical analysis

Step 5: Generate plausible conceptual models to limit statistical analysis

Step 6: Statistical analysis to generate robust relationships between expenditure and explanatory factors

Step 7: Expert review of the statistical models.

In practice, steps 5-7 may go through a number of iterations. There then follow 4 more steps leading from preliminary assessment of relative efficiency through further review of special factors specific to individual companies to final judgements on relative efficiency.

The resulting models have a distinctly ad hoc feel and the theoretical basis for the specifications chosen is unclear. However, given the difficulty in finding good statistical relationships at company level based on theoretical considerations, one can have some sympathy with the approach taken by Ofwat. It is also relevant that Ofwat's objective is simply to make efficiency comparisons, not to establish propositions about the economics of urban water supply.

Against this background, the Ofwat models dealing with water distribution and water resources and treatment are set out below, together with some comments on the specifications adopted.

i. Operating expenditure – water distribution model

Modelled cost	Ln(distribution functional expenditure *less* power / resident population)	$R^2 = 0.261$
	Constant	-5.203 (S.E = 0.160)
Explanatory variable	Length of main > 300mm / length of main	5.165 (S.E = 1.943)

In this model, distribution expenditure per head is modelled as a function of the proportion of mains over 300mm in diameter. Power expenditure is excluded as it is modelled separately – see below. The rationale for this model is not easy to discern: the number of properties served does not feature, the reason for a log specification of

[121] In 2005/06, an additional step was added here: "Consider company specific special factors."

expenditure per head is not clear and the choice of explanatory variable is puzzling. The explanatory power of the model is only modest.

ii. Operating expenditure – water resources and treatment model

Modelled cost	Resources and treatment functional expenditure *less* power *less* Environment Agency charges / resident population	R^2 = 0.274
	Constant	1.485 (S.E = 1.927)
Explanatory variables	Number of sources / distribution input	16.770 (S.E = 6.268)
	Proportion of supplies from rivers[122]	5.124 (S.E = 2.449)

In this model resources and treatment expenditure per resident person is modelled as a function of the inverse of average supply per source and the proportion of supplies from rivers. Power costs and EA charges are excluded. One difficulty in this area is that the expenditure figures include both water acquisition costs and treatment costs, making it difficult to disentangle their separate effects. However the positive coefficient on number of sources / distribution input can perhaps be interpreted as indicating economies of scale in treatment works, as more sources implies more treatment works and smaller scale operations on average; the coefficient on proportion of supplies from rivers then picks up the higher costs of treating river waters. It is not clear why in modelling resource and treatment expenditure Ofwat has not made use of the data in the June returns on numbers and size of treatment works and on type of treatment, which, on the face of it, should give a more direct relationship between expenditure and output. The explanatory power of this relationship is again modest.

iii. Operating expenditure – water power model

Modelled cost	Ln(power expenditure)	R^2 = 0.989
	Constant	-9.081 (S.E = 0.245)
Explanatory variable	Ln(Distribution input x average pumping head)	0.940 (S.E = 0.023)

The reason for modelling power expenditure separately is evident in the good fit for this relationship. It appears that power expenditure is pretty well fully explained by pumping

[122] Changed to proportion of supplies from boreholes in 2005/06 report.

costs, with a 1% increase in (distribution input x average pumping head) resulting in an increase of costs of a little under 1%. Pumping costs here include pumping from source to treatment works and distribution pumping. Although some power costs are incurred within treatment works, these too are related to the volumes being treated. One question is why the relationship should be log-linear rather than linear. It would be of some interest to try to check whether the distance over which water is pumped has any effect in addition to the pumping head effect.

iv. Capital maintenance expenditure – water distribution infrastructure model

Modelled cost	Ln(annual average water distribution infrastructure expenditure / length of main)	R^2 = 0.496
	Constant	-4.802 (S.E = 0.542)
Explanatory variable	Ln(Total number of connected properties per length of main / total length of main)	0.888 (S.E = 0.200)

While this relationship performs moderately well, there are again some puzzling features. Number of connected properties divided by length of main provides a measure of property density and the relationship can be interpreted as evidence of density economies. But the explanatory variable is twice divided by length of main so complicating interpretation. If one started with the view that water infrastructure capital maintenance costs are likely to be affected (positively) by total length of main and the number of connected properties and (negatively) by property density, the model would be specified differently to throw light on the separate effect of each variable.

v. Capital maintenance expenditure – water distribution non-infrastructure model

Modelled cost	Ln(annual average water distribution non-infrastructure expenditure / pumping station capacity)	R^2 = 0.338
	Constant	-6.433 (S.E = 0.533)
Explanatory variable	Ln(Water service reservoir and water tower storage capacity / pumping station capacity)	0.664 (S.E = 0.207)

This relationship makes use of data from the June Returns which is not publicly available. It is interesting in suggesting considerable economies of scale in this aspect of water supply, such that a 1% increase in reservoir/storage capacity leads to an increase of only 0.664% in the relevant capital maintenance costs.

vi. Capital maintenance expenditure – water resources and treatment model

For this aspect of Ofwat's efficiency comparisons, a unit cost approach has been adopted. Each company's average annual expenditure on water resources and treatment capital maintenance is divided by total connected properties and then compared with the weighted average industry cost of 8.471. The use of connected properties rather than treatment capacity as the divisor seems surprising and comparison with the industry average would appear to make no allowance for differences in costs attributable to differences in the number and size of communities served by a company – a large number of small communities being presumably more expensive to service than a small number of large ones.

c. Duncombe & Yinger (1993)

In an original contribution to the literature on public production, Duncombe & Yinger develop a new analysis of returns to scale using a two-stage procedure. They consider that the notion of scale in public production has three fundamental dimensions: the quality of the services provided, the level of activity by the government agency and the number of people served. With multiple products, a fourth dimension, economies of scope, must also be considered. Building on the conceptual framework developed by Bradford *et al* (1969), they make a distinction between the direct services provided by a government and the outcome of interest to voters. "In the case of fire protection, for example, voters care about the saving of lives and property, not about the number of fire companies available, per se."

Duncombe & Yinger therefore divide the public production process into two stages. "The first stage of the process is similar to production for a private firm. Local governments produce an intermediate output, *G*, with a standard production function:

$$G = f(L, K, Z) \qquad \ldots\ldots\ldots\ldots\ldots \qquad \text{(B.30)}$$

Where *L* is labor, *K* is capital equipment or facilities, and *Z* is other factor inputs. Assuming cost minimisation, the associated first-stage cost function is:

$$TC = c(G, W) \qquad \ldots\ldots\ldots\ldots\ldots \qquad \text{(B.31)}$$

Where *W* represents a vector of factor prices.” Note that the assumption of cost minimisation here implies the separability of this stage of production from the next stage.

The government activities, *G,* are then viewed as “an intermediate output in the production of the final output or service outcome of interest to voters, *S*. The distinction between *G* and *S* is important because exogenous ‘environmental’ factors influence the transformation of *G* into *S*. Two communities of the same size may utilise the same technology and level of resources for fire protection, for example, but experience significant differences in property losses and casualties owing to differences in the harshness of their fire-fighting environment.”

Following Bradford *et al* (1969), the second stage of the production process can be represented as

$$S = h(G, N, E) \qquad \ldots\ldots\ldots\ldots\ldots \qquad \text{(B.32)}$$

Where *N* is the jurisdiction’s population and *E* represents a vector of environmental cost factors. This equation indicates the level of government activity, *G,* required to produce a given level of public services, *S,* taking into account the impact of population and the environment. Duncombe & Yinger explain that population is included in the final output function here to allow for the possibility of “congestion” in the provision of public services and that it is found to play a similar role to other environmental factors.

It may be noted at this point that this formulation of the second stage of production does not allow for possible additional inputs of *L, K* or *Z* at this stage. The inclusion of population as a quasi-environmental factor derives from Duncombe & Yinger’s view that the product under consideration is basically a (Samuelsonian) pure public good but that beyond some point additional population in the service area may impinge on the availability or quality of the service to existing residents, and it is in this rather specialised sense that they use the term congestion. The approach is therefore not suitable to be adopted without modification to, say, the provision of utility services although the idea of dividing supply into two stages, e.g. water treatment and water distribution, each with its own distinctive production function is attractive.

Duncombe & Yinger go on to propose a cost function which combines the two stages of production:

$$TC = c[h^{-1}(S, N, E), W, E'] \qquad \ldots\ldots\ldots\ldots\ldots\ldots \qquad (B.33)$$

Where *E'* is a sub-set of environmental variables that affect factor substitution. Based on this relationship, they go on to derive their three measures of returns to scale in public production.

Service quality is measured by the final government output, *S*, which represents the service effectiveness of interest to voters. Average cost, i.e. cost per unit of quality is measured by *TC/S*. Duncombe & Yinger then define *returns to quality scale* as the change in *TC/S* which results from a change in *S* holding population, *N*, and environmental factors, *E*, constant. This derivative is:

$$\frac{\partial(TC/S)}{\partial S} = \frac{(\partial TC/\partial G)(\partial G/\partial S) - (TC/S)}{S} = \frac{MC_S - AC_S}{S} \quad \ldots\ldots.. \quad (B.34)$$

where MC_S and AC_S are the marginal and average costs of producing *S*. Thus, increasing returns to quality scale exist if $MC < AC$, i.e. if the average cost curve is downward sloping. This can also be expressed in elasticity form:

$$\frac{\partial(TC/S)}{\partial S}\frac{S}{TC/S} = \left(\frac{\partial TC}{\partial G}\frac{G}{TC}\right)\left(\frac{\partial G}{\partial S}\frac{S}{G}\right) - 1 = \theta_1\theta_2 - 1 \qquad \ldots.. \qquad (B.35)$$

Economies to quality scale thus exist if the product of θ_1 and θ_2 is less then unity.

Duncombe & Yinger comment (p. 53) that “The first of these elasticities, θ_1, which we call the ‘first stage’ or ‘technical returns to scale’, is the notion most closely associated with the definition of returns to scale in private production. It represents the technical relationship between inputs and the intermediate output of government, *G*. This interpretation is possible because of the duality between production and cost functions: θ_1 equals the inverse of the elasticity of *G* with respect to the scale of inputs. Increasing (decreasing) technical returns to scale imply that θ_1 is less than (greater than) one. The second elasticity, θ_2, measures what we call the ‘second stage returns to scale’, i.e. the relationship between the intermediate and the final output of government. This effect captures the influence of the production environment on the translation of *G* into *S* and is likely to vary with community characteristics. Communities with a harsh environment require more *G* to obtain a given *S*, i.e. they have a higher θ_2, than do communities with a favourable environment. They are less likely therefore to face increasing returns to quality scale, even if all communities have the same technical returns to scale.”

In a similar way, Duncombe & Yinger define *returns to population scale* as the derivative of *TC/N* with respect to *N*, controlling for *S* and *E*:

$$\frac{\partial(TC/N)}{\partial N}=\frac{(\partial TC/\partial G)(\partial G/\partial N)-(TC/N)}{S}=\frac{MC_N-AC_N}{N} \quad \ldots \text{ (B.36)}$$

Economies (diseconomies) to population scale exist if the per capita cost curve is downward (upward) sloping. This result can also be expressed in elasticity form:

$$\frac{\partial(TC/N)}{\partial N}\frac{N}{TC/N}=\left(\frac{\partial TC}{\partial G}\frac{G}{TC}\right)\left(\frac{\partial G}{\partial N}\frac{N}{G}\right)-1=\theta_1\theta_3-1 \qquad \ldots\ldots \quad \text{(B.37)}$$

Economies (diseconomies) to population scale exist if this expression is less than (greater than) zero. As Duncombe & Yinger explain (p. 54-55): “As with returns to quality scale, the first elasticity, θ_1, is technical returns to scale. In this case, however, the other elasticity, θ_3, captures the relationship between government activity and population; that is, it measures congestion in the provision of public services.” They go on to say: “ … congestion has been introduced into local expenditure research through a ‘congestion function’ of the form $S = (G)(N^{-g})$, where g = 0 for a pure public good and g = 1 for a private good. While the congestion parameter, g, may differ between public services, it has been assumed to be the same for all communities. In fact, however, the impact of another person on the amount of *G* needed to maintain a given level of *S* may depend on *E*. In fire protection, for example, the impact of another person on the level of fire protection activity required to maintain a certain standard of service quality may depend on the existing condition of buildings in the community. Because fires may spread from one unit to another, the cost of assuring a certain quality of fire service for a new household is likely to be higher in communities with poor building condition than in communities with good building condition. Thus, ‘publicness' itself may depend on the environment, and we model the relationship between *G* and *N* to reflect this possibility.” In a further comment, Duncombe & Yinger go on to observe that “ … two public services may face the same technical returns to scale but have different returns to population scale because of differences in congestion” and they proceed to make a contrast between police protection (a public good) and garbage collection (generally regarded as a private good). Although both may exhibit constant technical returns to scale, the former is more likely to show increasing returns to population scale because of a lower θ_3.

While Duncombe & Yinger’s distinction between the three different dimensions of returns to scale in public production is undoubtedly illuminating, this particular part of their discussion strikes one as rather laboured; and their focus on ‘congestion’ (as they define it) is potentially misleading. What it does not seem to address is some much more obvious and possibly rather important differences in communities’ environments which are likely to affect the amount of *G* needed to maintain a given level of *S*. For example, what about population density, which could either favour increasing returns to population scale (because it is easier to service large numbers of people if they are close together) or the opposite (if high density leads to congested transport infrastructure making it more difficult for police or fire services to get to incidents)? Or what about geography? A community situated in a mountainous or otherwise fragmented area will be more costly to service than one on a flat plain, other things equal. In short, Duncombe & Yinger do not appear to recognise that distribution or access costs may play a part in what they describe as ‘congestion’. In consequence, their findings on returns to population scale have probably missed an important aspect of the problem. This is somewhat surprising, as at a couple of points in their article their findings might have alerted them to the issue: a footnote on p.55 says “The results of Craig (1987) show low congestion in the case of police services, but the results of Ladd and Yinger (1989), which apply to very large cities, suggest severe congestion for police services, and hence diseconomies to population scale”; and on p.66, commenting on their fire service results, they say: “The coefficient on population density [a variable introduced as a proxy for the risk that fires will spread] is negative, suggesting that reductions in fire response time with greater population density outweigh the increased potential for fires spreading between units.”

Duncombe & Yinger go on to estimate their model using data on 188 fire departments in New York State for the years 1984-86. To obtain an estimating equation, they need to specify both the first-stage cost function (B.31) and the second-stage production function (B.32). They assume that second-stage returns to scale can be modelled as $\theta_2 = (1 + \lambda^* E^*)$. Thus the final output function is of the form:

$$S = G^{-(1+\lambda^* E^*)} N^{-g} E^{-\nu} \qquad \text{……………..} \qquad \text{(B.38)}$$

or

$$G = S^{(1+\lambda^* E^*)} N^{g(1+\lambda^* E^*)} E^{\nu(1+\lambda^* E^*)} \qquad \text{……………..} \qquad \text{(B.39)}$$

To place as few restrictions on production technology as possible, they employ a translog cost function. They assume that the cost of *G* can be described by a translog cost function for two factors of production, labour and capital equipment (including facilities). The standard translog cost function is modified by substituting (B.39) for *G*. This form makes it possible to estimate all the three parameters of returns to scale of interest, θ_1, θ_2, and θ_3. In addition Duncombe & Yinger investigate whether there are economies of scope between the two primary activities of fire departments, fire suppression (reducing fire damage once a fire starts) and fire prevention (preventing a fire from starting), using a multi-product translog cost function of the same general form.

Duncombe & Yinger employ a number of ingenious devices to assemble suitable data and to control for possible bias. Their findings as regards returns to scale are summarised in **Table B.7** below:

	Single-product Cost model	**Multi-product Cost model**
Economies of quality scale		
Technical economies of scale $(\partial \ln TC / \partial \ln G = \theta_1)$	0.28	0.73
Second-stage returns to scale $(\partial \ln G / \partial \ln S = \theta_2)$	1.11	1.08
Economies of quality scale $(\partial \ln (TC/S) / \partial \ln S = \theta_1\theta_2 - 1)$	-0.69	-0.22
Economies of population scale		
Technical economies of scale $(\partial \ln TC / \partial \ln G = \theta_1)$	0.28	0.73
Congestion elasticity $(\partial \ln G / \partial \ln N = \theta_3)$	3.84	1.51
Economies of population scale $(\partial \ln (TC/N) / \partial \ln N = \theta_1\theta_3 - 1)$	0.06	0.10
Economies of scope	N/A	-0.13

Table B.7: Duncombe & Yinger's estimates of returns to scale in fire protection in New York State (Duncombe & Yinger (1993, p.68)

Although Duncombe & Yinger acknowledge that most of their coefficients are not statistically significant, they generally have the expected sign and are mostly of a plausible magnitude. Overall, the effort involved in trying to disentangle the different dimensions of returns to scale appears to be vindicated.

d. Public facilities location

Here we encounter a rather different approach to the kind of problem addressed in this research. The intellectual foundations of this approach are found in operations research, more specifically in the use of linear programming to solve location problems.

One early contribution is Bos (1965) in his monograph on the spatial distribution of economic activity. Interestingly, his analysis foreshadows the kind of new economic geography developed in the wake of Krugman (1991). Bos summarises his findings as follows (p.89):

> "The analysis has been based on three assumptions:
>
> 1. Agricultural production and population are spread over a given area;
> 2. The production of non-agricultural industries is characterised by indivisibilities leading to economies of scale;
> 3. Transport of goods and services gives rise to transportation costs.
>
> These three elements are sufficient to explain that non-agricultural production is concentrated in production units of various sizes and that the number of production units is not the same for all industries."

Bos also extended the analysis to try to investigate a hypothesis of Tinbergen concerning the hierarchy of production centres, finding that both the type of industry and transport costs could affect this hierarchy. However, Bos found the technical demands of the analysis very challenging: (pp.91-2) "The problems which have been studied have, in principle, all been very simple and have omitted various features of reality. Even these highly simplified problems have been shown to have no single solutions and to require very complex methods of analysis." (Bos made use of a mixed integer linear programming model but was only able to run illustrative numerical examples as "no method of solution for determining an optimum dispersion was available.")

Perhaps Bos was too ambitious in the scope of the problem he tried to solve. More fruitful has been the use of linear programming to study more limited problems such as the location of a facility in a partial equilibrium framework. A useful survey can be found in Thisse & Zoller (1983) and a textbook treatment of the subject is provided by Love *et al* (1988).

Thisse & Zoller note in their introduction how the difficulty of identifying which services are truly public has led to a shift of attention from activities to facilities, i.e. from services to infrastructure. They comment (p. 1) " … the few pieces of theory which have been devoted to the locational analysis of public services deal directly with facilities. In so doing, the fact has been explicitly recognised that the public outputs are not everywhere equally available in most real cases. In other words, *space introduces some types of exclusion* ("impurities") *in public goods* and, therefore, a public-space theory is needed." Such a theory, they suggest, would need to bring together elements of hitherto somewhat separate fields: public goods and product differentiation theory from economics; central place theory from geography; and locational decision analysis from operations research. The key points that Thisse & Zoller take from these disparate fields are:

a. User benefits are either location dependent or distance dependent (and therefore the services concerned are not pure public goods);

b. Product differentiation, combined with differences in tastes or incomes, results in a partitioning of consumers by the facilities they patronise, with economies of scale in production implying a finite system of facilities;

c. Central place theory recognises the fundamental trade-off between increasing returns to scale and transportation costs: "stated differently, a decrease in the number of facilities provides a saving in the installation costs, but leads to an increase in the travel costs"(p.5);

d. Operations research type location models provide a method of solving quite complex problems, involving (for example) non-linear transport costs, fixed and variable production costs, alternative price policies, and a variable number of plants. However, the profit-maximisation or cost-minimisation objective commonly used in these models implies a rather simplistic (utilitarian) social utility function.

We focus here on point (d), which is examined in more detail by Hansen *et al* (1983) and, later, by Love *et al* (1988). Hansen *et al* remark (p.223): "We are concerned exclusively with *public services* (police and fire protection, postal service, emergency medical care, social services, education, recreation services, parks, libraries, wastewater treatment, solid waste disposal, etc) … We focus on services which are made available at some *facilities* … Two categories of services are distinguished: *fixed* services – that is to say, services consumed at the facilities where they are supplied – and *delivered* services – which are used at the places where they are demanded." They go on to note that in modelling these situations, both the objectives (e.g. minimisation of access or delivery cost, or minimisation of the combined cost of installation and travel) and the constraints (e.g. on the location of facilities, the capacity of facilities, the number of facilities, the extent to which demand is satisfied, available budget, etc) can vary considerably from case to case. In a single facility location problem, demand is located at a number of fixed points and the problem is to find the best location from which to service these requirements. More commonly with public services, the location of several facilities and the assignment of users to those facilities needs to be determined simultaneously.

We turn to Love *et al* (1988) to get a feel for this more complex problem: "When the locations of several new facilities are to be determined simultaneously with the allocation of flow between each new facility and the existing facilities, the problem is referred to as a *location-allocation* problem." Possible sites for the new facilities may be fixed points (the finite model) or any point within a defined space (the continuous model). In an interesting illustration (from the point of view of this research), Love *et al* (p.3) set out the elements of a large farm water supply problem as follows:

> "The existing facilities are points of end use for the water, such as livestock barns, irrigation systems, or houses. The new facilities are the deep wells to be drilled. If a new system is to be designed, the relevant questions are: How many wells should there be? Where should they be located? Which subset of users should each well serve? An extreme design is to locate a well at each user location. In this case, piping costs are minimised but the drilling cost may be prohibitive. Another configuration is to have one large well. A single well minimises drilling costs but piping and pumping costs may be prohibitive. Using one well would entail solving a single facility location problem. When two wells are considered, drilling costs are increased, but piping and pumping costs are reduced. The allocation question is thus introduced. Where should the two wells be located and to which set of users should each one be connected? If the two well problem can be solved, then a three well problem can be considered, and so on, until the most economical number of wells has been found."

Other problems that can be addressed using similar methods include:

- The location of emergency service facilities such as ambulance bases or fire stations, where it may be desirable to minimise the maximum distance from the new facility to any of the points served;

- The location of abnoxious facilities (such as garbage dumps or sewage works), so as to minimise nuisance to existing inhabitants;

- The location and number of radio or TV transmitters to ensure adequate coverage over a defined area.

In general, the solution to such problems involves the minimisation of a weighted distance function involving existing and new facilities (where the weights may be, for example, costs per km), subject to constraints deriving from the nature of the problem. Love *et al* comment (p.144) that set up in this way, the problem "has a non-linear objective function that is neither concave nor convex, and generally contains many *local* minima. This means standard non-linear programming algorithms may fail to produce a global minimiser." Moreover, they add, not all location-allocation problems can be adequately represented in this way (p.144): "Among the prominent factors that may impair the use of the model are the following: The requirements … may depend on the new facility locations. Transport costs may not be adequately expressed as weights times distances. The total cost may involve other significant components besides transportation costs. There may be flows between the new facilities. Finally, it may be more appropriate to maximise profit." Nevertheless, there are problems for which the location-allocation model is applicable, including the large farm water supply problem outlined above.

How useful, in the context of this research, might location decision analysis be? Unfortunately, less than might at first sight appear to be the case. Although the method is able to throw light on situations where there is a trade-off between production costs and distribution costs (as in the large farm water supply example), and the issue of multiple local minima has affinities with the question of non-separability between water supply and water distribution (see **Section 6(c)** below), it is at heart a highly specific, *ex ante* appraisal tool, with rather demanding information requirements, and it would not be appropriate to assume that the observed organisation of water treatment works and distribution networks was the outcome of location decisions reached using this method.

6. Lessons from the literature surveyed

All the literature surveyed can be seen as wrestling in one way or another with the implications of Schmalensee's (1978) observation that: "When services are delivered to customers located at many points, cost must in general depend on the entire distribution of demands over space." This lies at the heart of the economics of distribution, and is what distinguishes it from the economics of production.

a. Is there a trade-off between production and distribution?

In several of the references reviewed here, the possibility of a trade-off between production and distribution is mentioned – e.g. Nerlove (1963), Clark & Stevie (1981), Thisse & Zoller (1983), Kim & Clark (1988), and Torres & Morrison Paul (2006). However, only Clark & Stevie attempt to investigate this trade-off in a systematic way and their approach is open to criticism as too *ad hoc*. It seems likely that in general there is a trade-off but there is plenty of scope for it to be further explored.

b. Measuring scale economies in distribution

Duncombe & Yinger (1993) have pointed out that the notion of scale in public production has more than one dimension. In their study of fire protection services, they identify three fundamental aspects: the quality of the services provided, the level of activity by the government agency and the number of people served. With multiple products, they observe, a fourth dimension, economies of scope, must also be considered. It would be possible to adapt these ideas to apply to water distribution as follows:

- *Quality of service*: In water distribution this includes reliability, adequate pressure, etc as well as minimizing deterioration of water quality in the distribution system. In the UK all companies meet substantially the same (high) standards so that differences in standards are not an important factor in cost analysis[123]. However, it remains the case that the cost of achieving these standards may vary from company to company because of environmental factors, such as soil conditions, softness or hardness of water supplies and hilliness of the terrain.
- *Level of activity*: This can be taken to be the volume of water put into distribution, with the economies of scale in water treatment investigated in **Chapter IV** being equivalent to the 'first stage' or 'technical returns to scale' identified by Duncombe & Yinger.
- *Number of people served*: Here we see a need to extend Duncombe & Yinger's framework to recognize that the size of the area served and the distribution of properties within it as well as the number of people in the area affect distribution

[123] However, Stone & Webster Consultants' (2004) findings (p.24) "suggest that improvements in output quality, as well as the significant costs that have been borne in order to bring about these improvements, must be accounted for to properly assess economies of scale and scope in the water industry".

costs. The costs of serving a dense population will be different from the costs of serving the same population spread less densely over a larger area.

- *Economies of scope*: Although some authors have portrayed water supply as a multi-product activity, by distinguishing between residential and non-residential supply (Kim & Clark (1988)), or between water delivered to customers and water lost through leakage (Garcia & Thomas (2001)), we see this as an unnecessary complication in the present context. (On the other hand, treating the water supply and sewerage activities of companies that do both as distinct products would seem entirely justified.)

A better starting point however is provided by Roberts' (1986) analysis of scale economies in electricity production and delivery (see **section 3(b)** above). Roberts proposes a cost function for a firm's total cost of supplying electricity in the form:

$$C(P_I, P_{KD}, P_{MD}, Q, A, N) \qquad \ldots\ldots\ldots\ldots \qquad \text{(B.40)}$$

Where P_I is the price of input electricity, P_{KD} is the price of distribution capital and P_{MD} is the price of distribution materials, Q is the quantity of electricity supplied, A is service area and N the number of customers. Among the various advantages Roberts reasonably claims for his cost model are that it enables three distinct measures of economies of scale to be identified, viz:

1. Economies of output density: $R_Q = \dfrac{1}{\varepsilon_Q}$, where $\varepsilon_Q = \dfrac{\partial \ln C}{\partial \ln Q}$, applicable when there is an increased demand for power from a fixed number of customers in a fixed service area;

2. Economies of customer density: $R_{CD} = \dfrac{1}{\varepsilon_Q + \varepsilon_N}$, where $\varepsilon_N = \dfrac{\partial \ln C}{\partial \ln N}$, applicable when more power is delivered to a fixed service area as it becomes more densely populated, while output per customer remains fixed;

3. Economies of size: $R_S = \dfrac{1}{\varepsilon_Q + \varepsilon_N + \varepsilon_A}$, where $\varepsilon_A = \dfrac{\partial \ln C}{\partial \ln A}$, applicable when the size of the service area increases while holding customer density and output per customer constant.

Note, however, that Roberts' cost function incorporates both the production and distribution of electricity and assumes constant returns to scale in electricity production, which seems questionable.

Roberts' approach is further developed by Torres & Morrison Paul (2006) in their treatment of output density in US water supply (see **section 4(e)** above). 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.

The starting point is a short run cost function:

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

Where *Y* is a vector of outputs (wholesale water, Y_w, and retail water, Y_r, are distinguished), *P* is a vector of variable input prices (e.g. labour, electricity, purchased water), $\overline{X}$ is a vector of quasi-fixed inputs (e.g. storage and treatment capacity – this is what makes the approach short run) and *Z* is a vector of technical and environmental characteristics.

1. *Economies of volume scale* are then defined as:

$$\varepsilon_{CY} = \frac{\partial VC}{\partial Y_w}\frac{Y_w}{VC} + \frac{\partial VC}{\partial Y_r}\frac{Y_r}{VC} \qquad \ldots\ldots\ldots \qquad \text{(B.41)}$$

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, which need not concern us here.

2. *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

$$\varepsilon_{CN} = \frac{\partial VC(Y, P, \overline{X}, Z)}{\partial N}\frac{N}{VC} \qquad \ldots\ldots\ldots \qquad \text{(B.42)}$$

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

3. *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

$$\varepsilon_{CS} = \frac{\partial VC(Y, P, \overline{X}, Z)}{\partial Sa}\frac{Sa}{VC} \qquad \text{………..} \qquad \text{(B.43)}$$

and here *Sa* is service area, also a component of *Z*. This is not a measure used by Roberts – and indeed one might ask in what circumstances volume would increase with area if the number of customers has not increased. Finally,

4. *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.

As with Roberts' measures, those used by Torres & Morrison Paul incorporate the effects of both the production stage and the distribution stage of water supply but they do not assume constant returns to scale in water production.

Whatever the precise measures used, it is clear that it is important to bring out in any analysis of distribution the different cost effects of volume expansion, increase in number of connections and increase in service area; and in any discussion about scale effects, to be clear about which dimension, or dimensions, are under consideration.

c. Separating distribution from production using production/cost functions

A key issue in the economic analysis of water supply is how best to bring out the distinctive features of water distribution. Among those using production and/or cost functions, two broad approaches can be identified in the literature:

(a) Model water supply as a single activity but include variables intended to pick up distribution effects, such as miles of pipes (Kim & Clark (1988)), number of connections (Stone & Webster (2004)), or service area (Torres & Morrison Paul (2006)). It would also be possible to use some composite of these, such as connections/mile of pipe or connections/service area, i.e. measures of density, although this is not done directly in the studies mentioned. The main problem with this approach is that it may fail to expose fully the distinctive economics of the distribution stage.

(b) Develop a two stage model of production and supply, either based on network costs (Clark & Stevie (1981)) or on a two stage production function – e.g. Roberts (1986) and Thompson (1997) for electricity supply, Duncombe & Yinger (1993) for fire protection, with distribution effects being directly identified in the second stage. The main problem here is how to deal with the situation if the two stages are not separable (in the formal economic sense)[124].

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

d. Specifying functional form

Having selected an approach, the question of specification arises. Whereas in early empirical work on industrial production, the starting point might have been a specification of the production function (commonly the Cobb-Douglas) from which a cost function would then be derived[125], recent work has tended to specify the production function only in a very general form, proceeding then to a flexible form (e.g. translog) specification of the cost function. The duality between the cost function and the production function still allows the parameters of interest to be estimated, while the flexible form specification avoids unnecessarily restrictive assumptions about the form of the production function (such as the implied restriction of the Cobb-Douglas that all elasticities of substitution are equal to 1), leaving the data (as it were) to speak for itself. The arguments for using a flexible form cost function are indeed attractive. There are, however, some counter-arguments, some general to any application, some specific to application to water supply.

As regards the general limitations of flexible form functions, Chambers (1988) draws attention, *inter alia*, to two (see pp.174-179): (a) "Perhaps more serious than the above is the fact that generalized quadratic forms (e.g. the generalised Leontief, the translog, and the quadratic mean of order *p*) are very inflexible in representing separable technologies." (b) "Even if flexible forms are not restrictive, their ability to approximate arbitrary technologies is limited. The notions of approximation relied upon are local in

[124] See **section 3(d)** for a fuller discussion.

[125] See for example Nerlove's study of economies of scale in the US electric power industry, as described by Greene (2003), pp.124-127.

nature: either a point approximation to the function, gradient and Hessian or a second-order Taylor series expansion. Neither are truly global, and approximations based on them cannot be exact for a wide range of observations.” Chambers concludes:

> “The best way of interpreting these caveats and limitations is that the main attraction of flexible forms does not lie in their ability to closely approximate arbitrary technologies. They simply do not have this property. Therefore, it is probably counterproductive to think of a general linear form in terms of approximating the unknown, but true, structure. Rather, it seems more productive to recognize that estimation requires the specification of some functional form. In a classical statistical sense, specifying a functional form in empirical analysis is tantamount to an assumption that the underlying technologies are wholly consistent with that form. Therefore, the most likely contribution of the flexible forms lies not in their approximation properties but in the fact that they apparently place far fewer restrictions prior to estimation than the more traditional Leontief, Cobb-Douglas, and CES technologies. In most instances, they let measures like the elasticity of size and elasticities of substitution depend on the data. Hence, they can vary across the sample and need not be parametric as they are for the more traditional forms. [But] … one should not expect more of them than they are capable of giving.”

As regards application to water supply, if the economic characteristics of water distribution are rather different from those of water acquisition and treatment, trying to represent both activities in a single function, whether flexible or not, may obscure features of interest. If the two activities are treated separately, the question then arises how to bring them together. The algebra involved in combining two flexible form specifications is daunting, and adoption of some simpler specification may be necessary for reasons of tractability.

e. Other issues

A number of other more detailed points emerge from the surveyed literature:

Multiple outputs: How important is it to distinguish between different types of outputs, e.g. residential/non-residential (Kim & Clark(1988)), water supply/sewerage (Stone & Webster (2004), Saal & Parker (2005))?

i. *Treatment of leakage*: Should this be treated as an output (Garcia & Thomas (2001), Stone & Webster (2004)) or as part of distribution cost?

ii. *Effect of density*: Is the favourable effect of increasing density reversed at very high densities (Saal & Parker (2005))? Should density be modeled as declining away from urban centers (Clark & Stevie (1981))?

iii. *Treatment of capital costs*: How should the long-lasting nature of most water assets be reflected in the analysis – by treating some assets as quasi-fixed

(Torres & Morrison Paul (2006)) or by making separate short run and long run estimates (Stone & Webster (2004))?

iv. *Aggregation problems*: The theoretical models tend to assume a system consisting of a treatment works with associated distribution system. However, particularly in England & Wales, water companies serve quite large areas encompassing many largely independent systems. There is thus a question about how to adapt the models, or the data, to reflect this reality.

This review of the literature has thus produced a substantial list of issues that need to be confronted, if not overcome, in developing our own approach in **Chapter III**.

Appendix C

ASSET VALUES FOR WATER COMPANIES IN ENGLAND & WALES

Two different asset values for water companies are available from Ofwat data:

- Regulatory Capital Value (RCV): This is a value established by Ofwat, deriving from the companies' opening balance sheets at privatisation, adjusted year by year subsequently for new investment (after depreciation) and a "capital efficiency" factor. The resulting RCV provides the base for the rate of return on capital allowed by Ofwat in its quinquennial price reviews. This value covers all the companies' activities that are subject to Ofwat regulation and is not further sub-divided into (e.g.) a water supply RCV and a sewerage RCV. In the June Returns, Ofwat (2003a), RCV is included in Table C (line 8) of the Board Overview section. It also features in Ofwat's annual report on the companies' financial performance – e.g. Ofwat (2003c), Table 9, p.28.
- Gross Replacement Cost (GRC) (also sometimes referred to as Modern Equivalent Asset (MEA) value): For water service assets, this is reported in the first four columns of Table 25 of the June Returns (a similar analysis for the sewerage assets of WaSCs appears in further columns of Table 25). Each year the opening balance is adjusted for inflation, disposals and additions during the year, and any adjustments arising from the current Asset Management Plan. Depreciation is then subtracted to give the end-year balance. This value provides the base for capital maintenance charges (incl. depreciation) as recorded in Table 21 of the June Returns. The GRC value of water supply assets is sub-divided between:
  - Water service infrastructure assets;
  - Water service operational assets; and
  - Water service other tangible assets.

  The definitions of "infrastructure assets" and "operational assets" in the Ofwat guidance notes state:

  "**Infrastructure assets** cover the following: underground systems of mains and sewers, impounding and pumped raw storage reservoirs, dams, sludge pipelines and sea outfalls."

> "**Operational assets** cover the following: intake works, pumping stations, treatment works, boreholes, operational land, offices, depots, workshops, etc …"
>
> Thus the former include some assets related to water acquisition (e.g. dams and reservoirs) although the majority relate to water distribution (e.g. mains), while the latter relate almost entirely to water acquisition and treatment:

As GRC values are typically five or more times as large as RCV, there is a real question which to use in economic analysis of water company activities, when a capital value is required. Fortunately, despite the big difference in values, the two measures are closely correlated (See **Figure C.1**), so that when all that is required is an index of capital value which is consistent across companies, as in a cost function, either can be used. Similarly, the cost of capital (allowed rate of return plus capital maintenance) can be expressed as a percentage of either value.

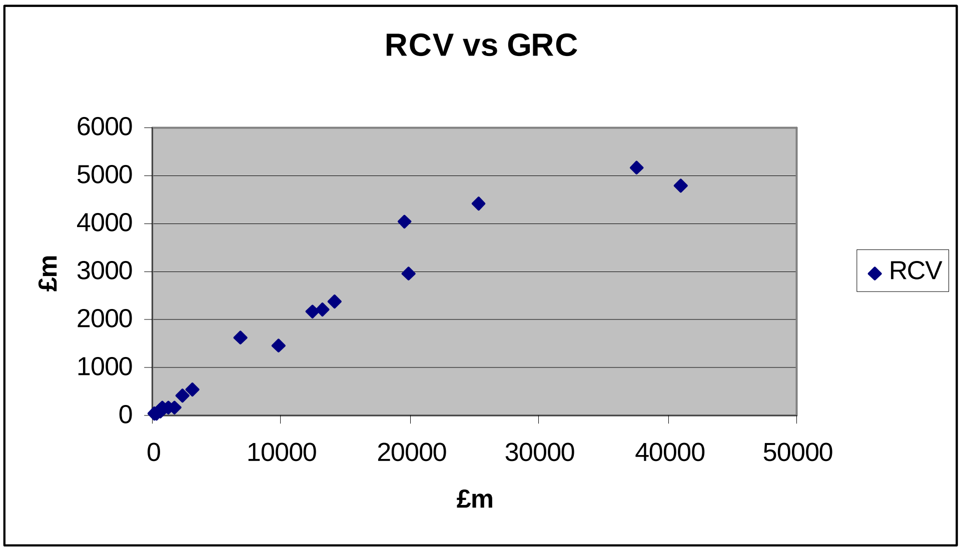

Figure C.1: RCV relative to GRC asset values for water companies in England & Wales

As RCV is used by Ofwat in determining allowed rates of return, it seems preferable to use RCV in the cost functions analysed in **Chapters IV** and **V**. However, this requires a method to estimate the proportion of RCV attributable to water production and water distribution respectively (and excluding the part attributable to sewerage and sewage treatment in the case of WaSCs). For this purpose, an allocation based on GRC values was developed. The steps in the allocation process are listed below and the resulting figures are set out in **Table C.1**:

- Step 1: Take GRC value of Water Operational Assets and add an allocation of Water Infrastructure Assets attributable to water production, based on infrastructure renewals expenditure as recorded in JR Table 21. This gives the GRC value of assets used in water production (*GK_T*).
- Step 2: Take GRC value of Water Infrastructure Assets and subtract the amount allocated in Step 1 to water production. This gives the GRC value of assets used in water distribution (*GK_D*).
- Step 3: (For WaSCs) Take the GRC value of Sewerage and Sewage Treatment Assets (*GK_S*).
- Step 4: Take RCV figures (for all services) from JR Board Overview, Table C and calculate amounts attributable to water production and distribution as:
    - $$\overline{K_T} = \frac{GK_T}{TotalGRC}.RCV$$
    - $$\overline{K_D} = \frac{GK_D}{TotalGRC}.RCV$$

The return on the regulatory value of capital employed by the water companies in 2002-2003, taken from Ofwat (2003c), Table 9, p.28, is given in the table below:

Company	Regulatory capital value (£m)	Return on capital employed (%)
WaSCs **Anglian**	3935.7	4.9
Dwr Cymru (Welsh Water)	2246.4	4.6
Northumbrian	2112.9	4.7
Severn Trent	4270.9	6.4
South West	1550.9	6.3
Southern	2132.1	6.0
Thames	4668.8	6.5
United Utilities (N West Water)	4948.0	5.5
Wessex	1416.4	6.9
Yorkshire	2837.4	6.3
WOCs **Bournemouth & W Hants**	96.2	6.3
Bristol	176.0	7.9
Cambridge	41.3	9.2
Dee Valley	43.3	7.0
Folkestone & Dover	43.7	9.5
Mid Kent	159.9	5.9
Portsmouth	90.1	8.7
South East	422.6	6.9
South Staffs	138.7	8.5
Sutton & E Surrey	104.6	10.1
Tendring Hundred	50.7	8.0
Three Valleys	519.9	6.2

References

Ofwat (2003a) *Water Company June Returns for 2002-2003*

Ofwat (2003c) *Financial performance and expenditure of the water companies in England & Wales, 2002-2003 report.*

Company Accro	Gross Replacement Cost (£m)								Regulatory Capital Value (£m)		
	Water Op Assets	Water Infr Assets	Other water assets	S&ST Assets	Total (all services)	Water Infr Adjustment	GK_T	GK_D	Total (all services)	$\overline{K_T}$	$\overline{K_D}$
WOCS											
BWH	177	414	7	0	598	0	177	414	95.99	28.41	66.45
BRL	473	1243	5	0	1721	128	601	1115	183.13	63.95	118.65
CAM	52	262	14	0	328	0	52	262	41.89	6.64	33.46
DVW	96	207	6	0	309	3	99	204	44.5	14.26	29.38
FLK	56	154	4	0	214	0	56	154	46.98	12.29	33.81
MKT	168	542	19	0	729	24	192	518	166.29	43.80	118.16
PRT	133	484	7	0	624	0	133	484	91.14	19.43	70.69
MSE	666	1644	24	0	2334	52	718	1592	434.77	133.75	296.55
SST	240	1035	41	0	1316	4	244	1031	146.37	27.14	114.67
SES	199	483	14	0	696	0	199	483	110.63	31.63	76.77
THD	68	141	7	0	216	0	68	141	51.06	16.07	33.33
TVN	813	2295	43	0	3151	0	813	2295	526.06	135.73	383.15
WASCS											
ANH	1394	4715	266	13123	19498	20	1414	4695	4032.26	292.42	970.94
WSH	1268	4812	113	7988	14181	110	1378	4702	2362.26	229.55	783.26
YKY	1843	6637	150	11266	19896	2328	4171	4309	2957.12	619.93	640.44
NES	1627	4563	105	6130	12425	312	1939	4251	2171.06	338.81	742.79
SWT	695	2174	59	3896	6824	121	816	2053	1630.32	194.95	490.48
SVT	1749	6480	248	16874	25351	308	2057	6172	4396.96	356.77	1070.49
SRN	854	2570	114	9726	13264	544	1398	2026	2191.84	231.02	334.79
TMS	2555	7053	161	31277	41046	356	2911	6697	4777.62	338.83	779.51
NWT	2719	11006	295	23596	37616	2259	4978	8747	5156.59	682.41	1199.08
WSX	499	2020	31	7163	9713	38	537	1982	1474.4	81.51	300.86

Table C.1: Derivation of capital values for water companies in England & Wales, 2002-2003

Appendix D

COMPONENTS OF WATER DELIVERED

<table>
<tr><td colspan="10"><b>Distribution input (100%)</b></td></tr>
<tr><td colspan="2">Distribution system</td><td colspan="8">Customers installations</td></tr>
<tr><td colspan="2" rowspan="3"></td><td colspan="6">Water delivered and billed</td><td colspan="2">Unbilled water</td></tr>
<tr><td colspan="2">Measured Households (8.3%)</td><td colspan="2">Measured Non-households (41.7%)</td><td>Unmeasured Households (37.8%)</td><td>Unmeasured Non-households (0.9%)</td><td rowspan="2">Taken legally (0.6%)</td><td rowspan="2">Taken illegally (0.0%)</td></tr>
<tr><td colspan="4">Billed measured (50.0%)</td><td colspan="2">Billed unmeasured (38.6%)</td></tr>
<tr><td colspan="2">Water not delivered</td><td colspan="8">Water delivered to customers (89.3%)</td></tr>
<tr><td rowspan="2">Distribution system operational use (0.7%)</td><td>Distribution losses (10.0%)</td><td>Underground supply pipe losses</td><td colspan="2">Plumbing losses</td><td colspan="3">Customer use</td><td colspan="2" rowspan="2"></td></tr>
<tr><td colspan="2"><b>Total leakage (14.1%)</b></td><td colspan="5"><b>Consumption (84.6%)</b></td></tr>
</table>

[**Source**: Adapted from Ofwat guidance notes relating to Table 10 of the June Returns. The % figures are taken from Table 10 of the Bournemouth & West Hants Water Co June Return for 2003 and should be regarded as indicative rather than representative.]

COMPANY LEVEL ESTIMATES OF RETURNS TO SCALE IN WATER PRODUCTION (ENGLAND & WALES)

a. Introduction

Although not directly relevant to the purposes of this thesis as it is unlikely that a company level analysis will throw much light on settlement level effects, it is of some interest to use the methods described in **Chapter IV** to estimate returns to scale at company level for the water companies in England & Wales which report to Ofwat. This is what is done in this Appendix: Note that the analysis here is for water production only; distribution is not included.

b. Data issues

The source for the data on the water companies in England & Wales is described in **Appendix A**. For the analyses reported in this Appendix, data for 2002/3 were assembled covering the water supply operations of the 12 WoCs and the 10 WaSCs that made June Returns to Ofwat in 2003. Some basic figures can be found in the tables in **Chapter II**, together with a key to the company acronyms. **Tables E.1A** and **E.1B** below show the data used in the regressions. In these tables, *VCP* is variable cost of production, *CMP* is capital maintenance cost, *FCP* is financing cost[126] and *TCP* is total cost. $\overline{K_P}$ is the regulatory value of each company's assets used for water production – the derivation of these figures is set out in **Appendix C**. Note also that although there are some imports and exports of bulk water between companies, these are mainly of untreated water, so that the quantity put into distribution by each company (*QDI*) is a good measure of the quantity treated by that company. The other variables in the tables are number of treatment works (*TN*), proportion of surface water (*SP*), resource pumping head (*PHR*) and proportion of water treated to level 4 (*W4P*).

[126] Taken here to be equal to the return on capital employed for each company as reported in the Ofwat financial performance report for 2002/03 (see **Appendix C**) times the regulatory value of water production assets for ($\overline{K_P}$).

Company[127]	VCP (£m)	CMP (£m)	FCP (£m)	TCP £'000	$\overline{K_P}$ (£m)	QDI (Ml/d)	TN	SP (prop)	PHR (m)	W4P prop
BWH	4.082	3.889	1.811	9.782	28.75	157.6	7	0.838	45.0	0.834
BRL	12.679	8.832	5.067	26.578	64.14	291.3	23	0.868	35.4	0.991
CAM	1.635	0.386	0.638	2.659	6.94	73.2	14	0.000	30.4	0.061
DVW	2.909	1.718	1.018	5.645	14.54	69.5	[128] 9	0.936	72.8	0.370
FLK	2.343	1.600	1.190	5.133	12.53	49.5	18	0.000	91.9	0.420
MKT	6.004	2.851	2.653	11.508	44.97	140.7	29	0.115	39.5	0.118
PRT	3.888	1.135	1.710	6.733	19.65	177.2	20	0.130	33.7	0.000
MSE	12.089	5.710	9.325	27.124	135.14	355.2	65	0.302	42.5	0.567
SST	7.756	5.592	2.381	15.729	28.01	330.9	29	0.574	48.0	0.648
SES	7.013	3.304	3.260	13.577	32.28	159.9	11	0.145	77.9	0.937
THD	1.627	1.212	1.329	4.168	16.61	30.1	2	0.139	74.0	0.139
TVW	19.831	25.502	8.532	53.865	137.61	796.0	99	0.448	17.4	0.676

Table E.1A: Key data for WoCs used in this Appendix

Company[2]	VCP (£m)	CMP (£m)	FCP (£m)	TCP (£m)	$\overline{K_P}$ (£m)	QDI (Ml/d)	TN	SP (prop)	PHR (m)	W4P prop
ANH	35.66	46.09	14.52	96.27	296.4	1150	143	0.488	91.2	0.640
WSH	39.19	32.25	10.64	82.08	231.4	883	105	0.963	79.8	0.125
YKY	39.91	53.47	29.36	122.74	624.6	1299	90	0.787	99.0	0.352
NNE	44.37	21.85	23.01	89.23	359.6	1201	67	0.897	47.6	0.439
SWT	16.90	11.02	12.39	40.31	196.6	447	40	0.897	43.4	0.446
SVT	56.00	55.10	21.62	132.72	360.3	1958	173	0.680	56.3	0.472
SRN	16.03	28.78	15.15	59.96	233.0	595	102	0.296	28.3	0.459
TMS	53.60	64.20	18.71	136.51	340.2	2804	99	0.781	37.8	0.862
NWT	42.84	61.89	50.15	154.87	726.8	1952	137	0.915	33.6	0.277
WSX	9.20	11.05	5.15	25.40	81.8	368	119	0.260	28.8	0.166

Table E.1B: Key data for WaSCs used in this Appendix

c. Specification and results

Based on (4.8) in **Chapter IV**, the specification adopted here is:

$$\ln VCP = \alpha_0 + \alpha_1 \ln QP + \alpha_2 (\ln QP)^2 + \alpha_3 \ln \overline{K_P} + \alpha_4 \ln(1+SP) + \alpha_5 \ln PHR + \alpha_6 \ln(1+W4P) \quad \ldots \text{(E.1)}$$

The results obtained are shown in **Table E.2**:

[127] For key to company acronyms, see **Tables 3.1A** and **3.1B** in **Chapter III**.

[128] The June Return gives 11 but it was found that 2 of these relate to supplies from another company.

Coefficents	All companies		10 WaSCs		12 WOCs	
	With $(\ln QP)^2$	without	With $(\ln QP)^2$	without	With $(\ln QP)^2$	without
α_0 (Const)	-3.996***	-3.279***	-23.71***	-2.966**	-3.384	-3.245*
S.E.	*1.135*	*0.510*	*5.394*	*0.777*	*1.936*	*1.375*
α_1 (ln*QP*)	0.820	0.592***	7.226***	0.723	0.621	0.547**
S.E.	*0.336*	*0.096*	*1.688*	*0.162*	*0.680*	*0.162*
α_2 $(\ln QP)^2$	-0.020	Dropped	-0.460**	Dropped	-0.008	Dropped
S.E.	*0.028*		*0.119*		*0.074*	
α_3 $(\ln\overline{K_P})$	0.279***	0.279***	-0.270**	-0.044	0.387**	0.380**
S.E.	*0.086*	*0.085*	*0.099*	*0.168*	*0.133*	*0.110*
α_4 (ln1+*SP*)	-0.133	0.142	1.021***	0.826	-0.051	-0.047
S.E.	*0.234*	*0.230*	*0.225*	*0.463*	*0.320*	*0.291*
α_5 (ln*PHR*)	0.247**	0.223**	0.075	0.308*	0.181	0.192
S.E.	*0.104*	*0.096*	*0.088*	*0.137*	*0.262*	*0.224*
α_6 (ln1+*W4P*)	0.186	0.211	0.126	-0.039	0.315	0.306
S.E.	*0.229*	*0.222*	*0.227*	*0.472*	*0.459*	*0.413*
R^2	0.9809	0.9802	0.9939	0.9634	0.9718	0.9718

Table E.2: Regression results, water production, Ofwat data, using (E.1)
(Significance levels: * = 1%; ** = 5%; * = 10%; relative to 1 for α_1)**

Looking at the first ("All companies") columns of **Table E.2**, there is some evidence of scale economies in that the coefficient on ln*QP* is less than 1 (but not significant), while the coefficient on $\ln\overline{K_P}$ is also less than 1 (and significantly so). Dropping the $(\ln QP)^2$ term gives a much stronger indication: using the relationships from **Chapter IV, section 1 (a)** for returns to scale gives RTS_S = 1.69 and RTS_L = 1.22. However, these values seem very high. Pumping head is found to have a significant effect on costs but surface water proportion and treatment to level W4 apparently do not.

To test for difference between WaSCs and WoCs, a WoC dummy was tried but found not to be significant. However, running (E.1) for WaSCs and WoCs separately produced the rather striking differences shown in the second and third pairs of columns of **Table E.2**. Of course, the number of degrees of freedom in these regressions has become extremely small, but the results nevertheless seem to indicate some important difference between WaSCs and WoCs. Whereas the coefficients for WoCs are roughly as might be expected, those for WaSCs look distinctly odd (and the implied returns to scale are again high, e.g. if the $(\ln QP)^2$ term is dropped, then for WOCs RTS_S = 1.83 and RTS_L = 1.13 while for WaSCs RTS_S = 1.38 and RTS_L = 1.44). With WaSCs there is also a switch in the sign on $\ln\overline{K_P}$**.** It is not obvious why this should be so. One possibility is a systematic difference in accounting treatment for capital assets as between WoCs and

WaSCs although adherence to Ofwat guidance should obviate this. Another possibility is that the controls do not adequately deal with the fact that WoCs are about twice as reliant as WaSCs on groundwater from boreholes (64% compared with 33%)[129]. To test whether average works size might have an effect, (E.1) was re-run with an additional term in ln*TN* (number of works)[130] but this was found not to be significant (although positive), either for all companies or for WaSCs and WoCs separately.

An alternative specification, based on (4.2) in **Chapter IV**, treating water production capital as variable rather than quasi-fixed, is:

$$\begin{aligned}\ln TCP = \alpha_0 + \alpha_1 \ln QP + \alpha_2 (\ln QP)^2 + \alpha_3 \ln(\delta + \tau) + \alpha_4 \ln(1 + SP) \\ + \alpha_5 \ln PHR + \alpha_6 \ln(1 + W4P)\end{aligned} \quad \ldots \quad \text{(E.2)}$$

Here the dependent variable is the full cost of water production ($TCP = VCP + CMP + FCP$) and δ and τ together make up the cost of capital $\left(\delta = \frac{CMP}{\overline{K_P}}; \tau = \frac{FCP}{\overline{K_P}}\right)$. The results obtained using (E.2) are shown in **Table E.3**.

Coefficents	**All companies**		**10 WaSCs**		**12 WOCs**	
	With $(\ln QP)^2$	without	With $(\ln QP)^2$	without	With $(\ln QP)^2$	without
α_0 (Const)	-3.428*	-4.193***	-19.28*	-3.272**	0.888	-3.291
S.E.	*1.924*	*0.958*	*9.41*	*1.470*	*2.728*	*3.051*
α_1 (ln*QP*)	0.705	0.967	6.015*	0.938	-1.691	0.799
S.E.	*0.572*	*0.071*	*2.965*	*0.174*	*0.989*	*0.288*
α_2 (ln*QP*)^2	0.022	Dropped	-0.373	Dropped	0.260**	Dropped
S.E.	*0.049*		*0.218*		*0.101*	
α_3 (ln**δ+τ**)	-0.307	-0.276	-0.129	-0.438	-0.698	-0.371
S.E.	*0.322*	*0.307*	*0.375*	*0.401*	*0.406*	*0.537*
α_4 (ln1+*SP*)	0.134	0.127	-0.065	-0.442	0.140	-0.071
S.E.	*0.401*	*0.391*	*0.614*	*0.699*	*0.463*	*0.634*
α_5 (ln*PHR*)	0.268	0.297*	-0.024	0.190	0.419	0.164
S.E.	*0.173*	*0.157*	*0.178*	*0.155*	*0.362*	*0.484*
α_6 (ln1+*W4P*)	0.362	0.310	-0.130	-0.357	1.017	1.116
S.E.	*0.456*	*0.431*	*0.461*	*0.539*	*0.708*	*0.984*
R^2	0.9534	0.9528	0.9738	0.9482	0.9504	0.8846

Table E.3: Regression results, water production, Ofwat data, using (4.12) (Significance levels: * = 1%; ** = 5%; * = 10%; relative to 1 for α_1)**

[129] Thus, among the WoCs, Cambridge (CAM) operates only boreholes and Portsmouth (PRT) has 19 boreholes out of 20 sources; on the other hand, Folkestone (FLK) which also only operates boreholes has relatively high capital maintenance charges.

[130] The number of reported works ranges from 2 for THD to 173 for SVT.

As can be seen, the coefficients obtained here are mostly not significant, and dropping the (lnQP)^2 term does not improve matters. However, the returns to scale in water production indicated now appear more reasonable with RTS_L = 1.03 for all companies, RTS_L = 1.25 for WOCs and RTS_L = 1.07 for WaSCs.

A summary of these results, including those when the (lnQP)^2 term is included, evaluated at QP = 1055 Ml/day (average value for WaSCs) and QP = 220 Ml/day (average value for WoCs), is shown in **Table E.4**. A value greater than 1 suggests scale economies, a value less than 1 diseconomies (and a value less than 0 is invalid). Although a rather wide range of values emerges, those calculated without the term in (lnQP)^2 are consistently greater than 1, suggesting that there probably are economies of scale in water production at company level in England & Wales.

	RTS$_S$		**RTS$_L$**	
	With (lnQP)^2	Without	With (lnQP)^2	Without
Using (E.1) **All Cos (evaluated at *QP* = 1055 Ml/day)**	1.47	1.69	1.06	1.22
WaSCs (evaluated at *QP* = 1055 Ml/day)	0.25	1.38	0.34	1.44
All Cos (evaluated at *QP* = 220 Ml/day)	1.40	1.69	1.01	1.22
WoCs (evaluated at *QP* = 220 Ml/day)	1.73	1.83	1.06	1.13
Using (E.2) **All Cos (evaluated at *QP* = 1055 Ml/day)**			1.17	1.03
WaSCs (evaluated at *QP* = 1055 Ml/day)			0.29	1.07
All Cos (evaluated at *QP* = 220 Ml/day)			1.21	
WoCs (evaluated at *QP* = 220 Ml/day)			-3.46	1.25

Table E.4: Company level returns to scale indicated using (4.11) and (4.12) and Ofwat data

Appendix F

SIMPLIFIED MODELS OF WATER DISTRIBUTION

The water distribution system of any settlement tends to be a reflection of history and local geography rather than technical or economic optimization, making generalization difficult. However, by constructing simple models of distribution systems, some results can be derived which can be used to help guide empirical investigation. Having regard to the data to be used, the model development considers both capital and operating costs. The key questions on which the models are designed to shed some light are how these cost elements vary with water usage per property, number of connected properties and size of service area. To explore these questions, a model of distribution costs is developed first for a linear settlement, and then for a square settlement. These models indicate, *inter alia,* that the distribution cost per litre of water can be expected to increase as the size of the service area increases. This is essentially because as the service area increases, the average distance over which water must be delivered also increases. However, higher density of demand, whether due to more connected properties per hectare or higher usage per property will tend to offset this effect, to an extent that depends on the relative size of the various cost parameters.

a. Linear settlement

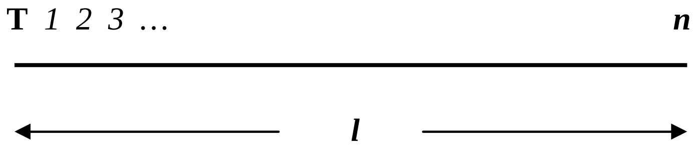

Figure F.1: The linear settlement

Figure F.1 shows a linear settlement of length ***l*** with ***n*** properties, which are equally distributed along the settlement. Each property is connected to a feeder pipe which runs the length of the settlement. Water is pumped from the point **T** and each property is assumed to consume ***w*** units of water per annum.

We start with annual *capital costs* (capital maintenance *plus* return on capital) which we suppose to be linearly related to two components of the system: the feeder pipe, with costs m_f per unit length, and the connection with cost m_c per connected property. Hence, annual capital costs are given in this case by:

$$CMD = l.m_f + n.m_c \qquad \text{……………………..} \qquad \text{(F.1)}$$

And the capital maintenance cost per unit volume (dividing by the volume of water used, ***n.w***) will be:

$$ACMD = \frac{1}{w}\left(\frac{l.m_f}{n} + m_c\right) \qquad \text{……………………..} \qquad \text{(F.2)}$$

which can be expressed in terms of property density, ***d*** (= ***n/l***) as

$$ACMD = \frac{1}{w}\left(\frac{m_f}{d} + m_c\right) \qquad \text{……………………..} \qquad \text{(F.3)}$$

From (F.3) it may be seen that in this model capital costs show constant returns to scale with respect to the number of properties connected (density held constant) but that higher property density or higher usage per property leads to savings in this average cost. (In the linear case, property density can only be measured as properties/km of mains; in the square settlement case below, an alternative measure, properties/sq.km of service area, is also available.)

Turning to *operating costs*, it is likely that some part of these will also be related to length of pipes and number of connections – these costs are denoted by $\boldsymbol{c_f}$ and $\boldsymbol{c_c}$ respectively - but in addition there will be volume related costs, notably pumping costs. To model pumping costs, we suppose that water is pumped directly into distribution from the point T and that pumping cost is $\boldsymbol{p_f}$ per unit volume per unit distance of feeder pipe.

For a linear settlement, using this technology, annual pumping costs will then be:

$$p_f.\frac{l}{n}.w(1+2+3+...+n) \approx \frac{1}{2}p_f.w.l.n \text{ [131]} \qquad \text{……………} \qquad \text{(F.4)}$$

Adding in the part of operating costs related to ***l*** and ***n***, and dividing by ***n.w*** then gives an expression for average operating cost per unit volume:

$$AVCD = \frac{1}{w}\left(\frac{l.c_f}{n} + c_c\right) + \frac{1}{2}p_f.l \qquad \text{…………….} \qquad \text{(F.5)}$$

which can be expressed in terms of property density ***d*** as:

$$AVCD = \frac{1}{w}\left(\frac{c_f}{d} + c_c\right) + \frac{1}{2}p_f.\frac{n}{d} \qquad \text{…………….} \qquad \text{(F.6)}$$

[131] Using the approximation $\sum_1^n n = \frac{1}{2}n(n+1) \approx \frac{1}{2}n^2$ for large *n*.

From this latter expression, it can be seen that average distribution operating cost is increasing in the number of connected properties but decreasing in property density and usage per property.

The effect of bringing capital and operating costs together for a linear settlement using (F.3) and (F.6) is illustrated in **Figures F.2 and F.3** below:

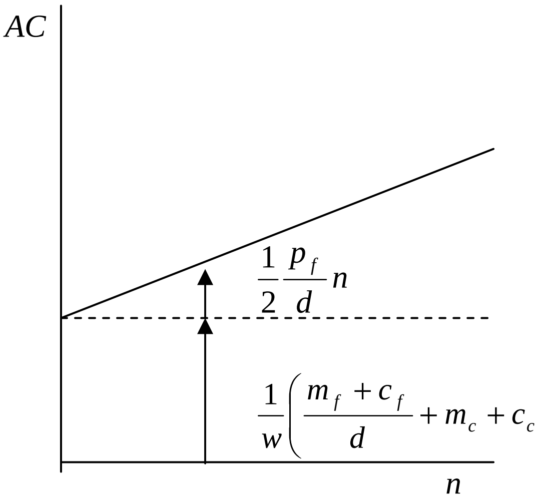
AC
$\frac{1}{2}\frac{p_f}{d}n$
$\frac{1}{w}\left(\frac{m_f + c_f}{d} + m_c + c_c\right)$
n

Figure F.2: Relationship between average distribution cost and number of properties (linear settlement

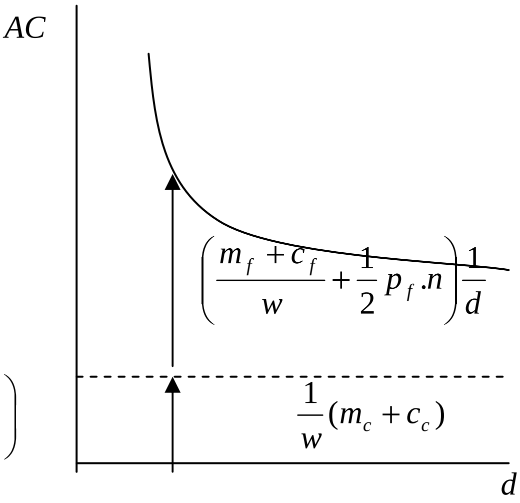
AC
$\left(\frac{m_f + c_f}{w} + \frac{1}{2}p_f.n\right)\frac{1}{d}$
$\frac{1}{w}(m_c + c_c)$
d

Figure F.3: Relationship between average distribution cost and property density (linear settlement)

Figure F.2 shows how, in this model, increasing settlement size (more properties, density and usage held constant) results in higher average water distribution costs because more water has to be pumped over greater distances. There is thus a diseconomy related to settlement size. This diseconomy may not be very great if pumping cost (***p*$_f$**) is low but even if pumping cost is zero, there are no scale economies, only constant returns to scale.

Figure F.3 however shows how a more compact settlement (higher property density) results in savings in both mains costs and pumping costs per unit volume. It seems natural to call this effect *density economies*.[132]

[132] A different definition is offered by Stone & Webster (2004) p.16 “The scale expansion of a water service firm is most appropriately defined by the expansion of throughput (volumes) and customers served (connections) … Economies of production density (EPD) inform us as to the relationship between costs and production, when holding the number of customers or connections constant. Economies of customer density (ECD) inform us as to the relationship between costs and scale when the number of customers is not held constant.” This formulation however seems to miss the important effect on costs of differences in density (properties/hectare), while conflating water treatment and distribution.

Next, we may note that pumping costs can be reduced, with this technology, by locating the supply point **T** within the community. In fact, pumping costs can be halved by relocating **T** to the midpoint of ***l***. We will make use of this point in the next section but it does not alter the fundamentals of the situation as expressed in **Figures F.2 and F.3** above.

A different perspective on pumping costs emerges if we suppose that water for distribution is pumped into a water tower at **T** and then fed by gravity along the feeder pipe to the ***n*** properties in the linear community. In this case, pumping costs reduce to $\boldsymbol{n.w.p_t}$ – where $\boldsymbol{p_t}$ is the cost of pumping one unit of water into the tower. In this case, average cost is unaffected by population size (for a given density) but there is still a density economy as the cost of the feeder pipe is spread over a larger volume of water.

In practice, pumping costs are likely to lie somewhere between the two cases outlined above. Most communities rely to a large extent on gravity feed from water towers or service reservoirs but the number and/or size of such facilities is likely to be related to the size of the community because of factors such as loss of pressure in the distribution system, and some distribution pumping is still likely to be required. In urban areas, the cost of getting water to high rise buildings may add to distribution costs.

b. Square settlement

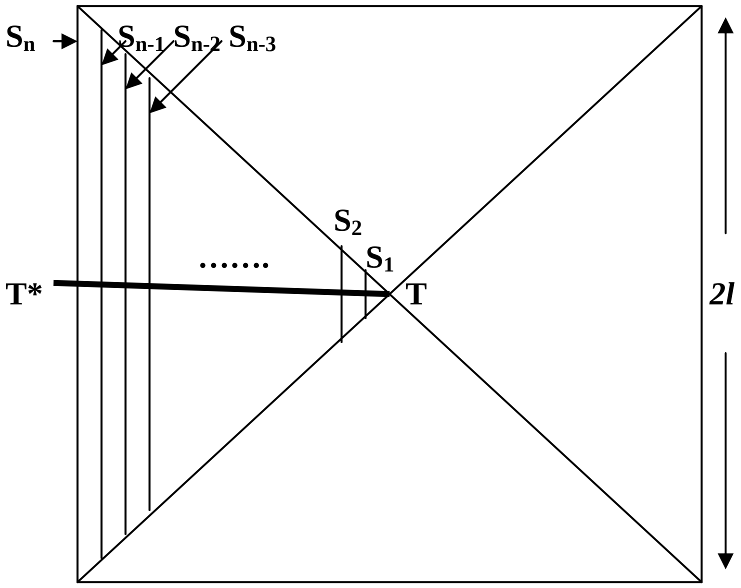

Figure F.4: The square settlement

For greater realism, we need a two-dimensional model and **Figure F.4** shows a square settlement**.** This may seem at first an odd case to investigate but it follows naturally from the linear case; nor is it particularly unrealistic given that many urban settlements are built on a grid pattern[133] (and qualitatively similar results can be obtained using a circular model.) The settlement is divided into 4 triangular sectors (water supply areas). Each sector contains a main supply pipe from the point **T** at the center of the settlement and a number of "streets" **S** along each of which runs a feeder pipe to which properties are connected. Each street can then be viewed as a linear settlement. If the side of the settlement measures ***2l***, then, in each segment of the settlement the length of the main supply pipe from **T** will be ***l*** and the length of the longest street ***2l***. Properties are assumed to be equally distributed along each street at the same rate as in the linear settlement, i.e. ***n*** properties per length ***l*** of street. The streets are also assumed to be at a distance ***l/n*** apart; there are therefore ***n*** streets in each segment. (Evidently, different assumptions could be made, which would complicate the arithmetic without affecting the general character of the results.)

For *capital costs,* there are three parts of the infrastructure to consider: the main supply pipe, the feeder pipes and the connections. Calculation shows that the total length of the streets in each triangular sector is approximately ***n.l*** and the number of properties is $\boldsymbol{n^2}$ (and the density of properties is therefore $\boldsymbol{n^2/l^2}$). If the capital cost per unit length of the main supply pipe is $\boldsymbol{m_m}$, it then follows (other symbols as in the linear settlement case) that infrastructure costs per sector are given by:

$$CMD = l.m_m + l.n.m_f + n^2.m_c \qquad \ldots\ldots\ldots\ldots\ldots\ldots \qquad \text{(F.7)}$$

For *operating costs,* using the direct pumping technology, the pumping costs that need to be assessed include both the cost of pumping along the feeder pipes in each street and the cost of pumping along the main supply pipe. These can be shown to amount respectively to approximately[134] $w.l\left(\frac{1}{3}n^2 + n\right)p_f$ and $\frac{2}{3}p_m w.l.n^2$ per sector**,** where $\boldsymbol{p_m}$ is unit pumping cost in the main supply pipe.

[133] Reflected in the use of rectilinear ("Manhattan") distances as one of the standard approaches in the facilities location literature – see, for example, Hansen *et al* (1983), p. 227-8.

[134] Using the approximation $\sum_1^n n^2 = \frac{n}{6}(2n^2 + 3n + 1) \approx \frac{1}{3}n^3$ for large *n*.

To allow for the likelihood that there will also be elements of operating costs related to the length of pipes and the number of connections, further terms, $l.\,c_m, n.l.c_f$ and n^2c_c, may be added. Adding these to pumping costs, leads to an expression for distribution operating costs per sector:

$$VCD = l.c_m + l.n.c_f + n^2.c_c + w.l\left(\frac{1}{3}n^2 + n\right)p_f + \frac{2}{3}w.l.n^2 p_m \qquad \text{........} \qquad \text{(F.8)}$$

Putting (F.7) and (F.8) together gives an expression for the total cost of distribution:

$$TCD = l(m_m + c_m) + l.n(m_f + c_f) + n^2(m_c + c_c) + w.l.n\left(\frac{1}{3}n + 1\right)p_f + w.l.n\left(\frac{2}{3}n\right)p_m$$

$$\text{....} \qquad \text{(F.9)}$$

To inform empirical work, it is helpful to adapt the expression (F.9) to use quantities likely to be observed in practice. Thus the number of properties served, $N = n^2$; the size of the service area, $A = l^2$; and the length of mains (including both main and feeder pipes), $M = l(n + 1)$. While both A and M provide a measure of the area served, it is likely in practice to be preferable to use M (where data on length of mains is available) because very often parts of the area measured by A will be unoccupied or unserviced. Further simplification is possible by treating the prices m, c and p as constants and assuming that:

$$(m_m + c_m) = (m_f + c_f) = \alpha_1$$
$$(m_c + c_c) = \alpha_2$$
$$p_f = p_m = \alpha_3$$

Proceeding in this way leads to:

$$TCD = \alpha_1 M + \alpha_2 N + \alpha_3 w.M\sqrt{N} \qquad \text{...........} \qquad \text{(F.10)}$$

Using A rather than M, assuming no unoccupied or unserviced land) this expression would be:

$$TCD = \alpha_1\sqrt{A}(1 + \sqrt{N}) + \alpha_2 N + \alpha_3 w.N\sqrt{A} \qquad \text{...........} \qquad \text{(F.11)}$$

It may be noted that (F.11) shows distribution costs in this model to be a function of average water usage (w), the number of properties served (N) and service area (A). This is consistent with distribution output being a function of these same variables, as was found in **Section 3** above. However, it differs in that $w.N$ (total consumption) does not enter into the relationship in a simple multiplicative way and we now have a linear combination of terms.

Some further insight can be obtained by examining the partial derivatives of (F.10) with respect to usage per property (*w*), number of connected properties (*N*) and length of mains (*M*). The precise form of these partial derivatives is a consequence of the particularities of the model but – see derivation in the **Annex** to this **Appendix** – it is possible to make some assessment of their likely range of values and the implications of these for economies of scale, as shown in **Table F.1** below.

Elasticity	Likely range	Implication for economies of scale
ε_W	$0 < \varepsilon_W < 1$	Economies of scale wrt *w*, given *N* and *M*.
ε_N	$0 < \varepsilon_N < 1$	Economies of scale wrt *N*, given *w* and *M*.
ε_M	$0 < \varepsilon_M < 1$	Economies of scale wrt *M*, given *w* and *N*.
$\varepsilon_S = \varepsilon_N + \varepsilon_M$	$0 < \varepsilon_S < 2$	Could be economies or diseconomies, depending on values of ε_N and ε_M
ε_D	$-1 < \varepsilon_D < 0$	There are density economies

Table F.1: Signs of partial derivatives (Square settlement, using length of mains)

If the square settlement model has succeeded in representing something of the real economics of water distribution systems, these are the general effects that one might expect to observe empirically.

As a further refinement, visual inspection of **Figure F.4** suggests that distribution costs would be reduced if water were pumped from T* rather than T. In fact, on the same assumptions as above, the cost of pumping along the main supply pipe is approximately halved (other costs are not affected). Whether this is the most economic solution overall will depend on the extent of economies of scale in water treatment. Unless the saving in production costs with a single treatment plant with capacity $\boldsymbol{4n^2}$ compared with 4 plants of capacity $\boldsymbol{n^2}$ is more than this difference in pumping costs, i.e. $4x\frac{1}{3}w.l.n^2 p_m$, it will be more economic to have 4 smaller treatment plants located at the edges of the settlement (at points such as **T***) rather than one large works at a central location. This is an example of the kind of trade-off that we hope our empirical work will throw light on.

Turning to the case where the water supply is pumped to a tower and then distributed by gravity, the pumping cost simplifies to $w.n^2$ (water usage) x p_t (unit pumping cost), so that distribution cost is now less affected by the size of the service area (although it should be noted that this result takes no account of differences in the cost of providing water towers for settlements of different sizes).

A further comment worth making at this point is that **Table F.1** shows that in the square settlement model the elasticity of distribution cost with respect to number of properties ε_N is less than 1 whereas if water distribution costs were to be modeled as commuting costs are in Arnott (1979), ε_N would be greater than 1.

c. How close to reality is this model?

Apart from identifying the likely drivers of distribution costs, one useful outcome of the development of this model is the derivation of expectations as to the strength of scale elasticities with respect to water usage per property, numbers of properties and length of mains. However, these results have been modeled on service areas which are fully occupied with properties at a uniform density. In reality it is more likely that densities will tend to decline away from the centre of each settlement (with suburbanization, for example) and that service areas may include several settlements with perhaps quite large more or less unoccupied space in between settlements. Unoccupied space is reasonably addressed by the use of length of mains in place of geographical area but the other issues raise questions about how reliable a guide the models will prove to be when confronted with real data.

To examine the effect of a suburban fringe, a modified version of the square settlement pictured in **Figure F.4** was considered. The length of the side of the square was increased by a factor k ($k > 1$), and the distance between properties and streets in the added area was assumed to be $l.k/n$, i.e. less dense than in the central square where these distances are l/n[135] . The effect on the expression for total distribution costs previously obtained can be derived by calculating these costs for the larger less dense square and then substituting the previous results in the central area. The new expression, which can be compared with (F.9) above, is:

[135] Attempts to find a more elegant representation of declining density, e.g. by having the distance between streets and houses increase by a factor (1 + k) at each step led to expressions more difficult to interpret than (F.11).

$$TCD = l.k(m_m + c_m) + l.n\left(\frac{k^2+k-1}{k}\right)(m_f + c_f) + n^2\left(\frac{2k^2-1}{k^2}\right)(m_c + c_c)$$
$$+\frac{1}{3}w.l.n\left(\frac{k^3(n+3)+k^2(n+3)-3k-n}{k^2}\right)p_f + \frac{2}{3}w.l.n\left(\frac{n.k^3+n.k^2-n}{k^2}\right)p_m$$

…. (F.12)

Adopting the same constants as in (F.10) leads to:

$$TCD = \alpha_1\left[l.k + l.n\left(\frac{k^2+k-1}{k}\right)\right] + \alpha_2\left[n^2\left(\frac{2k^2-1}{k^2}\right)\right] + \alpha_3 w.n.l\left[(k+1)(n+1) - \left(\frac{k+n}{k^2}\right)\right]$$

………… (F.13)

Unfortunately, although the coefficient on α_1 is the new length of mains and the coefficient on α_2 is the new number of properties, the coefficient on α_3 cannot easily be expressed in terms of these quantities. However, some of the implications can be exposed by means of numerical example.

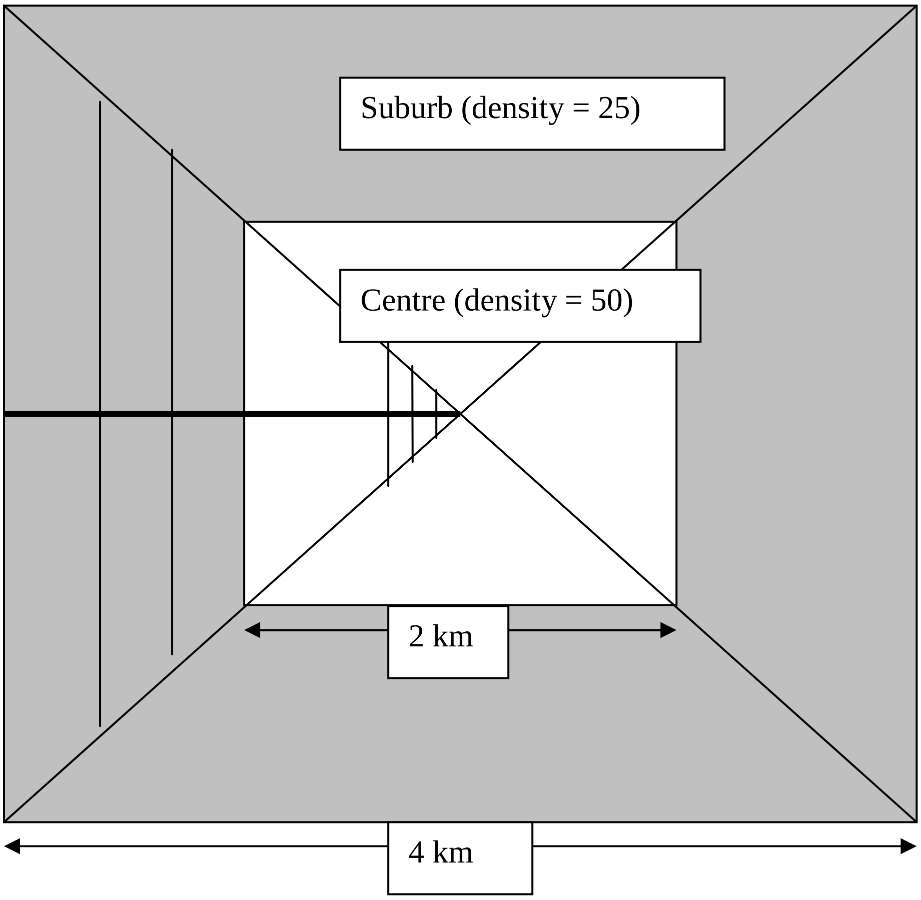

Figure F.5: Square settlement, enlarged by addition of a lower density “suburb”

In **Figure F.5**, the size of the settlement depicted in **Figure F.4** has been enlarged by the addition of a surrounding “suburb”. The central square has side 2 km (i.e. *l* = 1 km), while the enlarged area has side 4 km (i.e. *k* = 2). Property density (*n/l*) is taken to be 50 properties/km in the centre and 25 properties/km in the “suburb”. Then, using (F.13), the results shown in **Table F.2** can be obtained.

	Base case (Figure 3.6)	**(a) Side doubled, Constant density**	**(b) Side doubled, Density halved in “suburb”**
Length of side	2 km	4km	4km
Service area (*A*)	4x1 sq km	4x4 sq km	4x4 sq km
Property density (*n/l*) Centre square	50 props/km	50 props/km	50 props/km
Property density (*n/l*) “suburb”	n.a.	50 props/km	25 props/km
No of properties (*N*)	4x2,500 =10,000	4x10,000 =40,000	4x4,375 =16,500
Length of mains (*M*)	4x51 =204km	4x202 =808km	4x127 =508km
Density A (*N/A*)	400 props/sq.km	400 props/sq.km	175 props/sq.km
Density M (*N/M*)	49.02 props/km	49.5 props/km	34.45 props/km
Distribution cost/litre			
α_1 x	0.0204/w	0.0202/w	0.029/w
α_2 x	1/w	1/w	1/w
α_3 x	1.02	2.02	1.603

Table F.2: Square settlement, effects of enlargement (a) at constant density, (b) at lower density

The effect of doubling the side of the square is to increase area fourfold; the effect of halving the density measured as *n/l* is to reduce population density from 2,500 properties/sq.km in the centre to 625 properties/sq.km in the “suburb”. The table also brings out how these assumptions affect average property density over the whole settlement, measured as properties/sq.km or properties/km of mains (the effect on the latter is less pronounced[136]). The main interest of the table however lies in the effect on distribution cost per litre in the bottom rows. This calculates the coefficients on α_1 (mains related cost), α_2 (connection related cost) and α_3 (pumping related cost). These effects are:

a. Enlargement at constant density: It can be seen, comparing column (a) in **Table F.2** with the base case, that enlargement at constant density makes little difference to mains and connection cost per litre, but pumping cost per litre is almost doubled due to the greater distances involved in serving the “suburb”. (In comparing the value of these

[136] These measures are of course linked by the identity $\frac{N}{A} = \frac{N}{M}.\frac{M}{A}$.

coefficients, it should be kept in mind that α_1 is likely to be very substantially larger than α_3.)

b. Enlargement at lower density: Comparing column (b) with the base case, it can be seen that mains related cost has increased by about 42% (although the increase in property numbers is 75%, the increase in length of mains is over 150%); pumping cost per litre is also significantly higher, although not so much as in column (a).
Hence, two predictions follow:

(a) When comparing distribution costs for urban areas with similar density, we can expect to find that the average cost of distribution (particularly operating costs) will be higher in the area which is larger;

(b) When comparing distribution costs for areas with a high proportion of urban land with areas with a lower proportion, we can expect to find higher average operating costs in the former but higher average capital costs in the latter.

The situation where there are several settlements in each service area is more difficult to address. Much will depend on the actual water supply arrangements in each area. If each settlement has its own supply so that there are no connections between them, it might be possible to subtract unoccupied (or unserviced) areas from the service area and then apply the models developed above to the settlements individually (noting that it is likely to make a significant difference how many settlements there are, and whether they are large or small[137]). If however, one settlement acts as a hub, supplying water to other settlements in the area, some way of identifying the costs which relate to the connecting reticulation would need to be found. Either way, using the raw data for service area (or property density based on this area) is liable to be misleading. While using length of mains rather than service area should help, a further possibility, where the necessary additional information can be obtained, is to divide the service area into “urban” and “rural” components and then either (i) introduce the urban proportion of the service area as a control variable, or (ii) test the relationships using just the “urban” part of the service area, or (iii) develop separate relationships for “urban” and “rural” areas.

In the empirical work reported in **Chapter V**, there was not sufficient time or resources to assemble data of this kind for all the water companies which report to Ofwat. However, it did prove possible to derive the proportion of urban land for the 178 water

[137] For example, the implications for distribution costs would not be the same if there was one settlement occupying 25% of the service area or 5 settlements each occupying 5%.

quality zones of one company; and in the case of US water utilities, the majority appear to serve single communities so that the problem there may not be particularly severe. This part of the analysis was therefore able to take this factor into account.

Annex to Appendix F

DERIVATION OF DISTRIBUTION ELASTICITIES

In **Appendix F**, expressions were derived for the distribution costs of a square settlement. The expression for total distribution costs that emerged was – see (F.10):

$$TCD = \alpha_1 M + \alpha_2 N + \alpha_3 wM\sqrt{N} \quad \ldots\ldots\ldots\ldots \quad \text{(AF.1)}$$

The distribution elasticities implied by (AF.1) are:

$$\varepsilon_W = \frac{w}{TCD}.\frac{\partial(TCD)}{\partial w} \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad \text{(AF.2)}$$

$$\varepsilon_N = \frac{N}{TCD}.\frac{\partial(TCD)}{\partial N} \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad \text{(AF.3)}$$

$$\varepsilon_M = \frac{M}{TCD}.\frac{\partial(TCD)}{\partial M} \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad \text{(AF.4)}$$

Hence:

$$\varepsilon_W = \frac{w}{TCD}.\left(\alpha_3 M\sqrt{N}\right) = \frac{\alpha_3 wM\sqrt{N}}{(\alpha_1 M + \alpha_2 N + \alpha_3 wM\sqrt{N})} \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad \text{(AF.5)}$$

As all the terms in this expression are positive, it follows that $0 < \varepsilon_W < 1$, that is to say, it implies economies of scale with respect to consumption per property (*w*), for given *N* and *M*.

And:

$$\varepsilon_N = \frac{N}{TCD}.\left(\alpha_2 + \frac{1}{2}\alpha_3 \frac{M}{\sqrt{N}}\right) = \frac{\alpha_2 N + \frac{1}{2}\alpha_3 wM\sqrt{N}}{(\alpha_1 M + \alpha_2 N + \alpha_3 wM\sqrt{N})} \quad \ldots\ldots \quad \text{(AF.6)}$$

Which will be less than 1 if $\alpha_1 M + \alpha_3 wM\sqrt{N} > \frac{1}{2}\alpha_3 wM\sqrt{N}$, which is clearly the case. It follows that $0 < \varepsilon_N < 1$, implying economies of scale with respect to number of properties (*N*), for given *w* and *M*.

Finally:

$$\varepsilon_M = \frac{M}{TCD}.\left(\alpha_1 + \alpha_3 w\sqrt{N}\right) = \frac{\alpha_1 M + \alpha_3 wM\sqrt{N}}{(\alpha_1 M + \alpha_2 N + \alpha_3 M\sqrt{N})} \qquad \ldots\ldots \quad \text{(AF.7)}$$

Again, it is clear, since $\alpha_2 N$ is positive, that $0 < \varepsilon_M < 1$ implying economies of scale with respect to length of mains (*M*), for given *w* and *N*.

However, if numbers of properties and length of mains increase together, for given *w* (e.g. if settlements of different sizes but similar density are compared), (AF.6) and (AF.7) could combine to produce diseconomies of scale. If, for example, $\varepsilon_N = 0.4$ and $\varepsilon_M = 0.7$, then $\varepsilon_S = \varepsilon_N + \varepsilon_M = 1.1$.

Using these elasticities suggests two different measures of density effects. If density increases because there are more properties (*N*) in a given area, ε_N applies. For example, if $\varepsilon_N = 0.4$, a 10% increase in *N* leads to a 4% increase in *TCD*. On the other hand, if density increases because the same number of properties are served with a lower length of mains, ε_M is the relevant measure. For example, if $\varepsilon_M = 0.7$, a 10% reduction in *M* leads to a 7% reduction in *TCD*. Either way, there are density economies, of about 5% in the former case, or about 7% in the latter case – see **Table AF.1** below. The larger savings in the second case can be attributed to savings in infrastructure costs (mainly pipes) with a smaller service area.

	Start position	**10% increase in *N***	**10% decrease in *M***
No of properties, *N*	100	110	100
Length of mains, *M*	100	100	90
Density (*N/M*), *D*	1	1.1	1.11
Total distn cost,*TCD*	100	1.04	93
Unit cost, *TCD/N*	1	0.95	0.93

Table AF.1: Evaluation of density economies if $\varepsilon_N = 0.4$ and $\varepsilon_A = 0.7$

This latter effect is rather clearer if (AF.1) is re-stated with property density, *D_M* (= *N/M*) in place of *M*. It then becomes:

$$TCD = \alpha_1 \frac{N}{D_M} + \alpha_2 N + \alpha_3 w \frac{N}{D_M}\sqrt{N} \qquad \ldots\ldots\ldots.. \qquad \text{(AF.8)}$$

The related elasticity of distribution cost with respect to density is:

$$\varepsilon_D = \frac{D_M}{TCD}.\frac{\partial(TCD)}{\partial D_M} = \frac{D_M}{TCD}\left(-\alpha_1 \frac{N}{D_M^2} - \alpha_3 w \frac{N}{D_M^2}\sqrt{N}\right)$$

$$= -\left[\frac{\alpha_1 N / D_M + \alpha_3 wN\sqrt{N} / D_M}{\alpha_1 N / D_M + \alpha_2 N + \alpha_3 wN\sqrt{N} / D_M}\right] = -\left[\frac{\alpha_1 + \alpha_3 w\sqrt{N}}{\alpha_1 + \alpha_2 D_M + \alpha_3 w\sqrt{N}}\right]$$

……….. (AF.9)

Which is clearly negative so that $-1 < \varepsilon_D < 0$.

Appendix G

ALTERNATIVE COST FUNCTION ESTIMATION FOR WATER DISTRIBUTION (WITH APPLICATION TO BWC ZONES)

a. Cost function derivation

Following the standard approach based on production theory, and assuming that the technical options for water distribution can be represented by a normal production function, cost minimisation (or profit maximisation) would lead to a cost function for water distribution having the general form (See **Chapter IV, section 1(c)**):

$$VCD = VC_D(DO, p_{LD}, \overline{K_D}, Z_D) \qquad \ldots\ldots\ldots\ldots \qquad (G.1)$$

Where *VCD* is the variable cost of water distribution, *DO* is a measure of distribution output, p_{LD} is a price for variable inputs, $\overline{K_D}$ is a measure of water distribution capital[138] and Z_D is a vector of control variables. Assuming that the variation between cases in p_{LD} is small, this term can be dropped (this assumption appears reasonable for BWC zones, questionable for companies reporting to Ofwat and very questionable for US utilities).

Drawing on the discussion in **Chapter III**, **sections 4** and **5**, *DO* can be expressed as a function of average consumption per property (*w*), number of properties (*N*) and a measure of the average dispersion of properties (φ).

$$DO \ = f(w, N, \phi) \qquad \ldots\ldots\ldots\ldots \qquad (G.2)$$

Hence, (G.1) becomes:

$$VCD = VC_D\left\{f(w, N, \phi), \overline{K_D}, Z_D\right\} \qquad \ldots\ldots\ldots\ldots \qquad (G.3)$$

The values for $\overline{K_D}$ and the control variables included in Z_D will be determined by the availability of data.

Because the cost function for distribution may be quite complex, there is a case for adopting a flexible form specification, such as the translog, provided the number of observations is sufficient to enable this to be done. If all the RHS variables of (G.3) are treated equally, this would lead to:

[138] Following Garcia & Thomas (2001), capital in this formulation is taken to be “quasi-fixed”.

$$\begin{aligned}\ln VCD &= \delta_0 + \delta_1 \ln w + \delta_2 \ln N + \delta_3 \ln\phi + \delta_4 \ln \overline{K_D} + \delta_5 \ln Z_D \\ &+ \frac{1}{2}\delta_6 (\ln w)^2 + \frac{1}{2}\delta_7 (\ln N)^2 + \frac{1}{2}\delta_8 (\ln\phi)^2 + \frac{1}{2}\delta_9 (\ln \overline{K_D})^2 + \frac{1}{2}\delta_{10} (\ln Z_D)^2 \\ &+ \delta_{11} \ln w.\ln N + \delta_{12} \ln w.\ln\phi + \delta_{13} \ln w.\ln \overline{K_D} + \delta_{14} \ln w.\ln Z_D \\ &+ \delta_{15} \ln N.\ln\phi + \delta_{16} \ln N.\ln \overline{K_D} + \delta_{17} \ln N.\ln Z_D \\ &+ \delta_{18} \ln\phi.\ln \overline{K_D} + \delta_{19} \ln\phi.\ln Z_D + \delta_{20} \ln \overline{K_D}.\ln Z_D\end{aligned}$$

………… (G.4)

A simpler approach is to adopt a translog specification for (G.2), i.e:

$$\begin{aligned}\ln DO &= \alpha_0 + \alpha_1 \ln w + \alpha_2 \ln N + \alpha_3 \ln\phi + \frac{1}{2}\alpha_4 (\ln w)^2 + \frac{1}{2}\alpha_5 (\ln N)^2 + \frac{1}{2}\alpha_6 (\ln\phi)^2 \\ &+ \alpha_7 \ln w.\ln N + \alpha_8 \ln w.\ln\phi + \alpha_9 \ln N.\ln\phi\end{aligned}$$

…………. (G.5)

And then substitute this into a generalised Cobb-Douglas specification of (G.3), giving:

$$\begin{aligned}\ln VCD &= \beta_0 + \beta_1 (\alpha_0 + \alpha_1 \ln w + \alpha_2 \ln N + \alpha_3 \ln\phi + \frac{1}{2}\alpha_4 (\ln w)^2 + \frac{1}{2}\alpha_5 (\ln N)^2 \\ &+ \frac{1}{2}\alpha_6 (\ln\phi)^2 + \alpha_7 \ln w.\ln N + \alpha_8 \ln w.\ln\phi + \alpha_9 \ln N.\ln\phi) + \beta_2 \ln \overline{K_D} + \beta_3 \ln Z_D\end{aligned}$$

……….. (G.6)

which can be expressed as:

$$\begin{aligned}\ln VCD &= \gamma_0 + \gamma_1 \ln w + \gamma_2 \ln N + \gamma_3 \ln\phi + \gamma_4 \ln \overline{K_D} + \gamma_5 \ln Z_D + \gamma_6 (\ln w)^2 + \gamma_7 (\ln N)^2 \\ &+ \gamma_8 (\ln\phi)^2 + \gamma_9 \ln w.\ln N + \gamma_{10} \ln w.\ln\phi + \gamma_{11} \ln N.\ln\phi)\end{aligned}$$

…………. (G.7)

Compared with (G.4), this eliminates a large number of second order terms while still providing a reasonable degree of flexibility in the relationship.

From the specification (G.7), expressions for certain short and long term distribution scale elasticities can be derived. However, because *N* and *φ* are not independent, being linked through *λ* and *R*, the elasticities that can be derived are rather restricted in scope (this issue is discussed more fully in **Chapter V, section 2(c)**). Thus ε_N measures the response of costs to changes in numbers of properties *with φ constant* (a form of densification), while ε_φ measures the response of costs to changes in average distance to properties *with N constant* (a form of dispersion). With this limitation in mind, the elasticities derivable from (G.7) are:

Short term elasticities

$$\text{(a)}\quad \varepsilon_W^S = \frac{\partial(\ln VCD)}{\partial(\ln w)} = \gamma_1 + 2\gamma_6 \ln w + \gamma_9 \ln N + \gamma_{10} \ln \phi \qquad \ldots\ldots \quad \text{(G.8)}$$

$$\text{(b)}\quad \varepsilon_N^S = \frac{\partial(\ln VCD)}{\partial(\ln N)} = \gamma_2 + 2\gamma_7 \ln N + \gamma_9 \ln w + \gamma_{11} \ln \phi \qquad \ldots\ldots \quad \text{(G.9)}$$

$$\text{(c)}\quad \varepsilon_\phi^S = \frac{\partial(\ln VCD)}{\partial(\ln \varphi)} = \gamma_3 + 2\gamma_8 \ln \phi + \gamma_{10} \ln w + \gamma_{11} \ln N \qquad \ldots\ldots \quad \text{(G.10)}$$

Long term elasticities

$$\text{(d)}\ \varepsilon_W^L = \frac{\varepsilon_W^S}{1-\varepsilon_K} \qquad \ldots\ldots \quad \text{(G.11)}$$

$$\text{(e)}\ \varepsilon_N^L = \frac{\varepsilon_N^S}{1-\varepsilon_K} \qquad \ldots\ldots \quad \text{(G.12)}$$

$$\text{(f)}\ \varepsilon_\phi^L = \frac{\varepsilon_\phi^S}{1-\varepsilon_K} \qquad \ldots\ldots \quad \text{(G.13)}$$

Where

$$\varepsilon_K = \frac{\partial(\ln VCD)}{\partial(\ln \overline{K_D})} = \gamma_5 \qquad \ldots\ldots \quad \text{(G.14)}$$

b. Application to data for BWC zones

The data assembled for BWC's 184 water quality zones as described in **Appendix H** go a long way towards providing sufficient information to estimate (G.7). There are however a couple of further points to consider first.

Direct information on $\overline{K_D}$ is lacking but length of mains (*M*) provides a good proxy. As for control variables, information on differences in geographical conditions (such as the effect of topography on pumping head) is not available but the proportion of urban land (*UAP*) in 178 zones has been obtained and would seem worth testing (at the expense of dropping the 6 cases for which *UAP* was not available).

With these adjustments, (G.7) then becomes:

$$\ln VCD = \gamma_0 + \gamma_1 \ln w + \gamma_2 \ln N + \gamma_3 \ln \phi + \gamma_4 \ln M + \gamma_{56} \ln(1+UAP) + \gamma_6 (\ln w)^2 + \gamma_7 (\ln N)^2$$
$$+ \gamma_8 (\ln \phi)^2 + \gamma_9 \ln w.\ln N + \gamma_{10} \ln w.\ln \phi + \gamma_{11} \ln N.\ln \phi)$$

$$\ldots\ldots \qquad \text{(G.15)}$$

The results obtained using this specification are shown in the first column of **Table G.1** and are clearly not very satisfactory. The key coefficients are not significant, and some of the values and signs look implausible. The consequential elasticity estimates presented in **Table G.2** below confirm this impression. The negative elasticity for the dispersion variable (φ) is not tenable and the long term elasticities, driven by the high estimated value for ε_K, are not acceptable either. As the second column of this table shows, a simpler specification omitting second order variables produced more plausible results with little change in R^2. One possibility here is that the true relationship between the variables concerned is linear rather than log linear because about one third of *VCD* value has had to be estimated by allocation; another possibility is that interaction between *M* and φ is present (dropping the term in ln*M* reversed the sign on lnφ).

	Using (G.15)		**Omitting second order variables**	
Variable	**Coefficient**	**S.E.**	**Coefficient**	**S.E.**
lnw (γ_1)	4.868	*5.16*	0.327	*0.111*
lnN (γ_2)	4.237	*3.58*	0.640	*0.200*
lnφ (γ_3)	-11.52	*7.93*	-0.235	*0.941*
lnM (γ_5)	0.948	*0.381*	0.408	*0.282*
Ln(1+UA)(γ_6)	0.044	*0.047*	0.074	*0.042*
(lnw)^2 (γ_7)	-0.385	*0.323*	-	-
(lnN)^2 (γ_8)	-0.442	*0.234*	-	-
(lnφ)^2 (γ_9)	-2.832	*1.12*	-	-
Lnw.lnN (γ_{10})	-0.039	*0.341*	-	-
Lnw.lnφ (γ_{11})	0.237	*0.770*	-	-
lnN.lnφ (γ_{12})	2.267	*1.01*	-	-
R^2 (d.f.)	0.9329 (166)		0.9283 (172)	

Table G.1: Regression results for 178 BWC distribution zones, using (G.15) with and without second order variables

To check further on the plausibility of these results, the various elasticities derived at (G.8) to (G.14) were calculated, using mean values for *w* (430.6 litres/property/day), *N* (17,849 properties) and φ (11.355 x 100 metres). The resulting estimated elasticities are shown in **Table G.2**:

Calculated elasticity	**Using (G.15)**	**Omitting second order variables**
ε_W^S	0.392	0.327
ε_N^S	0.854	0.640
ε_φ^S	-1.65	-0.235
ε_K	0.948	0.408
ε_W^L	7.54	0.552
ε_N^L	16.43	1.08
ε_φ^L	-31.75	-2.79

Table G.2: Calculated elasticities (G.8) – (G.14)

As already noted, many of these values are not very plausible, particularly the long term elasticities. However, ε_W^S looks acceptable, and the restricted scope of ε_N and ε_φ should be kept in mind.

b. An alternative approach

In view of the evidence above for a more linear relationship, this avenue needs also to be explored. In **Appendix F**, it was found that in the constant density version of the square settlement model, distribution costs could be represented as:

$$TCD = \alpha_1 M + \alpha_2 N + \alpha_3 w.M\sqrt{N} \qquad \ldots\ldots\ldots\ldots \qquad (G.15)$$

A similar expression emerged when a lower density suburb was added to the square settlement. This suggests that distribution costs can be modeled as a linear combination of terms to pick up the separate effects of length of mains, numbers of properties and pumping costs. The RHS variable in (G.15) is *TCD* (total distribution costs) but the derivation in **Appendix F** indicates that a similar relationship should hold for *VCD* (distribution operating costs). To test whether this is the case, the specification (G.16) below was run.

$$VCD = \alpha_0 + \alpha_1 M + \alpha_2 N + \alpha_3 PMP + \alpha_4 UA \qquad \ldots\ldots\ldots \qquad (G.16)$$

Where *M* is length of mains (km), *N* is number of properties, *PMP* is a composite measure intended to capture pumping costs ($PMP = (w+l)M\sqrt{N}$, where *w* is average water usage and *l* average leakage per property (litres/property/day)) and *UA* is the area of urban land in each zone (in hectares).

The results obtained using (G.16) are shown in **Table G.3**:

Variable	**Using VCD**	
	Coefft	**S.E.**
M (α_1)	-268.9	182.4
N (α_2)	9.73***	0.971
PMP (α_3)	0.007***	0.0016
UA (α_4)	-5.61	23.1
R^2/d.f.	0.8447 (173)	

Table G.3: Regression results for 178 BWC zones, using (G.16)

These results suggest a strong association between distribution costs and number of properties, while the coefficient on pumping costs although numerically small is also highly significant. The negative coefficients on length of mains and urban area look puzzling although not significant (the former perhaps because its effect has been absorbed in the pumping costs variable). To test for non-linearity, terms in N^2 and M^2 were also tried but found not to be significant.

The distribution cost data provided by BWC is for operating costs but the missing capital costs can be estimated by taking distribution capital maintenance (*CMD*) plus return on capital (*CCD*) for BWC as a whole and allocating this total to zones in proportion to length of mains in each zone. Running (G.15) with total distribution costs (*TCD*) so calculated simply loads the extra costs onto the *M* coefficient (which then becomes strongly and significantly positive), leaving the other coefficients unchanged. While this is consistent with the general observation that capital costs are the dominant element in infrastructure costs[139], the regression adds no new information given the way this part of the data has been constructed.

d. Conclusions

In view of the problems encountered with the approaches described in this appendix, it was decided to adopt instead the rather different approach set out in **Chapter V** of the main text.

[139] For example, Speir & Stephenson (2002) in their comparison of the effects of different patterns of housing development on the costs of providing water and sewerage note that “On average, … water distribution and sewer collector mains within the development tracts make up 78% of costs across all scenarios … [and] … water distribution makes up a much greater proportion than wastewater collection.”

Appendix H

PROCESSING THE BWC DISTRIBUTION SYSTEM DATA

The basic unit of BWC's distribution side is the District Meter Area (DMA). There are over 3000 of these. Within each county, DMAs are grouped together in Water Quality Zones (WQZ), of which there are nearly 200 altogether. The company is able to identify which treatment works or boreholes supply water to each zone, which is important for water quality control or in case of an interruption in supply. Most WQZs obtain their water from several different sources – for security of supply as well as water quality reasons – but from the distribution point of view they constitute reasonably coherent units although they are not self-contained distribution systems. Initially therefore, WQZs were taken as the unit of analysis. For further analysis, some WQZs were then grouped together to correspond better to urban areas.

a. Data sources

Four sources of data have been used to produce the information required for the BWC distribution analysis:

a. **Leakage monitoring system** data. The raw data provide weekly readings for some 3000 DMAs. This is the source for information on DFT (daily flow totals) and leakage (LKG); it also includes a count of property numbers enabling W (water supplied per property, in litres/prop/day) and *w* (water used per property) to be estimated; and km of mains can be calculated by dividing 'leakage (m)' by 'leakage (l/km)'.

b. Information on **direct costs of distribution (excl. power)** by DMA for 2002/03 (sheet 1 of an XL file), together with a full list of DMAs with property counts for that year (sheet 2 of the same file).

c. **BWC's June Return** for 2002/03 which shows distribution costs in that year to be made up of Direct costs (excl. power) £29.0m, Power £10.9m and General & support £14.6m – making £54.5m in all.

d. The area (in sq. metres) of each DMA was obtained from **ArcGIS files** provided by the company (the total area so obtained agreed closely with the area reported by Water UK for this company). WQZ areas were then obtained by dissolving DMA boundaries into WQZ boundaries and calculating the resulting polygon areas using ArcGIS. This value divided by 10,000 to convert from square metres, gives the area (*A*) in hectares for each WQZ. To obtain the variable *UP* measuring the proportion of urban land in each WQZ (to use as a control variable), the same boundaries were then applied to carry out an OS "Strategi" land use analysis in ArcMap to determine the proportion of urban land in each WQZ. (For this purpose, the inner and outer limit features[140] for both large and small urban areas from Strategi were plotted onto a map of the WQZ boundaries, enabling the urban proportion of each WQZ to be calculated following rasterisation of the resulting polygons[141].) This analysis yielded an average figure of 11.4% for the proportion of urbanized land[142], a figure that varied between 54% for the most urbanized county to 4.8% for the most rural.

b. Processing leakage monitoring system data

In connection with its leakage control activities, the company logs flows into each DMA with flow rates being recorded at 15 minute intervals. Leakage is estimated by subtracting night use by measured customers and an estimate of night use by unmetered households (default = 2 litres/hour/household) from the minimum hourly night time flow. This hourly rate is grossed up to a daily rate by a "pressure adjustment factor", the value of which is specific to each area and varies between about 17 and 24 to reflect the effect of lower daytime pressures in moderating leakage rates. In principle, the following relationship between estimated household consumption and the other quantities should hold:

$$EHC = DFT - DFL - MET - LKG \quad \ldots\ldots\ldots\ldots \quad (H.1)$$

Where:

- EHC = Estimated household consumption (m^3/day) [143]
- DFT = Daily total flow (m^3/day)
- DFL = Daily flow logged customers (m^3/day)
- MET = Measured non-household daily use (m^3/day)
- LKG = Estimated leakage (m^3/day) – obtained as described above.

[140] "The line features which form the limits of an area are given feature codes indicating whether they are the **outer limit** or **inner limit** of such a classified area. For example, the outer limit of large urban areas are bounded by a line which has a Feature Code of 5420; "islands" within this outer limit which are not classed as part of the urban area are bounded by a line with a Feature Code of 5492, representing the inner limit." Strategi Guidance Notes.

[141] I would like to acknowledge here the assistance of a fellow student, Alejandra Castrodad-Rodriguez, in carrying out this analysis.

[142] The difference between this figure and the 8.6% obtained using ONS data is presumably due, at least in part, to the inclusion of urban areas with population <5k.

[143] 1 m^3 = 1,000,000 cc = 1,000 litres; so 1 Ml/day = 1,000 m^3/day.

The raw data from this system consisted of 52 weeks observations in 24 columns for each DMA for 2004/05 (?). Initially, some 185,000 lines of data were found on the disk. To render this enormous volume of information manageable, some winnowing down was clearly needed. First, the data was loaded onto 3 sheets of an Excel workbook and the first sheet containing 18 weeks data (the weeks with start dates between 1st and 10th of the month for all months of the year) was selected to provide the basis for analysis[144]. This first sheet had at this stage some 65,000 lines. However, some 17,000 lines relating to some 500 temporary and invalid DMAs could also be removed and a further 3,000 lines could not be used because no daily flow total had been recorded for one reason or another. The remaining observations were then used to obtain an average weekly value for each of the valid DMAs left; zone codes were added for each DMA and the average values were then summed to give a total for each of 182 zones. The key quantities so obtained are summarized in **Table H.1** below:

Data item	**Abbreviation**	**Units**
Connected properties (all)	PROPS	Numbers
Connected households	HHLDS	Numbers
Leakage	LKG	m^3/day
Daily flow total	DFT	m^3/day
Daily flow logged customers	DFL	m^3/day
Daily flow metered customers	MET	m^3/day
Estimated household consumption	EHC	m^3/day
Length of mains	M	km

Table H.1: Key data items obtained from BWC's leakage monitoring system

Examination of the data indicated that a number of observations at DMA level were problematic. Some leakage figures were negative, as were some estimated household consumption figures; others were implausibly high (a possible explanation here could be mains bursts, losses from which cannot be separately distinguished). In very few cases did the relationship (H.1) between EHC and the other quantities hold exactly – in fact, in most cases there was a sizeable discrepancy, in some cases very large indeed. However, it appeared in general that the values for DFT and LKG were quite plausible, particularly when aggregated to zone level. In consequence, it was decided to disregard

[144] During the data processing, one week's data for about half the DMAs was accidentally deleted so that for these DMAs the average values are based on 17 rather than 18 weeks data. For a similar reason, the data for DFL has been taken from the second set of weekly data (17 weeks beginning 10th -19th of each month).

the EHC figures, using DFT – LKG as measure of consumption (incl. measured consumption) when required.

c. Processing direct costs data

The steps were

a. Match costs from sheet 1 to list of DMAs and property counts on sheet 2.

b. Add data from leakage monitoring system, so that matching DMAs are aligned. At this stage, there were 3580 lines of data, many incomplete.

c. Delete DMAs not in leakage monitoring system, and which have <100 properties and no recorded costs (440 cases, 3144 properties); delete DMAs from sheet 1 which are not in sheet 2 (183 cases, £897k of costs); delete DMAs for which zone is not identifiable (171 cases, 40,433 properties, £1,632k costs).

d. To fill gaps in c.150 of remaining 2786 cases: (i) where leakage system data missing, use property count from costs data to calculate DFT and leakage assuming average *W* and *w* for relevant zone; (ii) calculate km mains using property count and average km/property for relevant zone.

d. Additional processing

Comparison of the totals for these 2786 DMAs with JR figures gives:

Item	**JR figure**	**Total of 2786 DMAs**	**Difference (%)**
Direct costs (excl. power)	£29.0m	£26.4m	-9.0%
Properties	3,279,000	3,284,202	+0.2%
Water supplied	1958 Ml/day	1818 Ml/day	-7.2%
Km mains	45,674km	39,087km	-14.4%
Area	19,745 sq km[145]	19,124 sq km	-3.1%

In general, it seems likely that these differences can largely be attributed to industrial supplies. Where these appear in the leakage monitoring system, the records are often incomplete and inconsistent. Establishing precisely what is going on here would be very time-consuming, it was therefore decided to omit the missing amounts from the analysis, on the argument that the remaining information should give a reasonable assessment of distribution costs for non-industrial supplies.

[145] However, Water UK gives BWC a gross area of 21,650 sq km incl a WOC of 1,507 sq km, a net area of 20,143 sq km.

It remained to allocate costs of power, general support and capital maintenance to zones. In the Ofwat econometric models, power costs are closely related to volumes supplied. Power costs are therefore allocated in proportion to DFT. For lack of a better basis, general and support costs are allocated in proportion to numbers of properties. Capital maintenance is allocated in proportion to length of mains. In all three cases, allowance is made for the amounts attributable to the omitted mainly industrial supplies: thus the amount of power costs allocated to the identified DMAs is £10.9m x 1818/1598 = £10.122m; the amount of general and support costs allocated is £14.6m x 26.4/29.0 = £13.310m; and the amount of capital maintenance allocated is £76.6m x 39,087/45,674 = £65.553m.

URBAN AREAS: AREA/LAMBDA/DENSITY TABLE

Radius	Area	Do	λ =	λ =	λ =	λ =	λ =	λ =	λ =	λ =	λ =	λ =
100m	Ha	Prop/Ha	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.1
5	78.53982	30	29.02	28.07	27.16	26.28	25.44	24.62	23.84	23.08	22.35	21.65
6	113.0973	30	28.83	27.70	26.63	25.61	24.62	23.69	22.79	21.93	21.10	20.32
7	153.938	30	28.64	27.34	26.11	24.95	23.84	22.79	21.79	20.84	19.94	19.08
8	201.0619	30	28.45	26.98	25.61	24.31	23.08	21.93	20.84	19.81	18.84	17.93
9	254.469	30	28.26	26.63	25.11	23.69	22.35	21.10	19.94	18.84	17.81	16.85
10	314.1593	30	28.07	26.28	24.62	23.08	21.65	20.32	19.08	17.93	16.85	15.85
11	380.1327	30	27.89	25.94	24.15	22.50	20.97	19.56	18.26	17.06	15.95	14.92
12	452.3893	30	27.70	25.61	23.69	21.93	20.32	18.84	17.49	16.25	15.11	14.06
13	530.9292	30	27.52	25.27	23.23	21.37	19.69	18.15	16.75	15.47	14.31	13.25
14	615.7522	30	27.34	24.95	22.79	20.84	19.08	17.49	16.05	14.75	13.57	12.49
15	706.8583	30	27.16	24.62	22.35	20.32	18.49	16.85	15.38	14.06	12.87	11.79
16	804.2477	30	26.98	24.31	21.93	19.81	17.93	16.25	14.75	13.41	12.21	11.13
17	907.9203	30	26.81	23.99	21.51	19.32	17.38	15.66	14.14	12.79	11.59	10.52
18	1017.876	30	26.63	23.69	21.10	18.84	16.85	15.11	13.57	12.21	11.01	9.95
19	1134.115	30	26.46	23.38	20.71	18.38	16.34	14.57	13.02	11.66	10.46	9.41
20	1256.637	30	26.28	23.08	20.32	17.93	15.85	14.06	12.49	11.13	9.95	8.91
21	1385.442	30	26.11	22.79	19.94	17.49	15.38	13.57	12.00	10.64	9.46	8.44
22	1520.531	30	25.94	22.50	19.56	17.06	14.92	13.09	11.52	10.17	9.01	8.00
23	1661.903	30	25.77	22.21	19.20	16.65	14.48	12.64	11.07	9.73	8.58	7.59
24	1809.557	30	25.61	21.93	18.84	16.25	14.06	12.21	10.64	9.31	8.17	7.20
25	1963.495	30	25.44	21.65	18.49	15.85	13.65	11.79	10.23	8.91	7.79	6.84
26	2123.717	30	25.27	21.37	18.15	15.47	13.25	11.39	9.84	8.53	7.43	6.50
27	2290.221	30	25.11	21.10	17.81	15.11	12.87	11.01	9.46	8.17	7.09	6.18
28	2463.009	30	24.95	20.84	17.49	14.75	12.49	10.64	9.11	7.83	6.77	5.88
29	2642.079	30	24.78	20.58	17.17	14.40	12.14	10.29	8.77	7.51	6.47	5.60
30	2827.433	30	24.62	20.32	16.85	14.06	11.79	9.95	8.44	7.20	6.18	5.34
31	3019.071	30	24.46	20.06	16.55	13.73	11.46	9.62	8.13	6.91	5.91	5.09
32	3216.991	30	24.31	19.81	16.25	13.41	11.13	9.31	7.83	6.64	5.66	4.86
33	3421.194	30	24.15	19.56	15.95	13.09	10.82	9.01	7.55	6.37	5.42	4.64
34	3631.681	30	23.99	19.32	15.66	12.79	10.52	8.72	7.28	6.12	5.19	4.43
35	3848.451	30	23.84	19.08	15.38	12.49	10.23	8.44	7.02	5.88	4.97	4.23

Appendix I: Average density of a circular settlement with radius *R* whose density declines at a rate λ from the centre where density is 30 properties/Ha (Extracted from full table, approx 4 times as large)

BIBLIOGRAPHY

References

AWWA (American Water Works Association) *1996 Water Stats Survey, USA* (from CD-ROM produced by AWWA)

Arnott R J (1979) “Optimal city size in a spatial economy” *Journal of Urban Economics* 6:65-89

Bairoch P (1988) *Cities and Economic Development: From the dawn of History to the Present*, translated by C. Braider, University of Chicago Press

Bartik T J and V Kerry Smith (1987) “Urban Amenities and Public Policy” in ES Mills (Ed) *Handbook of Urban and Regional Economics, Vol 2* North-Holland

Basso L J and S R Jara-Diaz (2006) “From economies of density and network scale to multioutput economies of scale and scope: A synthesis” Paper presented to European Transport Conference 2006, Strasbourg

Beckmann M J (1988) “An economic model of urban growth” in Ausubel JH and R Herman (1988) *Cities and their vital systems: Infrastructure, past, present and future*, Washington, National Academy Press

Bertaud A and S Malpezzi (1998) *The spatial distribution of population in 35 world cities,* Working paper. World Bank and Center for Urban Land Economics Research, University of Wisconsin.

Biehl D (1986) *The Contribution of Infrastructure to Regional Development* Report to EC Commission

Biehl D (1991) “The role of infrastructure in regional development” in Vickerman RW *Infrastructure and Regional Development*, London, Pion

Bos H C (1965) *Spatial Dispersion of Economic Activity* Rotterdam University Press

Bradford D, R Malt and W Oates (1969)”The rising cost of local public services: Some evidence and reflections” *National Tax Journal* 22:185-202

Bromwich J (1996) *The Roman Remains of Southern France: A guidebook* (paperback), Routledge, London

Brueckner J K, J-F Thisse and Y Zenou (1999) “Why is central Paris rich and downtown Detroit poor?” *European Economic Review* 43: 91-107

Buchanan J M (1965) “An economic theory of clubs” *Economica* 33:1-14

Burchfield M, H G Overman and D Puga (2006) “The causes of sprawl: A portrait from space” *The Quarterly Journal of Economics* 121(2): 587-633

Byatt I, T Balance and S Reid (2006) “Regulation of water and sewerage services” in Crew M A and D Parker (Eds) *International Handbook on Economic Regulation*, Edward Elgar

Cairncross F (1997) *The death of distance: How the communications revolution will change our lives* London: Orion

Carruthers J I and G F Ulfarsson (2003) “Urban sprawl and the cost of public services” *Environment and Planning B: Planning and design* 30: 503-522

Caves D W, L R Christensen and M W Tretheway (1984) “Economies of density versus Economies of scale: Why trunk and local airline costs differ” *Rand Journal of Economics* 15:471-89

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

Cheshire P and S Shepherd (1995) "On the price of land and the value of amenities" *Economica 62: 247-67*

Christensen L R and W H Greene (1976) "Economies of scale in US electric power generation" *Journal of Political Economy,* 84:655-676

Ciccone A and R E Hall (1996) "Productivity and the Density of Economic Activity" *American Economic Review* 86:54-70

Clark R M and R G Stevie (1981) "A water supply cost model incorporating spatial variables" *Land Economics,* 57:18-32

Deloitte, Haskins & Sells (1990) *Water industry: Comparative efficiency review for Department of the Environment and Welsh Office*

DiPasquale D and W C Wheaton (1996) *Urban Economics and Real Estate Markets* Prentice Hall

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

Dupuit A (1844) "On the the measurement of the utility of public works", translated from the French, in *International Economic Papers, No 2* (London, 1952)

Downing P B and R D Gusteley (1995) "The public service costs of alternative development patterns: A review of the evidence" in Downing P B (Ed) *Local service pricing policies and their effect on urban spatial structure* pp. 63-86, University of British Columbia Press

Duranton G and D Puga (2004) "Micro-foundations of urban agglomeration economies" in Henderson J V and J-F Thisse (Eds) *Handbook of Regional and Urban Economics, Vol 4* Elsevier North-Holland

Eberts R W & D P McMillen (1999) "Agglomeration economies and urban public infrastructure" in Cheshire P & E S Mills (Eds) *Handbook of Regional and Urban Economics, Vol 3* Elsevier North-Holland

Elis-Williams D G (1987) "The effect of spatial population distribution on the cost of delivering local services" *Journal of the Royal Statistical Society A*, 150(2): 152-166

Environment Agency (2007) *Hidden Infrastructure*

Frank J E (1989) *The costs of alternative development patterns: A review of the literature* Urban Land Institute, Washington DC

Fujita M (1989) *Urban Economics* Cambridge University Press

Fujita M, Krugman P and A J Venables (1999) *The Spatial Economy: Cities, Regions, and International Trade* The MIT Press

Fujita M and J-F Thisse (2002) *Economics of Agglomeration: Cities, industrial location and regional growth* Cambridge University Press

Garcia S and A Thomas (2001) "The structure of municipal water supply costs: Application to a panel of French local communities" *Journal of Productivity Analysis,* 16:5-29

Glaeser E L and J D Gottlieb (2006) "Urban resurgence and the Consumer City" *Urban Studies* 43(8): 1275-99

Glaister S (1996) *Incentives in Natural Monopoly: The case of water* Research Papers in Environmental and Spatial Analysis No. 30, London School of Economics

Gramlich E M (1994) “Infrastructure investment: A review essay” in *Journal of Economic Literature*, XXXII: 1176-96.

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

Grigg N S (1986) *Urban Water Infrastructure: Planning, Management and Operations* John Wiley & Sons

Gyourko J and J Tracy (1991) “The structure of local public finance and the quality of life” *Journal of Political Economy* 99:774-806

Hamer M (2006) “Every home should have one” *New Scientist*, 21 January 2006

Hansen P, D Peeters and J-F Thisse (1983) “Public facility location models: A selective survey” in Thisse J-F and H G Zoller *Locational Analysis of Public Facilities* North-Holland

Henderson J V and J-F Thisse (Eds) (2004) *Handbook of Regional and Urban Economics, Vol 4* Elsevier North-Holland

Herman H and JH Ausubel (1988) “Cities and infrastructure: Synthesis and perspectives” in Ausubel JH and R Herman (1988) *Cities and their vital systems: Infrastructure, past, present and future,* Washington, National Academy Press

Kim H Y (1985) “Economic modeling of water supply: An econometric analysis of the multiproduct firm” Project Report, USEPA, Cincinnati, Ohio.

Kim H Y and R M Clark (1988) “Economies of scale and scope in water supply” *Regional Science and Urban Economics*, 18:479-502

Knox Lovell CA and P Schmidt (1988) “A comparison of alternative approaches to the measurement of productive efficiency” in Dogramaci A and R Fare *Applications of Modern Production Theory*, Kluwer Academic Publishers

Krugman P (1991) “Increasing returns and economic geography” *Journal of Political Economy* 99:483-99

Ladd H F (1992) “Population growth, density and the cost of providing public services” *Urban Studies* 29: 273-295

Lea A C (1979) “Welfare theory, public goods, and public facility location” *Geographical Analysis* 11:217-39

Li F, D Tolley, N P Padhy and J Wang (2005) *Network benefits from introducing an economic methodology for distribution charging* , A study for Ofgem by the Dept of Electronic & Electrical Engineering, University of Bath

Love R F, J G Morris and G O Wesolowsky (1988) *Facilities Location: Models and Methods* North-Holland

Lundberg M and L Squire (2003) “The simultaneous evolution of growth and inequality” *Economic Journal* 113 (April): 305-25

Mays LW (2002) “Urban water infrastructure: A historical perspective” in Mays LW (Ed) *Urban Water Supply Handbook* McGraw-Hill

McCann P and D Shefer (2004) “Location, agglomeration and infrastructure” *Papers in Regional Science* 83:177-196

McDonald J F (1997) *Fundamentals of Urban Economics* Prentice Hall

Mills E S (Ed) (1987) *Handbook of Regional and Urban Economics, Vol 2* Elsevier North-Holland

Mills E S and P Cheshire (Eds) (1999) *Handbook of Regional and Urban Economics, Vol 3* Elsevier North-Holland

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

Newman P, J Kenworthy and F B Laube (1999) *An International Sourcebook of Automobile Dependence in Cities, 1960-1990* University Press of Colorado

Nijkamp P (Ed) (1986) *Handbook of Regional and Urban Economics, Vol 1* Elsevier North-Holland

Ofgem (2004) *Electricity Distribution Price Control Review: Final Proposals* (Ref 265/04)

Ofgem (2005) *Enduring transmission charging arrangements for distributed generation: A discussion document* (available on the Ofgem website)

Ofwat (2002, revised 2003) *Future approaches to leakage target setting for water companies in England & Wales*

Ofwat (2003a) *2003 June Returns for the Water Industry in England & Wales* (CD-ROM available from Office of Water Services)

Ofwat (2003b) *Tariff structure and charges, 2002-2003 report*, Office of Water Services, Birmingham

Ofwat (2003c) *Financial performance and expenditure of the water companies in England & Wales, 2002-2003 report*, Office of Water Services, Birmingham

Ofwat (2004) *Water and sewerage service unit costs and relative efficiency, 2002-2003 report*, Office of Water Services, Birmingham

ONS (2001) *Census 2001: Census area statistics for urban areas in England & Wales* (CD-ROM available on request from ONS)

ONS (2004) *2001 Census: Key statistics for urban areas in England & Wales*, HMSO

OXERA (1996) *Infrastructure in the UK: Public projects and private money*

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

Richardson H J and C-H C Bae (2004) "Transportation and Urban Compactness" in Hensher D A, K J Button, K E Haynes and P R Stopher (Eds) *Handbook of Transport Geography and Spatial Systems* Elsevier.

Rushton A, J Oxley and P Croucher (2000 Edition) *The Handbook of Logistics and Distribution Management*, Kegan Page Ltd

Saal D S 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.*

Saal D S, D Parker and T Weyman-Jones (2004) "Determining the contribution of technical, efficiency and scale change to productivity growth in the privatized English and Welsh water and sewerage industry" *Aston Business School Research Paper RP0433*.

Samuelson PA (1954) "The pure theory of public expenditure" *Review of Economics and Statistics* 36:387-9

Sayers H M (1938) *The Economic Principles of Electrical Distribution* (London, Pitman)

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

Shepherd E S (1980) “Location and the demand for travel” *Geographical Analysis* 12:111-28

Shy O (2001) *The Economics of Network Industries* Cambridge University Press

Sole-Olle A and M H Rico (2008) *Does urban sprawl increase the cost of providing local public services? Evidence from Spanish municipalities* Document de Treball 2008/6, Institut d’Economia de Barcelona

Speir C and K Stephenson (2002) “Does sprawl cost us all?” *Journal of the American Planning Association* Vol 68(1): 56-70

Starrett D A (1988) *Foundations of Public Economics* Cambridge University Press

Stone & Webster Consultants (2004) *Investigation into evidence for economies of scale in the water industry in England & Wales* (Final report to Ofwat)

Strategic Management Consultants (2002) *Optimum entity size in the water industry of England and Wales: A review of factors which influence the size of companies* Report to Ofwat

The Economist (2006) “Why hospitals must be allowed to fail” 11th March, p.11

Thisse J-F and H G Zoller (1983) “Some notes on public facility location” in Thisse J-F and HG Zoller (Eds) *Locational Analysis of Public Facilities* North-Holland

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

Tiebout C M (1961) “An economic theory of fiscal decentralization” in NBER *Public Finances: Needs, Sources and Utilisation* Princeton University Press

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

Twort A C, D D Ratnayaka and M J Branch (2000) *Water Supply (5th edition)* Arnold/IWA Publicity

Water UK (2004) *Capital maintenance: Building on our inheritance* Report issued under cover of Press Notice dated 17 June 2004.

Wenban-Smith H B (2000) *Urban Futures: Report of the Cities and Transport Group of the Chatham House Forum* Royal Institute of International Affairs