Economies of scale in the water sector: a survey of the empirical literature
Gustavo Ferro, Emilio J. Lentini and Augusto C. Mercadier

ABSTRACT
The issue of the most favorable size and optimal industry structure in the water sector is a relevant topic in many countries, due to fragmentation of the water sector and the key role played by municipalities. Important debates are taking place worldwide about how to provide universal access to the water supply and offer an efficient service. Regarding efficiency, the possibility of exploiting economies of scale would imply better resource allocation, the potential for lower water charges, and greater geographical coverage. By surveying the empirical research from different parts of the world, we aim to shed some light on the topic of economies of scale, and to provide a synthesis of the literature. We also aim to determine whether there is a tradeoff between centralization and decentralization. Our survey shows that, for several countries, variations in efficiency of water provision due to economies of scale do exist. Increases in efficiency related to economies of scale are found for populations in the range of 100,000 to 1 million people served. For larger populations, volume- or density-constant returns to scale are observed, followed by decreasing returns to scale; the reverse occurs for smaller values, suggesting that cost savings are derived from consolidation of providers. [Returns to scale refer to changes in output resulting where all inputs increase by a constant factor. If output increases by that same proportional change, then there are constant returns to scale. If output increases by less than that proportional change, there are decreasing returns to scale. If output increases by more than that proportional change, there are increasing returns to scale. Returns to scale is a technological phenomenon, due to the relationship between inputs and outputs in the production function. Economies of scale refer to reductions in unit cost as the scale of production increases. Diseconomies of scale are the opposite: increasing in unit costs as the scale of production increases. Scale changes mean a proportional increase in all factors of production. Economies of scale are an economical phenomenon, due to the relationship between unit costs and the level of production.]

Key words | decentralization, economies of scale, efficiency, water

ABBREVIATIONS AND ORGANIZATIONS

CONICET National Council of Scientific and Technical Research
CPI consumer price index
ECLAC Economic Comisión of Latin America and the Caribbean of the UN
ERAS Water and Sanitation Regulator of the Metropolitan Area Argentina
EVI economy of vertical integration
FCV-UBA Faculty of Veterinary, Universidad de Buenos Aires
GDR Grupo Internacional de Investigación ‘Agua, ciudad y territorio. Gobernabilidad y acceso al agua en las Américas’
Ofwat Water Services Regulation Authority in the UK

Gustavo Ferro (corresponding author)
Instituto de Economía UADE and CONICET,
Universidad Argentina de la Empresa and
Consejo Nacional de Investigaciones Científicas y Técnicas – National Council of Scientific and Technical Research,
Lima 717,
(1073) Buenos Aires,
Argentina
E-mail: gferro@uade.edu.ar

Emilio J. Lentini
Centro de Estudios Transdisciplinarios del Agua,
FCV-UBA, and GDR «Res-Eau-Ville»,
Centro de Estudios Transdisciplinarios del Agua,
Faculty of Veterinary-Universidad de Buenos Aires
and Grupo Internacional de Investigación ‘Agua, ciudad y territorio. Gobernabilidad y acceso al agua en las Américas’,
Av. Chorroarín 280,
(1427) Buenos Aires,
Argentina

Augusto C. Mercadier
ERAS and UNLP,
Ente Regulador de Agua y Saneamiento Water and Sanitation Regulatory Agency and
Universidad Nacional de La Plata,
Callao 982,
(1023) Buenos Aires,
Argentina

This paper surveys empirical evidence regarding the existence of economies of scale in the water sector, in an attempt to generate a useful synthesis to meet research and policy goals. Economies of scale and the optimal size of the water industry have become relevant topics in many countries, because the water sector is highly fragmented, with a key role still being played by municipalities.

Researchers and policy-makers worldwide have been discussing ways to provide more equitable access and more efficient provision. One issue related to efficiency is the possibility of taking advantage of economies of scale.

Studies on economies of scale in sanitation are scarce, but there are many studies devoted to economies of scope between both water and sanitation. [Economies of scope refer to lowering average cost for a firm in producing two or more products (instead of producing them separately in two or more firms). They are based on the common and recurrent use of proprietary know-how or on an indivisible physical asset.]

Certain features make the water sector unique and give prominence to the issue of economies of scale. First, the industry is capital-intensive and a significant portion of its capital is tied up long term and technological change has also been relatively slow as compared with other sectors. Accordingly, water providers have an economic incentive to extract monopoly rents but at the same time are vulnerable to political pressures to keep prices low, preventing adequate returns that would permit capital replacement and attract new investment.

Second, the water industry is vital to the population, and tariff levels have been very sensitive to political pressure, especially in developing countries; this implies difficulties in cost recovery, and consequently disincentives to the private sector.

Finally, as long as the water sector is a local service, it generally falls under municipal jurisdiction, but the optimal (efficient) scale of provision may not coincide with the size of some local communities.

The three aspects mentioned above have given rise to a wide range of industrial structures in the sector, which in turn have undergone a number of changes owing to the consolidation of highly atomized services, or the breakup of geographically concentrated providers, or changes in the property pattern (i.e. privatization).

Theoretical discussion of economies of scale, and empirical results regarding these economies in the water sector, are useful in determining the tradeoffs between the concentration or aglomeration and decentralization of the services.

In order to answer the questions of ‘what should be measured; and why, how and which results are produced’, the next section of this paper presents issues regarding measurements related to economies of scale in the water and sanitation sector. The following section reviews the literature, and then the results are discussed. Our conclusions are presented in the final section.

**WHAT ARE WE MEASURING?**

From the cost function \( c = f(q, \text{clients}, \text{area}) \), where \( c \) means total costs, \( q \) is output, ‘clients’ describes the amount of served connections, and ‘area’ is the serviced surface, we can define the output-density economies as the inverse of the percentage increase in costs resulting from a 1% increase in output:

\[
E_{OD} = \left( \frac{\partial \ln c}{\partial \ln q} \right)^{-1} = \frac{1}{\varepsilon_q}
\]

(1)

The formula measures the reaction of costs to increase in output, while holding the parameters of customers and serviced area as constants.

The economies of customer density are defined as the inverse of the percentage increase in costs resulting from a 1% increase both in the quantity of output and in the number of customers:

\[
E_{DC} = \left( \frac{\partial \ln c}{\partial \ln q} + \frac{\partial \ln c}{\partial \ln \text{clients}} \right)^{-1} = \frac{1}{\varepsilon_q + \varepsilon_{cl}}
\]

(2)
This measure helps to analyze the existence of serviced areas which become more densely populated over time.

Economies of scale measure the reaction of costs when the output, the number of customers and the area size increase proportionally (where customer density could be measured using the area in square kilometers, the length of the network, or the number of covered jurisdictions in the denominator and inhabitants or connections in the numerator).

We can assume that customer density and output per customer is given by:

\[
E_{\text{scale}} = \left( \frac{\partial \ln c}{\partial \ln q} + \frac{\partial \ln c}{\partial \ln \text{clients}} + \frac{\partial \ln c}{\partial \ln \text{area}} \right)^{-1}
\]

\[
= \frac{1}{\varepsilon_q + \varepsilon_{\text{cl}} + \varepsilon_{\text{area}}}
\]

(3)

A detailed discussion on these measures can be found in Filippini et al. (2008) and Bottasso & Conti (2009).

**A REVIEW OF THE LITERATURE**

Since the structure of the water sector is constrained by its institutional and historical framework and the environment in which it operates, the studies in this field have been motivated mainly by policy considerations. We can group the literature into five subsets of papers: (1) studies on the highly atomized US sector; (2) papers concerned with the English and Welsh providers, consolidated in the 1970s, privatized in the 1980s, and subject to mergers in the 1990s; (3) articles on the Italian experience, which, departing from a fragmented sector in 1994, concentrated its providers to achieve economies of scale, creating 'optimal territorial scope providers'; (4) recent analysis of economies of scale in other countries where the service is atomized, in order to assess whether consolidation is or is not recommended; and (5) cross-country analysis for comparative purposes, run by international organizations.

**The United States: a fragmented sector**

Hayes (1987) concentrated on the cost structure of a joint-product water utility, producing both retail and wholesale water. He found that the costs of the industry could be reduced if providers chose to merge when a threshold of 18.9 million cubic meters per year was not surpassed. However, the study did not control for customers, serviced area, quality, or any form of density.

In a similar fashion, Kim & Clark (1988) considered water utilities as multiproduct firms providing services to residential and non-residential customers. They found constant overall economies of scale for the average utility; remarkable economies of scale for the smaller utilities; and moderate diseconomies of scale for the larger ones. They used controls and highlighted the importance of operating variables such as capacity utilization and service distance in the determination of marginal costs for water supply.

Feigenbaum & Teeples (1985) found output-density economies. Similarly, Bhattacharyya et al. (1994) found economies of output density for both public firms and private ones. In a complementary paper, Bhattacharyya et al. (1995) concluded that economies of output density are strong for private providers and do not exist for public ones, and that there are economies of scale only for private enterprises.

Shih et al. (2004) found economies of scale at all cost levels and within individual components of the costs.

Wolff & Hallstein (2005) presented the economies of scale that the Lansing Michigan Board of Water and Light achieved in their central operations by combining retail contracts with other operations, wholesale contracts of water sales, and assets transferred from some municipalities to the Board in order to be managed jointly. Other functions, such as water distribution, remained in the hands of the municipalities. Another project in Minnesota allowed the municipalities of Saint Michael, Albertville and Hanover to profit from the economies of scale achieved by a private enterprise, which served the three localities.

García et al. (2007) challenged the idea that vertical disintegration of water utilities can result in cost-efficiency losses because the production stages are characterized by strong economies of vertical integration (EVIs). The EVIs may be the consequence of market imperfections and monopoly power at the upstream stages of the production process. They show that disintegration of the production and distribution stages may lead to cost savings (with the
exception of the smallest services). With regard to economies of scale, the authors found significant and important short-term returns to density for the average vertically integrated (VI) utilities, and stated that the average VI utility is characterized by increasing short-term and long-term returns to scale. An increase in the service size (i.e. production, customers and network) will result in a decrease in the average cost.

Torres & Morrison-Paul (2006) developed and implemented a cost-structure model of the US water-utility industry. Their estimates reveal significant economies of output density, particularly for small utilities that tend to have less output per customer and per area. However, these economies were counteracted by simultaneous increases in customers and serviced area, especially for large utilities. Moreover, the cost savings from higher output levels were insufficient to offset diseconomies from distributing water to more customers through larger service areas. Overall, the results indicated that consolidation of small utilities might be beneficial, depending on the concurrent expansion of the network, but the consolidation of already large utilities without corresponding increases in output density is unlikely to be cost effective. In other words, fragmentation, rather than consolidation, may be economically justifiable for some larger utilities, unless it is possible to realize cost efficiencies by increasing output for a given network or decreasing network size for a given output in order to gain economies of volume or output density.

England and Wales: privatization and mergers

Hunt & Lynk (1995) tested the existence of economies of scale and scope before the privatizations, and found significant evidence of economies of scope in the provision of water and sanitation. This study also constituted one of the first attempts to estimate economies of scale in England.

After the privatization process, Ashton (1999) examined economies of scale in ‘water-only companies’ (known as WoCs, in order to differentiate them from vertically integrated ‘water and sewerage companies’, WaSCs) in England and Wales during the period 1990–1996. The study found evidence of slight diseconomies of scale in the WoCs. Additionally, there are significant diseconomies in the use of the capital stock and low levels of capital capacity utilization. Ashton concluded that the mergers and acquisitions that had taken place were not driven by efficiency considerations.

Bottasso & Conti (2009), like Ashton (1999), tried to determine economies of scale and test the technological change during the period 1995–2005 in the WoCs. The estimates suggested the existence of economies of output and customer density that were not exploited, and few economies of scale, which increased along with the population density. Moderate cost savings could be expected from prudent mergers, with greater benefits in highly populated areas. The estimates established that even relatively large companies can obtain small economies of scale.

Saal & Parker (2000, 2001) analyzed the impact of privatization in 1989 and the improvement in water quality and environmental standards for WaSCs in April 1994. They found that the WaSCs experienced diseconomies of scale, as well as capital-augmenting/labor-saving technological change. Saal et al. (2007) estimated productivity growth-rates for the period 1985–2000. The decomposition of the origin of changes in productivity allowed them to conclude that the study provides strong evidence that the WaSCs are indeed characterized by decreasing returns to scale, as noted by Saal & Parker (2000) and Stone & Webster Consultants (2004). Furthermore, the latter paper quantified the resulting impact on productivity growth-rates caused by the continued increase in WaSC scale. They suggested that, both before and after privatization, decreasing returns to scale have had a consistent negative effect on productivity growth in the industry. For this reason, further consolidation is unlikely to achieve cost reductions through scale effects.

Bottasso & Conti (2003) concluded that efficiency gaps have been steadily narrowing since the 1994 price review. They found economies of output density, depending on the model specifications. On the other hand, as the size of the enterprises grew through mergers, the economies of output density decreased. In the short term, economies of scale for smaller firms showed an increase, while these economies become constant as the size of the enterprises grew. Long-term economies suggested a U-shaped average variable costs function. Bottasso & Conti (2003) also found economies of scope between water and sanitation.
In contrast to Bottasso & Conti (2003, 2007, 2009) used a complex cost function to test the hypothesis that WaSCs and WoCs share the same technology. For the average firm, the results showed the existence of both aggregate economies of scale, and scope diseconomies. More interestingly, the hypothesis that the two groups share the same parameters was rejected.

In the light of the wave of mergers and acquisitions, the Water Services Regulation Authority (Ofwat) commissioned a report to Stone & Webster Consultants (2004) because of its concern that its position on any potential future proposed mergers (and other restructuring) in the water industry in England and Wales should be informed by an understanding of economies of scale in the sector. They suggested that there is evidence of scale diseconomies for WaSCs, although the diseconomies declined in the sample period. Also, the integration of the water and sanitation services was associated with diseconomies of scale. They found no evidence of underlying cost reductions after the mergers. The operative cost rankings of the regulator Ofwat showed a deterioration in relative efficiency when surpassing 2.5 million connections to the water supply.

Strategic Management Consultants (2002) concluded that returns to scale vanished above the level of 400,000 connections, and quantified in current values the savings estimates arising from mergers in water companies. Those savings are attributed to economies of scale.

Saal & Parker (2006) estimated the economies of scale after regulatory changes, and controlled for technical efficiency differences, the net density, the quantity of wells, the water quality, and merger episodes. They find no economies of scale for the average company in the sample.

Antonioli & Filippini (2001) focused on water-distribution companies, attempting to determine their cost structure and, consequently, the existence of economies of scale and scope. Their results indicated the existence of economies of output and customer density and the presence of small diseconomies of scale. Thus, they did not recommend mergers, because the increase in the enterprises’ size pushes up costs more than proportionally.

Fraquelli & Moiso (2005) explored how water and sanitation have improved since the reforms were applied in 1995. They found economies of output density for all providers, regardless of size. The magnitude of those density economies is inversely related to the volume of water dispatched. With this in mind, and calculating the economies of scale for the different enterprises, the authors found economies of scale, decreasing up to 90 million cubic meters per year (serving approximately 1 million inhabitants).

**In search of a more efficient scale?**

Kim & Lee (1998) analyzed the effects of spatial integration in the water sector in South Korea and found evidence of economies of scale for the average firm in the sample.

Renzetti (1999) analyzed the economies of scale for the services of water and sanitation in 77 municipal enterprises of the Province of Ontario (Canada) in 1991, finding economies of output density for potable water provision and for sanitation services.

García & Thomas (2001) examined the cost structure of municipal water enterprises in France and concluded that mergers in districts are profitable. They found economies of output density, economies of customer density, and economies of scale.

Mizutani & Urakami (2001) studied the density economies of networks and the economies of scale of water provision in Japan. Their results indicated few diseconomies of scale at the average point of the sample and at optimal size for enterprises servicing 776,000 inhabitants. They also conclude that economies of network density exist, but not economies of scale.

Urakami (2005) focused on vertical integration between the water-intake-purification and water-distribution stages, and calculated economies of vertical integration in Japan. The results showed that economies of vertical integration

**Italy: from atomization to an ‘optimal scale’**

In this context, Fabbri & Fraquelli (2000) concluded that most of the 6,000 operators are classifiable as smaller than the Italian average and that they fall in the range where it is possible to enjoy increasing economies of scale. Additionally, the authors observed that total costs decrease when the customer density increases. They pointed out that the economies of scale in their study refer mainly to delivery costs, and that considerable economies of scale could be achieved in the production and in managerial and financial activities.
exist between the water-intake-purification and water-distribution stages.

Sauer (2005) studied the cost structure of rural water provision in Germany. He found that the optimal size for a water enterprise was one serving 66,000 inhabitants and 3,600 cubic meters of water delivered per consumer per year. In terms of customer density, the optimal size is 23 connections per kilometer of the network, with diseconomies of customer density starting from a level of 36 connections per kilometer. The former implies a U-shaped average cost function.

Zschille & Walter (2011) and Walter et al. (2009) also analyzed the German water sector and found evidence for unexploited economies of scale and substantial economies of density. They supported merger initiatives.

Martins et al. (2006a) studied 218 municipal providers of water and sanitation in Portugal in 2002. They concluded that economies of scale do exist for small and medium-sized enterprises, but not for larger ones. They recommended mergers when possible.

Martins et al. (2006a, b) showed that the average production scale in Portugal is below the estimated minimum efficient scale, and that large utilities have moderate overall diseconomies of scale and scope. In addition, there are moderate economies of scope from the joint production of potable water and wastewater collection up to the minimum efficient scale, suggesting advantages in merging small and medium-sized contiguous water utilities.

Martins et al. (2008) analyzed the existence of economies of scale and scope in order to consider mergers of firms and services in Portugal. The results showed that the current average output scale was lower than the efficient scale, while the larger enterprises exhibited diseconomies of scale and scope. Modest economies of scope jointly produce water and sanitation up to a point, suggesting that smaller firms could benefit from mergers.

In contrast to these studies in Portugal, Monteiro (2010) assessed the cost structure of the water industry at the retail level. He found diseconomies of scale water supply and wastewater drainage and treatment activities at the average provider of the sample (even with evidence of economies of output density for the same average of the sample). Additionally, the results contrast, in that economies of scale are found to increase with the size of the utility, except for wastewater collection.

In Switzerland, Baranzini & Faust (2010) studied the cost structure of water utilities. They found moderate returns to production density when the size of the utility decreases.

Revollo Fernández & Londoño (2008) studied the economies of scale and scope of the water and sanitation services in Colombia, aiming to consider mergers or breakups of providers. The results showed that the water and sanitation service exhibited economies of scale both in the short and long term. There were economies of scale in the short and long term both in water and sanitation for small- and medium-sized enterprises (with up to 100,000 customers), while larger firms showed diseconomies of scale both in the short and long term. The output level associated with constant economies of scale corresponds with 28 million cubic meters, which is equivalent to 150,000 customers. With respect to economies of scope, they found that an increase in the production in one of the outputs reduced the total variable cost of production in both products.

Filippini et al. (2008) estimated the economies of scale and efficiency for a sample of providers in Slovenia during the period 1997–2003. The optimal size of firm was that of the sample median size, implying that economies of scale are found in smaller firms and that diseconomies appear in larger ones. When the firms were categorized by size, economies of output and customer density were present in the three groups (small, medium and large). Economies of scale were also present in the medium-sized enterprises. The larger firms in the sample seemed to have exhausted the economies of scale.

Frone (2008) studied some topics related to the opportunity to regionalize services, merging some town providers in Romania. She also developed a strategic guide to the regional operation of the services to supply water to areas with 100,000 inhabitants.

De Witte & Dijkgraaf (2007) assessed the effects on efficiency levels arising from mergers in the Netherlands. They concluded that mergers have neither improved efficiency nor achieved economies of scale.

Prieto et al. (2009), in connection with García-Valiñas (2007), found economies of scale for the municipalities of Spain. They concluded that the optimal firm size is larger than the average size of the firms in the sample.

Malmsten (2008) analyzed utilities in Sweden in an attempt to determine the cost structure of the industry.
The results showed that economies of scope exist between water and wastewater in the industry. The larger utilities exhibited diseconomies of production output density and diseconomies of customer density, while the smaller utilities in the sample exhibited economies of production output density and economies of customer density.

Tsegai et al. (2009) found returns to scale (greater than one), and they suggested that a merger of South African providers would be economically advantageous. In addition, they recommended that the process of transferring water-services’ authority to ‘local’ municipality level should be reversed.

SCL Econometrics (2009) estimated economies of scale for the Chilean water industry, and found economies of scale in the production and treatment processes for enterprises of all sizes. Economies of output density were found for medium-sized and large firms. There were economies of scale at the industry level.

Cross-country studies driven by policy considerations

Tynan & Kingdom (2005) conducted a cross-country study, where they showed that certain providers, particularly those servicing populations of 125,000 or fewer, can reduce their operating costs per customer by increasing the scale of operation. The study covered 33 countries in Africa; Indonesia; Peru; the US; and Vietnam. The authors suggested that economies of scale exist when output is measured by volume, but inconsistent results appeared when customers or connections are considered.

Nauges & van den Berg (2007) examined the economies of scale in water and sanitation enterprises from four countries: Brazil, Colombia, Moldova and Vietnam. Their results indicated economies of scale for medium-sized firms throughout all of the countries, with the exception of Brazil.

Nauges & van den Berg (2008) conducted a study similar to that of Nauges & van den Berg (2007) but increased the number of countries in the sample from four to 14; they found comparable results.

Iimi (2008) quantified the optimal size of public–private partnership contracts in the sector by estimating a cost function. He showed that economies of scale exist but tend to diminish quickly as production increases. When the amount of water sold exceeded about 40 million cubic meters, the statistical significance of economies of scale disappeared. And there was no rationale for auctioning water operations with an annual water delivery of more than 400 million cubic meters under a single contract.

**DISCUSSION OF THE RESULTS**

In this section we will compare the findings in the literature and introduce the methodological issues. We will briefly present the overall conclusions of the five literature branches that we have identified in Table 1, which allow us to

<table>
<thead>
<tr>
<th>Table 1</th>
<th>A synthesis of the literature</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Subset of the studies</strong></td>
<td><strong>Main results</strong></td>
</tr>
<tr>
<td>United States</td>
<td>The papers on the US water sector recognize local fragmentation. They find that economies of scale can be achieved by consolidating small providers, but they condition the existence of economies of scale in larger providers to increases of output density</td>
</tr>
<tr>
<td>England and Wales</td>
<td>The sector departed from considerable consolidation, which was reinforced with various mergers. These mergers were justified by economies on the grounds of scale. It is possible that the larger providers had surpassed the optimal dimensions and were entering the region of diseconomies of scale. In addition, for smaller providers, the studies support the existence of economies of scope and economies of scale</td>
</tr>
<tr>
<td>Italy</td>
<td>In the Italian case, a factor of interest is the enactment of the economic concept of optimal size as a goal of public policy in order to attain economies of scale. The evidence for its achievement is mixed</td>
</tr>
<tr>
<td>Other countries with fragmented provision</td>
<td>The literature reported in this subsection is consistent with a U-shaped average-cost function, where economies of scale are present at smaller firms</td>
</tr>
<tr>
<td>Cross-country studies</td>
<td>In general, results are rather heterogeneous across countries</td>
</tr>
</tbody>
</table>

Source: Own compilation.
conclude that economies of scale are predominant for small water providers, while it is very likely that in very large ones providers diseconomies of scale could appear, as some studies in England suggest.

In choosing the functional form to represent the underlying production technology, there is a tradeoff between flexibility and stability. Simpler functional forms tend to be stable in their parameters, but they involve assumptions that are hardly met in reality. On the other hand, more flexible forms provide flexibility in the assumptions made regarding the underlying technology, but they require larger samples.

Table 2 shows that most of the studies chose a translog specification, and within this group many of the studies tested the Cobb–Douglas specification, which in all cases was rejected as a hypothesis. The translog specifications could be considered as the benchmark for functional forms analysis. This is recognized by Antonioli & Filippini (2001), who stated that the Cobb–Douglas specification was chosen because of multicollinearity between variables. Martins et al. (2006a, b, 2008) used quadratic and cubic forms in order to avoid the problems related to a lack of data in multiproduct cases. For the same reason, Torres & Morrison (2006) estimate a generalized Leontief quadratic. Bottasso & Conti (2007) and De Witte & Dijkgraaf (2007) attempted to provide greater flexibility in the underlying production function by estimating other functional forms. In these papers, the translog specification is replaced for the Fourier and the general composite.

### Table 2 | Functional forms and output definition

<table>
<thead>
<tr>
<th>Author and date</th>
<th>Functional form</th>
<th>Function</th>
<th>Output</th>
<th>Volume of water</th>
<th>Other variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ashton (1999)</td>
<td>Translog</td>
<td>Variable</td>
<td>U</td>
<td>Supplied</td>
<td></td>
</tr>
<tr>
<td>Baranzini &amp; Faust (2010)</td>
<td>Translog</td>
<td>Variable</td>
<td>U</td>
<td>Delivered</td>
<td></td>
</tr>
<tr>
<td>Bhattacharyya et al. (1994)</td>
<td>Cobb–Douglas/translog</td>
<td>Variable</td>
<td>U</td>
<td>Delivered</td>
<td></td>
</tr>
<tr>
<td>Bhattacharyya et al. (1995)</td>
<td>Translog</td>
<td>Variable</td>
<td>U</td>
<td>Billed</td>
<td></td>
</tr>
<tr>
<td>Bottasso &amp; Conti (2009)</td>
<td>Translog</td>
<td>Variable</td>
<td>U</td>
<td>Delivered</td>
<td></td>
</tr>
<tr>
<td>Bottasso et al. (2007)</td>
<td>General composite</td>
<td>Total</td>
<td>M</td>
<td>Delivered</td>
<td>Equivalent sewerage population</td>
</tr>
<tr>
<td>Fabbri &amp; Fraquelli (2000)</td>
<td>Translog</td>
<td>Total</td>
<td>U</td>
<td>Delivered</td>
<td></td>
</tr>
<tr>
<td>Filippini et al. (2007)</td>
<td>Cobb–Douglas/translog</td>
<td>Total</td>
<td>U</td>
<td>Delivered</td>
<td></td>
</tr>
<tr>
<td>Fraquelli &amp; Moiso (2005)</td>
<td>Translog</td>
<td>Total</td>
<td>U</td>
<td>Delivered</td>
<td></td>
</tr>
<tr>
<td>García &amp; Thomas (2001)</td>
<td>Translog</td>
<td>Variable</td>
<td>M</td>
<td>Billed</td>
<td>Water losses</td>
</tr>
<tr>
<td>García et al. (2007)</td>
<td>Translog</td>
<td>Variable</td>
<td>M</td>
<td>Billed</td>
<td>Raw water</td>
</tr>
<tr>
<td>Hayes (1987)</td>
<td>Quadratic</td>
<td>Total</td>
<td>M</td>
<td>Produced</td>
<td>Sewerage/environmental services</td>
</tr>
<tr>
<td>Hunt &amp; Lynk (1995)</td>
<td>Translog</td>
<td>Variable</td>
<td>M</td>
<td>Sold</td>
<td>The number of water connections and the number of sewage connections</td>
</tr>
<tr>
<td>Iimi (2008)</td>
<td>Translog</td>
<td>Total</td>
<td>M</td>
<td>Sold</td>
<td></td>
</tr>
<tr>
<td>Kim &amp; Lee (1998)</td>
<td>Translog</td>
<td>Total</td>
<td>U</td>
<td>Produced</td>
<td></td>
</tr>
<tr>
<td>Martins et al. (2006a)</td>
<td>Quadratic</td>
<td>Total</td>
<td>M</td>
<td>Delivered</td>
<td>Residential/non-residential/water losses</td>
</tr>
</tbody>
</table>

(continued)
With respect to the output definition (q), the last two columns of Table 2 report different proxies used in the literature. As can be shown, many of them have chosen different definitions of water volume (distributed, billed, produced, delivered, or supplied). Ideally, one would expect the water billed to be the best proxy, but sometimes the service is not metered and this variable is not as precise as the water delivered into the network. The difference between the concepts is related to the network losses due to leakage.

Alternatively, there are other output definitions, such as number of customers, population served, or connections. These definitions are relevant when the service is not metered. The use of other variables refers to multiproduct cost functions or problems of data availability.

The majority of the studies have opted for cost-function estimates, with the common assumption being that firms minimize costs in competitive input markets or that the input prices are given.
Most of the studies have estimated a cost function and, in doing that, prices of input variables are needed. Here there are two alternatives, the first one is to consider a public index as a proxy, and the second one is to infer the price by the ratio of expenditure and output. The problem of the first alternative is that a public index tends to be the same across the firms, and the second is that information on output and expenditure is not available in all cases.

Labor is used in most of the studies not only because of its importance, but also because it is easy to compute. The price of labor is generally obtained by dividing personal expenses – usually based on balance-sheet data – by the number of workers. There are potential problems with this variable if there are part-time workers or if there are contracted services that increase the expenses but not the number of employees. Capital is proxied by different forms in the studies.

Table 3 presents the input variables used in the studies. ‘P’ refers to price of the input, while ‘Q’ stands for quantity and is relevant for capital. There are two main inputs: labor and ‘others’. Labor accounts for 5% (Ashton 1999) to 60% (Revollo Fernández & Londoño 2008) of variable costs, and 4% (Bottasso & Conti 2007) to 85% (Kim & Lee 1998) of total costs.

Perhaps the most difficult variable to compute is capital. In the case of total (variable) cost functions, the definition refers to capital prices (quantity). In the case of total cost functions, the differences in the definition of capital price implies that the capital share ranges from as little as 6%, as in Kim & Lee (1998), up to 85%, as in Saal & Parker (2000).

The variable ‘others’ is a residual category, and its price is estimated as the ratio of dividing the difference of costs (total or variable) and the expenses of the computed inputs by a physical variable. As long as this variable is a residual one, it is very heterogeneous, so the issue in defining the price is the relevant physical variable to place in the denominator. The reviewed studies seem to deal with two alternatives: the first one is to divide ‘other expenses’

### Table 3 | Independent variables in the cost estimates

<table>
<thead>
<tr>
<th>Author and date</th>
<th>Labor</th>
<th>Capital</th>
<th>Energy</th>
<th>Materials</th>
<th>Bulk water</th>
<th>Chemical</th>
<th>Services</th>
<th>Others</th>
</tr>
</thead>
<tbody>
<tr>
<td>Antonioli &amp; Filippini (2001)</td>
<td>P</td>
<td>Q</td>
<td>P</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ashton (1999)</td>
<td>P</td>
<td>Q</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bhattacharyya et al. (1994)</td>
<td>P</td>
<td>Q</td>
<td>P</td>
<td>P</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bhattacharyya et al. (1995)</td>
<td>P</td>
<td>Q</td>
<td>P</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>P</td>
</tr>
<tr>
<td>Bottasso &amp; Conti (2005)</td>
<td>P</td>
<td>Q</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>P</td>
</tr>
<tr>
<td>Bottasso &amp; Conti (2009)</td>
<td>P</td>
<td>Q</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>P</td>
</tr>
<tr>
<td>Feigenbaum &amp; Teeples (1983)</td>
<td>P</td>
<td>P</td>
<td>P</td>
<td>P</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Filippini et al. (2008)</td>
<td>P</td>
<td>P</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>P</td>
</tr>
<tr>
<td>García &amp; Thomas (2001)</td>
<td>P</td>
<td>P</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>P</td>
</tr>
<tr>
<td>García et al. (2007)</td>
<td>P</td>
<td>P</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>P</td>
</tr>
<tr>
<td>Hayes (1987)</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>Iimi (2008)</td>
<td>P</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>P</td>
</tr>
</tbody>
</table>

(continued)
by network length, as in Fabbri & Fraquelli (2000), Fraquelli & Moiso (2005), Filippini et al. (2008) and Bottasso & Conti (2007, 2009); and the second one is to use the quantity of water delivered, as in Kim & Lee (1998), García & Thomas (2001) and García et al. (2007). Ashton (1999) divides the ‘other expenses’ by total assets. Saal & Parker (2000) and Bottasso & Conti (2003) use the retail-price index (RPI), but the problem in doing this is that this variable does not vary across the firms at one point in time.

With regard to the rest of the inputs, energy is the most important one. To determine the energy price, the energy expenses in some cases have been divided by a representative unit, such as the quantity of kWh consumed or the volume of water delivered. In other cases, the cost index of energy from the official statistics has been used. The unit price of raw materials is normally estimated using some price index or sub-index (such as the consumer price index, CPI, or a part of it). Bulk water is easy to estimate because information on expenditure and quantities of bulk water is often available to generate the unit price in cases where there is no set price for a cubic meter of bulk water. Chemicals and services costs are also proxy by means of an index.

Once the outputs and inputs have been defined, it should be noted that firms do not operate in a homogeneous context. Therefore, to control for differences in the operating environment, hedonic (or environmental, or control) variables have to be included.

As control variables, in most cases we found the network length, customers or density variables in the

<table>
<thead>
<tr>
<th>Author and date</th>
<th>Labor</th>
<th>Capital</th>
<th>Energy</th>
<th>Materials</th>
<th>Bulk water</th>
<th>Chemical</th>
<th>Services</th>
<th>Others</th>
</tr>
</thead>
<tbody>
<tr>
<td>Malmsten (2008)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>P</td>
</tr>
<tr>
<td>Martins et al. (2006a)</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>Martins et al. (2006b)</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>Martins et al. (2008)</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>Nauges &amp; van den Berg (2008)</td>
<td>P</td>
<td>Q</td>
<td>P</td>
<td>P</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nauges &amp; van den Berg (2007)</td>
<td>P</td>
<td>Q</td>
<td>P</td>
<td>P</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prieto et al. (2009)</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>Renzetti (1999)</td>
<td>P</td>
<td>P</td>
<td>P</td>
<td>P</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Saal et al. (2007)</td>
<td>Q</td>
<td>Q</td>
<td></td>
<td>P</td>
<td></td>
<td></td>
<td></td>
<td>Q</td>
</tr>
<tr>
<td>SCL Econometrics (2009)</td>
<td>P</td>
<td>X</td>
<td>P</td>
<td>P</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Shih et al. (2004)</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>Tsegai et al. (2009)</td>
<td>P</td>
<td>P</td>
<td></td>
<td>P</td>
<td></td>
<td></td>
<td></td>
<td>P</td>
</tr>
<tr>
<td>Tynan &amp; Kingdom (2005)</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td></td>
<td>n/a</td>
<td></td>
</tr>
<tr>
<td>Zschille &amp; Walter (2011)</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
</tr>
</tbody>
</table>

Source: Own compilation.
estimates. Network length sometimes reflects the capital stock and can also be a proxy for the energy needed to transport the water. The density variables seek to capture cost differences arising from demand concentration or dispersion. Likewise, some authors have incorporated variables distinguishing between underground- and surface-water sources. The underground sources are cheaper to purify (fewer chemical inputs), but demand more energy to pump. Unaccounted-for water may generally represent a measure of service quality or shed light on the age and condition of the network. Another variable used is the percentage of non-residential customers as a proxy for type of customer. Firms with a higher percentage of industrial customers usually have lower costs than providers that tend to focus on residential ones, because the demand is more concentrated, commercial costs are lower, and so on. Finally, a set of variables exists that reflects differences in the operational context: breaks in the network, the type of provider (public or private), quality standards, and so on.

The quantitative results of the literature surveyed must be analyzed with caution, since fair comparisons should be made with similar contextual conditions, and in these cases, the operational environment and the regional setting differ significantly in some studies. Therefore, in comparing the values of the different concepts, not only has the size of the utilities to be taken into account, but also the environment in which they operate. In addition, given the slow technological progress of the sector and the relevance of sunk capital, the history and the institutional framework are very important in determining the optimal size. [Sunk costs refer to the difference between the cost of resources before its investment and after that. Some resources are not specific and could be reallocated without major costs. Other capital is specific and once put in operation its recovery value is strongly diminished. Water mains and pipes, once buried, are sunk capital.] For example, old cities would be expected to have networks of small-diameter pipelines, while newer ones tend to have networks of larger-diameter pipelines because they took into account forecasts of population growth.

On average, scale diseconomies appear among the larger providers. Economies of scale and constant economies of scale (considering elasticity values between 0.95 and 1.05 as commonly found in the literature) are widespread. The empirical evidence showing that returns to output density are larger than economies of scale may suggest that constraints on expanding production would not arise at the production stage, but rather at the network stage.

Economies of output density are observed for the whole range of the providers with the exception of the samples in Ashton (1999), Stone & Webster Consultants (2004) and De Witte & Dijkgraaf (2007). The reason for this is that both the Dutch and English companies are the largest ones being analyzed. This implies that in the majority of the industry, there is room for expanding the output without increasing unit costs.

In the case of economies of scale, firms producing an annual output of 100 million to 200 million cubic meters are said to be producing on their long-term minimum cost. For higher output levels constant economies of scale begin to weaken and finally disappear.

Table 4 shows a synthesis of the quantitative results achieved by the studies, and in the Appendix a more detailed list of these results is presented. In Figure 1 we also display in graphic form the results of the studies concerning economies of scale in terms of water production. Each point represents one study.

**CONCLUDING REMARKS**

Water and sanitation services present particular features: a local natural monopoly; slow technological progress; non-contestable markets; high political and social sensitivity; and few incentives for private participation. These factors have generated different institutional arrangements that do
not always respond to efficiency goals. One way to increase efficiency would be to take advantage of economies of scale, which explains why it is important to study empirically economies of scale in the water industry. We have aimed to shed some light on and synthesize the empirical studies on the subject, in order to determine whether there is a tradeoff between centralization and decentralization.

We have observed differing motives for carrying out research into economies of scale. The first studies from the US were linked to an excessive atomization of the sector in that country and the discussion of public/private provision. In England, the discussion targeted research involving possible efficiency gains for larger-scale operations resulting from mergers. Continental Europe and other countries showed a clear concern for the optimal firm size. Finally, in recent years, international organizations have promoted cross-country studies based on large (but heterogeneous) databases.

The studies from a significant set of countries show economies of scale (in different countries with different situations) in populations of 100,000 to 1 million (or in some cases covering many millions), with population densities of up to 250 inhabitants per square kilometer, or with volumes up to 100 million to 200 million cubic meters per year. With lower populations, volumes or densities, we can expect constant returns to scale, and the same applies to the larger enterprises with regard to diseconomies of scale; smaller providers can gain by agglomerating services.

Likewise, given the diverse motives for the studies, the studies differ in their definition of economies of scale and in the context of the analyses. For this reason, discretion is highly recommended when analyzing comparisons, and it is always essential to keep in mind how the different economies of scale, scope and density are defined.

No easy answers arise from the study of the literature. The optimal scale seems to be highly particular to the providers’ conditions. Nonetheless, examination of the literature provides useful methodological insights which can be used to conduct further applied research.

ACKNOWLEDGEMENTS

A preliminary version of this paper was presented at the XLIVth Annual Meeting of the Asociación Argentina de Economía Política in Mendoza, Argentina, in November 2009. A previous paper, written for the Economic Commission for Latin America and the Caribbean of the UN (ECLAC), studied economies of scale in water and sanitation in Latin America (Ferro & Lentini 2010). We thank two anonymous referees for very useful comments. Any remaining errors are ours.

REFERENCES


Revollo Fernández, D. & Londoño, G. 2008 Análisis de economías de escala y alcance en los servicios de acueducto and
alcantarillado en Colombia. Comisión de Regulación de Agua Potable and Saneamiento Básico (CRA). Congreso de Economía, 50 Años CEDE. Universidad de Los Andes, Bogotá, Colombia.


Strategic Management Consultants 2002 Optimum entity size in the water industry of England and Wales: a review of the factors which influence the size of companies. Prepared for OFWAT. April.


First received 25 November 2010; accepted in revised form 4 August 2011