Setting network tariffs with heterogeneous firms: The case of natural gas distribution

SETTING NETWORK TARIFFS WITH HETEROGENEOUS FIRMS: THE CASE OF NATURAL GAS DISTRIBUTION

Teresa Romano
Corresponding Author
Email address: teresa.romano@ofgem.gov.uk. Ofgem – 32 Albion Street Glasgow G1 1LH – United Kingdom. Tel.: +44 0203 263 3066. The views expressed in this article are of the authors and do not reflect Ofgem's views on the topic.

Carlo Cambini
Department of Management and Production Engineering - Politecnico di Torino - corso Duca degli Abruzzi, 24 - 10129 Turin - Italy

Elena Fumagalli
Copernicus Institute of Sustainable Development – Utrecht University – Princetonlaan 8a, 3584 CB Utrecht, The Netherlands

Laura Rondi
Department of Management and Production Engineering - Politecnico di Torino - corso Duca degli Abruzzi, 24 - 10129 Turin - Italy

Abstract

The proper treatment of firm heterogeneity plays a crucial role in the application of benchmarking analyses for regulatory purposes. Within the realm of two-step approaches, this paper challenges the widespread adoption of single-variable clustering: heterogeneity has often multiple sources, which calls for more sophisticated clustering methodologies. In fact, reliable cluster-specific rankings provide firms' management with more realistic objectives as well as freedom to identify the appropriate strategies to improve efficiency. In order to provide regulatory guidance on this issue, we use a unique dataset of detailed accounting data and unbundled network-related costs for a panel of Italian gas distributors and we test two alternative methods: a hybrid clustering procedure (HCP) and a latent class model (LCM). Our results show that HCP and LCM perform better than size segmentation in the identification of classes, thereby leading to more reliable production frontiers, but do not support a conclusive preference for one or the other method. While both methods are sensitive to outliers, LCMs seem to provide deeper insights on the drivers of firm inefficiency. However, they also present stationarity and convergence issues, which might favour the implementation of HCP methods. Furthermore, the degree of discretionary judgement in the modelling decisions (e.g., model specification and choice of the partition) is slightly higher with LCMs than with HCP. In this respect, the HCP, with its lower modelling and analytical complexity, may feature as a more appealing option, facilitating the interactions between regulator and firm managers.

Keywords: Data Envelopment Analysis; gas distribution; network tariff regulation; cluster analysis; heterogeneity

JEL Codes: L51, L95

1. INTRODUCTION

When National Regulatory Authorities (NRAs) set network tariffs for energy distribution systems, they face a variety of technical, methodological and policy issues, from information asymmetries to the transition to less carbon-intensive energy systems. In practice, incentive-based pricing schemes are often employed to promote efficiency and innovation "by rewarding good performance relative to some pre-defined benchmark" (Jamasb & Pollitt, 2001, p. 108). Naturally, the proper identification of such benchmark is crucial and, to this aim, benchmarking analyses are often the preferred analytical tool (Haney & Pollitt, 2009). Nonetheless, implementation issues often arise when economic efficiency depends not only on managerial decisions, but also on external conditions partially observable by the regulator and, sometimes, beyond the firm's control. In fact, because significant rent extraction may occur, observed and unobserved heterogeneity has frequently become ground for judicial controversy and lengthy negotiations between firms and NRAs, threatening the stability of the regulatory decisions. In this context, the availability of adequate tools to deal with heterogeneity is crucial.

The Italian gas distribution sector provides an excellent case to study the problem of implementing an incentive structure which is robust to the presence of significant heterogeneity across numerous distribution system operators (DSOs). Like several other NRAs, the Italian energy regulatory authority (ARERA, *Autorità di Regolazione per Energia Reti e Ambiente*) implements *ad hoc* segmentation methods to identify classes of similar operators and define expected efficiency gains for each class (ARERA, 2008). Using an original dataset collected by ARERA from the entire population of Italian DSOs, this paper tests how alternative clustering methods can improve the accuracy of efficiency estimates when firm heterogeneity is an issue, discussing how well such methods perform when applied to data used for regulatory purposes.

As a starting point, we rely on the literature that questions the ability of *ad hoc* segmentation methods to lead to homogeneous reference sets (Agrell & Brea-Solìs, 2017). The literature shows that the methodology to identify homogeneous classes of firms prior to benchmarking is crucial to improve the accuracy of the efficiency estimates (e.g., Agrell & Brea-Solis, 2017; Nieswand & Seifert, 2018) and thus avoid setting allowances not reflecting actual costs and potential efficiency gains. Yet, more practitioner-oriented contributions are still lacking. Moreover, the few applications focus solely on the electricity sector (Agrell et al., 2013; Agrell & Brea-Solìs, 2017; Bjørndal et al., 2018; Dai & Kuosmanen, 2014; Llorca et al., 2014; Orea & Jamasb, 2017; Silva et al., 2019). Indeed, none of the existing benchmarking studies on gas distribution employs clustering to control for heterogeneity (Carrington et al., 2002; Ertürk & Türüt-Aşik, 2011; Farsi et al., 2007; Tovar et al., 2015; Zorić et al., 2009), including the only two studies that use Italian data (Erbetta & Rappuoli, 2008; Goncharuk & lo Storto, 2017).

To fill this gap, we apply a two-step benchmarking approach, acknowledged as successfully able to "reduce the unexplained variance previously claimed as inefficiency" (Agrell et al., 2013, p. 16). In the first step, we partition the sample into classes of firms with similar contextual factors/production possibilities by applying two classification methods, a combination of partitioning and hierarchical clustering techniques (which we labelled Hybrid Clustering Procedure, HCP) and a Latent Class Model (LCM). Both based on existing literature, these methods are more sophisticated than the simple *ad hoc* segmentation currently in use, as the number of classes is not arbitrarily set in advance but endogenously derived.[1] In the second step, we perform a Data Envelopment Analysis (DEA) to derive DSOs' efficiency scores within each class.[2] In addition to assessing the performance of HCP and/or LCM against *ad hoc* segmentation or full-sample benchmarking (e.g., Agrell et al., 2013), we also highlight advantages and disadvantages of their potential implementation in the regulatory practice.

Our data include highly detailed accounting information gathered by ARERA for the period 2008-2011. From the original dataset, we extract a balanced panel of 105 DSOs covering 94% of the natural gas delivered in Italy. Our dataset is particularly appropriate to address the methodological problems implied by firm heterogeneity as gas distributors in Italy feature an unusually high degree of heterogeneity in terms of customers served.[3] Moreover, differently from existing studies that use aggregated distribution costs related to both regulated and unregulated activities, we can rely on unbundled, network-related costs, which allows us to exclude from the analysis all expenditures for deregulated services (e.g., retail costs). This feature is crucial to prevent us from estimating DSOs' efficiency by mixing regulated and unregulated services.

This paper provides an empirical contribution to the literature as well as to the regulatory practice. Our findings not only confirm the drawbacks of *ad hoc* segmentation methods, but also that, by providing a more accurate identification of classes in the first step, HCP and LCM allow a more reliable identification of production frontiers. Nonetheless, they also support a more critical review of these approaches. If on one side HCP and LCM endogenously derive a partition suitable for subsequent efficiency analysis, on the other side outliers and, in the case of LCM, partial stationarity (i.e., not all firms stay in the same class over the

[1] Applications of partitioning or hierarchical methods are scant in the energy literature and mainly relate to electricity customers' segmentation (e.g., López et al., 2011). As for gas distribution, the only contribution is Alaeifar et al. (2014), a non-benchmarking study that employs hierarchical clustering to identify the optimal size of Swiss distributors. To the best of our knowledge, this is the first time a partitioning or hierarchical method is applied to energy distribution data and coupled with regulatory benchmarking. See Cagliano et al. (2003) for a similar approach applied to e-business strategies. Applications of model-based methods in the energy sector are relatively recent and use Monte Carlo generated data or data from electricity distribution or transmission (Agrell et al., 2013; Agrell & Brea-Solìs, 2017; Dai & Kuosmanen, 2014; Llorca et al., 2014; Orea & Jamasb, 2017).

[2] DEA is largely applied in several regulated industries such as local public transportation (see the survey by Daraio et al., 2016), as well as in environmental regulation (Manello, 2017). It is also employed in productivity and efficiency studies of non-regulated industries (see Devicienti et al., 2017 for an application to the manufacturing sector). There are many DEA applications to the electricity sector (among others, Cambini et al., 2014; Giannakis et al., 2005; Jamasb & Pollitt, 2003), far less for gas distribution. The latter studies account for firm heterogeneity by controlling for the effect on efficiency of variables such as customer and output density, firm size, age and ownership, and climatic factors (Carrington et al., 2002; Farsi et al., 2007; Zoric et al., 2009; Ertürk & Türüt-Aşik, 2011; Tovar et al., 2015).

[3] The ratio between the number of customers served by the largest and the smallest firm in our sample is 2,380 (1,639 in Erbetta & Rappuoli, 2008). Among the studies employing two-step benchmarking, only Dai and Kuosmanen (2014) and Agrell and Brea-Solìs (2017) feature a similar heterogeneity.

entire observation period) challenge the implementation of such methods. This sheds light on a number of possible implications for the regulatory practice.

On the one hand, our study indicates that firm heterogeneity has multiple sources, which suggests that more sophisticated clustering methodologies should be employed based on multiple separating variables. Moreover, cluster-specific rankings can provide firms' management with more realistic indications of the best performing operators working in a similar context, and thus general guidance to identify the appropriate actions and strategies to increase efficiency (Dai & Kuosmanen, 2014).

On the other hand, our contribution highlights that the implementation of LCMs is complex and implies discretionary evaluation by the regulator as regards, for example, model specification and selection criteria (Agrell Brea-Solis, 2017). HCP also involves discretion in the choice of the separating variables and of the appropriate number of classes, but its implementation is relatively simpler as compared to LCM. Complexity in implementation might become a critical issue in the methodological choice by the NRA as well as in the interactions between managers and regulators during regulatory reviews.

The remainder of the paper is organized as follows. Section 2 briefly describes the Italian gas distribution sector and the main concerns about the regulatory incentive scheme currently in place. Section 3 and Section 4 describe the dataset and the methodology, respectively. Section 5 presents and discusses the results, while Section 6 concludes and derives policy implications.

2. THE ITALIAN GAS DISTRIBUTION SECTOR

Until 2000 in Italy, natural gas distribution and retail services used to be carried out by local municipalities, either directly or through an appointed (private or public) company. This explains the large number of distribution companies (over 700). The liberalization process led to unbundling network and commercial activities (supply and retail) and to mandatory competitive awarding of service contracts. Since then, many small local utilities merged into companies that now serve province- or region-wide areas, while the two largest firms operating at the national level further extended their distribution activities through acquisitions. As a result, today the sector includes about 200 DSOs, but the largest twenty companies distribute 85% of total natural gas (32 billion cubic meters in 2018), suggesting a highly concentrated but also heterogeneous industry.

Since 2000, the Italian NRA has the task to set the annual allowed revenues of each DSO (ARERA, 2000; 2004; 2008; 2013). These are classified as "large", "medium" or "small" firms depending on the number of customers served (more than 300,000, between 50,000 and 300,000, and below 50,000, respectively). Allowed revenues are meant to cover network-related, metering, and commercialization costs, all separately reported by the DSOs.[4] In this paper, we focus on network-related activities, i.e. on the core service provided by DSOs (delivery of gas). The corresponding allowed revenues are meant to cover capital expenditures, depreciation, and operational expenditures. The empirical analysis uses "network-related operational expenditures" (hereinafter, *opex*), including the cost of labor, services, and materials. Two further aspects about *opex* are specifically relevant to this work.

The first one regards the annual updates during the 4-year regulatory period. Similar to other countries, in Italy, *opex* for each regulated firm is adjusted to account for inflation and an X efficiency factor (i.e., the expected annual efficiency gain), whose value is set at the beginning of each regulatory period. The X factor is not the same for all firms, but is defined per class, so as to be higher for small DSOs than for medium and large ones (respectively, 5.4%, 4.6%, and 3.2%, to be achieved annually in the regulatory period 2009-2012). The second one is that the NRA acknowledges that DSOs in different classes operate with different degrees of cost efficiency. In fact, *opex* for each firm is estimated at the beginning of the regulatory period as the product of a "standard operational unit cost" (in € per customer) and the number of customers served. The standard operational unit cost is lower for large firms than for medium and small ones, in line with the assumption of economies of scale. Moreover, a relatively higher standard operational unit cost is assigned to firms in the same class, whenever they present a lower *customer density* (measured in number of customers per meter of network), pointing to economies of density. To give an example, in the regulatory period 2009-2012, the standard operational unit cost was 56.46 €/customer for small DSOs with low customer density and 39.30 €/customer for large DSOs with high customer density.[5]

The above regulatory framework is the object of interest of our analysis. In fact, the implications of having different expectations in terms of standard operational unit costs and efficiency gains based on firm size and/or density is potentially quite relevant with respect to the potential of rent extraction, as well as for the desirable size of a distribution company (and consequently merger and acquisition strategies).

3. DATA

We use an original dataset of all the Italian regulated firms (DSOs) collected by ARERA, from which we extracted a balanced panel of 105 DSOs tracked from 2008 to 2011, covering 90% of customers in 2011 (hereinafter referred to as 'Full Sample'). The empirical analysis focuses on network-related operational expenditures which are, on average, 65% of each firm's total expenditures.

Table 1 reports the summary statistics for the Full Sample including technical and (inflation adjusted) accounting variables, as well as network-related measures of partial productivity. Looking at variables

[4] Commercialization costs refer to network-related and metering-related services (not to retail activities).

[5] Table A.1 in Appendix A provides more details on the evolution of incentive regulation in Italy over the period 2000-2020.

capturing DSOs' output characteristics, number of customers (*customers*), volumes of distributed gas (*volumes*), and area served (*area*), we note a remarkable heterogeneity across DSOs (large standard deviations with respect to the mean). Similar high levels of heterogeneity characterize the typical contextual variables for gas distribution: *customer density* (*customers* per *network* length), which captures differences between urban and rural networks; *output density* (delivered gas volumes per customer), which accounts for both higher per capita consumption in colder areas and higher share of non-residential load; and average *altitude* of the service area, capturing both the difficulty of serving mountain areas and the potential for higher distributed volumes due to lower temperatures.[6]

Table 1. Descriptive statistics – Full Sample (420 observations).

Variables	Mean	Std. Dev.	p25	p50	p75
customers	171,012	566,098	11,220	28,795	91,996
volumes [million m^3]	270	836	18.40	51.90	156.00
area [km^2]	1,805	6,532	90	289	751
customer density [*customers*/m]	0.088	0.049	0.057	0.080	0.103
output density [m^3/*customers*]	1856	737	1,391	1866	2,281
altitude [m]	184	179	57	131	253
opex [million €]	6.30	16.10	0.48	1.40	3.51
opex/customers [€/*customers*]	51.84	31.74	30.58	43.08	65.72
opex/volumes [€/m^3]	0.032	0.022	0.018	0.024	0.040
network [km]	1,897	5,886	156.31	401.39	1,172.78

Network-related operational expenditures (*opex*) as well as partial productivity measures (*opex/customers* and *opex/volumes*) also exhibit large standard deviations, suggesting significant heterogeneity across firms. In turn, this leads us to expect relatively low efficiency scores when performing benchmarking on the Full Sample, and offers a rationale for the current incentive structure. Finally, the variable network length (*network*) is included to serve as a physical measure of capital.

In the following, we test whether the size segmentation approach used by the regulator is appropriate to account for such heterogeneity. Then, we propose two alternative solutions for the identification of homogeneous classes of firms. Our analysis and the ensuing discussion have a bearing on the reliable identification of production frontiers, and thus on the accuracy of the expected efficiency gains.

[6] The number of customers and the volumes of gas delivered are fundamental drivers of operation, maintenance and repair costs of the DSO, and are therefore expected to have a positive impact on *opex* via labor, services and materials' costs. The same costs are also expected to decrease, on average, with higher *customer density* and to increase with *altitude*. High *output density* implies a higher use of the installed network capacity, which is efficient in terms of capital expenditures, but at the extreme, might signal a saturation of the existing assets (the need for investments) and lead to higher operational costs. Although the literature suggests that size and pipelines' material can influence *opex* (Carrington at al., 2002), this information is not available for all firms in the sample and therefore not included. Nevertheless, over the observed period, steel with cathodic protection constituted 90% to 97% of the material used by the DSOs annually sampled by the NRA (e.g., ARERA, 2012).

4. Methods

To deal with firm heterogeneity, we implement a two-step analysis (e.g., Dai & Kuosmanen, 2014). The first step identifies the number and composition of classes of firms with similar characteristics, while the second step estimates efficiency scores within each class. As described below, two alternative clustering procedures are used in the first step and a non-parametric approach (DEA) in the second (see, for example, Llorca et al., 2014). This two-step analysis is in line with the method proposed by Agrell et al. (2013) and was preferred over alternative approaches, given the nature of the industrial sector under observation, the compatibility with the regulatory practice, and the structure and limitations of the dataset.

There are, indeed, several alternative ways to account for firm heterogeneity. Focusing on DEA, they consist of either observing the impact of contextual variables on efficiency (Simar & Wilson, 2007, 2011; Banker & Natarajan 2008) – a method highly debated in the literature – or the simultaneous use of clustering and DEA to set more targeted efficiency incentives (Thanassoulis, 1996 and Afsharian et al., 2019). In fact, our approach, by deriving the optimal number of clusters in the first step, complements Thanassoulis (1996), where the number of clusters is exogenous. Nevertheless, being developed to account for differences in the output mix, this method would present some conceptual difficulties when applied to the gas distribution sector (where firms provide the same service to all customers). Moreover, although semi-nonparametric stochastic methods (Johnson & Kuosmanen, 2011) can also identify classes with large samples, we opt for DEA given it has been generally preferred by regulators (Haney and Pollit, 2009). Alternatively, Stochastic Frontier Analysis (SFA) models including unobserved heterogeneity in the frontier estimation (via True Fixed-Effect and True Random-Effect Models, see Greene, 2005) have also been employed.[7] As for the present study, the limited availability of input price data for many firms could allow the application of SFA models to a much smaller sample that our current Full Sample. Therefore, to avoid losing too many observations and reduce the heterogeneity of our sample, we decided to use SFA only as an additional robustness check.

4.1 First step: hybrid clustering procedure

The literature proposes three types of clustering techniques: hierarchical, partitioning, and model-based (see Everitt et al. 2011 for a comprehensive overview). In this section, we focus on the first two methods, which we combine into what we labelled, for brevity, the Hybrid Clustering Procedure (HCP).

This method involves the use of a hierarchical method first, followed by the application of a partitioning algorithm. The hierarchical method generates multiple partitions based on measures of distance between

[7] A related problem highlighted by Silva et al. (2019) is limited and noisy data, which make information on production technology and heterogeneity difficult to extract. As a solution, the authors propose a Stochastic Frontier Analysis (SFA) with generalized maximum entropy, a methodology which seems to be robust even in very small samples.

pairs of observations (e.g., Euclidean distance). A function of these pairwise distances (linkage) defines the distance between sets of observations. We use the Ward's linkage, defined as the increase in the variance of the distance with respect to all separating variables.[8]

Specifically, DSOs' proximity is measured using *customers* and *customer density* as main separating variables, in line with current regulatory practice (see Section 2). In line with previous literature (e.g., Farsi et al., 2007), *volumes* of gas delivered, *output density*, *altitude*, and *area* are also considered in alternative combinations to further describe firms' output characteristics and operating environment. All variables are standardized to avoid biases due to differences in scale and unit measures.[9]

The optimal number of clusters is selected by means of two "stopping" rules for continuous data (Milligan & Cooper, 1985), namely the Calinski-Harabasz rule and the Duda-Hart rule, which indicate the ideal number of clusters based on the highest value of given ratios (i.e., the "stopping" point of an iterative procedure). More specifically, the Calinski-Harabasz ratio is defined as

$$\frac{\text{trace}(B)/(g\text{-}1)}{\text{trace}(W)/(N\text{-}g)},$$

where B is the matrix containing the between-cluster sums of squares and cross-products, W is the corresponding within-cluster matrix, g is the number of clusters and N is the sample size. The Duda-Hart stopping rule, based on the idea of dividing each group into two subgroups, is given by $\frac{Je(2)}{Je(1)}$, that is, the sum of squared errors in the two resulting subgroups ($Je(2)$) over the sum of squared errors within the group that is to be divided ($Je(1)$).[10]

Starting from the identified optimal number of clusters, we apply a partitioning algorithm, which changes the composition of clusters until a given criterion is satisfied. We use the k-means algorithm, whereby each initial cluster is described by its mean, and the clusters' composition is iteratively updated by reallocating each firm to the cluster with the closest mean (in Euclidean distance terms). This process produces the final DSOs' classification, which is then subject to post-clustering tests (ANOVA and Scheffé) to check the appropriateness of the variables selected for the clustering.

This procedure exploits the advantages, while removing the shortcomings of each of the two separate techniques: the hierarchical method allows us to determine the optimal number of clusters and the initial

[8] To generate the partitions, we used the Stata command "cluster wardslinkage" and relevant post-estimation commands.

[9] Before going further, it is worth discussing the issue of sample size for clustering methods. In this respect, the market segmentation literature (see for example, Dolnicar et al., 2013) recognizes the lack of rules of thumb to determine an adequate sample size for cluster analysis and only generically points to the importance of ensuring a reasonable ratio between number of observations and clustering variables. To the best of our knowledge, only few contributions explicitly refer to the $5x2^k$ (with k being the number of clustering variables) rule by Formann (1984) for the minimum sample size, while Dolnicar et al. (2013) suggest a 70 x k rule. In our work we consider up to 5 clustering variables, implying a minimum sample size of 160 as per Formann's rule and 350 as per Dolnicar et al. (2013) suggestion. The size of our sample (412 to 420 observations) is well above both these thresholds.

[10] It is worth noting that a given clustering criterion can generate multiple optimal partitions.

cluster composition instead of setting them in advance, while the iterative nature of the partitioning method ensures a stable cluster assignment. The only drawback of this "hybrid" procedure is its sensitivity to outliers.

4.2 First step: latent class model

The alternative clustering approach is based on a formal statistical model (a latent class model, see Lazarsfeld & Henry, 1968) assuming that any given cluster of firms is characterized by different multivariate probability density functions (i.e., finite mixtures) for the selected firm-level variables. Therefore, once generalized linear models are estimated, regression coefficients vary across clusters, thus capturing firm heterogeneity. The well-known advantage of LCMs is their ability to simultaneously perform endogenous partitioning and robust technology estimation (Agrell & Brea-Solis, 2017).[11] Nevertheless, they are prone to convergence problems and to multiple likelihood maxima (Everitt et al., 2011). Similar to the HCP method, selection criteria for the optimal number of classes are still needed.

In this paper, we apply LCMs to a panel of gas distribution operators by estimating a Cobb-Douglas cost driver function. Specifically, our main specification assumes a function with one input (*opex*), two outputs (*customers* and *volumes*), and a set of contextual variables (*customer density, output density* and *altitude*):

$$\ln(opex_{it}) = \alpha_{0j} + \alpha_{1j}\ln(customers_{it}) + \alpha_{2j}\ln(volumes_{it}) + \alpha_{3j} customer\ density_{it} + \alpha_{4j} output\ density_{it} + \alpha_{5j} altitude_{i} + \varepsilon_{it|j}, \tag{1}$$

where *i* identifies the DSO, *t* the year, *j* is the latent class, and $\varepsilon_{it|j}$ is the normally distributed error term. Since benchmarking is performed in the second step, the latter does not embody any assumptions in terms of inefficiency. Moreover, given the limited sample size, we decided not to employ a translog functional form to reduce the number of parameters to be estimated.

The model's parameters are estimated via maximum likelihood, and the corresponding posterior probabilities are used to determine cluster membership. The identification of the more appropriate number of clusters is based on the Akaike's (AIC) and the Bayesian (BIC) information criteria.[12] The selected model and number of clusters will have the lowest AIC and BIC values, though it is possible that more than one suitable model exists.

[11] If technology is homogeneous across firms, differences between latent classes can be mainly interpreted as differences in contextual factors (Orea & Kumbhakar, 2004; Llorca et al., 2014).

[12] The appropriateness of AIC and BIC as model selection criteria depends on regularity conditions, which in the case of finite mixture models might easily not be verified.

4.3 Second step: the DEA model

To estimate the relative performance of DSOs we use DEA, a non-parametric benchmarking approach that allows for multiple inputs and outputs and identifies the industry frontier without imposing a functional form for production (Coelli et al., 2015).

Because we account for firm heterogeneity in the first step, we can benefit from a simpler model specification in the second step. Given the focus of this study, in the main DEA model (OPEX) we use network-related operational expenditures (*opex*) as the only input. Connecting customers to the grid and transporting gas to final users are the main network-related activities of a DSO. Thus, the number of customers and the total volume of gas delivered are the output variables (*customers* and *volumes*).

For completeness and in line with the literature, we also consider a second model (OPNTW), which includes network length (*network*) as an input (e.g., Carrington et al., 2002).[13] This provides a reliable, physical measure of capital expenses, as mains are the major capital component of distribution networks, and information on their length is usually accurate. Finally, a third and fourth model (OPEXA and OPNTWA) include the *area* served as an additional output - "larger service areas generally require larger and more spread networks, thus more operating and maintenance costs" (Farsi et al., 2007, p. 70). For reasons of brevity, these results are only reported in the Supplementary Material.

All the estimated DEA models are input-oriented, as in the gas distribution sector it is reasonable to assume that demand is mostly beyond firms' control, and assume Variable Returns to Scale (VRS). Scores are calculated as input efficiency measures according to Farrel (1957) and are bias corrected via bootstrap replications.[14]

5. Results and Discussion

This section presents and discusses the results of our analysis. Section 5.1 examines class identification via the current size segmentation approach and applies the Hybrid Clustering Procedure (HCP) as an alternative. Section 5.2 implements a Latent Class Model (LCM). To test the ability of these approaches to account for firm heterogeneity, Section 5.3 compares the results of the efficiency analysis after clustering via the three alternative methodologies and discusses implementation issues.

[13] Although the Italian NRA applies an X factor to operational expenditures only, most efficiency studies consider a model with at least two inputs (e.g., Carrington et al., 2002; Zoric et al., 2009; Ertürk and Türüt-Aşik, 2011).

[14] Nonparametric tests of returns to scale using the "nptestrts" STATA command confirm the appropriateness of the VRS assumption. DEA is performed using the "teradialbc" command (Badunenko & Mozharovskyi, 2016) with 2,000 bootstrap replications.

5.1 Size segmentation and Hybrid Clustering Procedure (HCP)

The regulatory thresholds that classify firms based on size were defined in 2003. Since they have not changed, we start by employing the k-means method over the number of customers to verify whether this partition is still valid for the DSOs in our dataset which covers the period 2008-2011. Assuming that the number of classes (3) and the separating variable (*customers*) remain the same, we find that one of the three classes contains only two companies. These are indeed the two largest firms in this respect, as they serve a number of customers at least 13 times higher than the sample mean. Hence, as customary in clustering procedures, and only for the purpose of this preliminary analysis and the following HCP, we excluded the two outliers from the Full Sample ("103 Sample" hereafter).

The resulting delimiting thresholds (184,697 and 472,949 *customers*) significantly differ from those indicated by the regulator (50,000 and 300,000 *customers*), suggesting that classes identified at a given point in time might no longer reflect the structure of the same industry observed in later years. At the same time, differences in sample size might also contribute to explain the discrepancy. A comparison of the population of DSOs with the Full Sample used for the present paper indicates that the firms dropped due to missing or inconsistent data were mainly small.

When we then employ the hierarchical method, results suggest an optimal number of three classes (see **Table A.2** in Appendix A), which we use in the subsequent k-means clustering. As a robustness check, we applied the HCP to the Full Sample as well. This further analysis confirmed the presence of a fourth cluster consisting of the two excluded outlier firms.

Turning to the detailed results, **Table 2** shows that the average values of the classifying (and other) variables steadily increase/decrease from Class 1 to 3 (except for *output density*). Specifically, from Class 1 to Class 3 firms serve a greater number of *customers*, transport larger *volumes* of gas, and serve territories characterized by lower *altitude* but higher *customer density*. Moreover, going from Class 1 to Class 3, *area* decreases while *network* increases. In other words, Class 1 can be interpreted as the class of firms serving rural areas, Class 2 semi-urban areas, and Class 3 urban areas. The variable *output density* is larger in Class 2 and Class 1, which is compatible with a larger industrial consumption outside urban areas and larger residential consumption at relatively higher altitudes. As expected, average *opex* are also increasingly larger when going from Class 1 to Class 3. Nonetheless, the partial productivity measure *opex/customers* is consistently higher in Class 1 than in Class 2 and 3, while the same measure per volume (*opex/volumes*) indicates higher partial productivity in Class 2 than in Class 3 and Class 1. The identified classes are fully stationary - the DSOs' allocation does not change over the observation period.

Table 2. Descriptive statistics of the three classes: HCP.

Variable Mean (Std. Dev.)	103 Sample	Class 1 – HCP	Class 2 – HCP	Class 3 – HCP
Customers	102,922 (211,886)	42,404 (87,613)	92,254 (192,199)	229,476 (331,222)
volumes [million m^3]	169 (350)	58 (70)	174 (351)	289 (496)
area [km^2]	966 (2,211)	1,391 (2,459)	920 (2,312)	633 (1,037)
customer density [*customers*/m]	0.088 (0.050)	0.047 (0.020)	0.079 (0.023)	0.179 (0.057)
output density [m^3/*customers*]	1,862 (743)	1,891 (949)	1,985 (639)	1,259 (584)
altitude [m]	184 (181)	489 (166)	122 (86)	79 (84)
opex [million €]	4.62 (9.95)	2.12 (4.13)	4.50 (10.50)	8.40 (11.50)
opex/customers [€/*customers*]	52.32 (31.83)	63.73 (40.90)	51.82 (29.98)	39.98 (20.07)
opex/volumes [€/m^3]	0.032 (0.022)	0.039 (0.026)	0.028 (0.017)	0.040 (0.030)
network [km]	1,152 (2,281)	797 (1,253)	1,214 (2,530)	1,322 (2,069)
Observations	412	77	275	60

5.2 Latent Class Model (LCM)

We also use an LCM as a clustering alternative. With the specification in eq. (1), convergence is ensured up to five classes. As illustrated in **Table A.3** in Appendix A, both AIC and BIC show the largest improvement when moving from three to four classes. Consistently, we select four as the optimal number of classes and use the estimated posterior probabilities to define observations' class membership.[15] The estimated coefficients for each class are reported and discussed in Appendix A (see **Table A.4**).

Table 3 reports the descriptive statistics per class. Class 1 includes firms delivering relatively small outputs (*customers* and *volumes*) and characterized by low *customer density* and high *altitude* (average *output density*). Average *opex* are only slightly higher than for DSOs in Class 2, characterized by higher levels of outputs, higher *customer density*, and lower, but still relatively high *altitude* (and average *output density*).[16] Firms in Class 3 have the highest outputs across classes, a higher *customer density* and significantly higher average *opex* than firms in Class 1 and Class 2 (also lower *altitude* and lower *output density*). Firms in Class 4 appear peculiar in that they produce outputs aligned with those produced by firms in Class 2, but with much higher average *opex* (in the order of three times more than firms in Class 2). Nevertheless, firms in Class 4 also serve areas

[15] These probabilities show that firms can be distinguished with an acceptable degree of confidence – minimum probabilities are always greater than 0.42.

[16] When looking at the mean, the number of *customers* in Class 1 is lower than in Class 2, while *opex* is higher in Class 1 than in Class 2. When considering median values, Class 1 has a lower number of *customers* and lower *opex* (see Supplementary Material).

with higher *customer density* and noticeably higher *output density*. The latter two indicators suggest a saturation of production resources, which might partially explain the difference in average *opex* with respect to Class 2.

In order to investigate clusters' characteristics further, we performed both stationarity and outlier detection analyses[17], which confirmed that this partition is able to separate out firms' technological differences relatively well. Specifically, the generally high degree of stationarity (73% of DSOs is either in the same class over the 4 years or changes class at most once) suggests that the identified clusters tend to capture persistent technological differences (particularly Class 2 and 3). Nonetheless, only 30% of the firms in Class 4 are stationary. In our case, this seems to indicate that these firms are more likely to be undergoing a restructuring process due to mergers and acquisitions, although the interpretation of this result remains complex and requires specific knowledge of the sector.

Interestingly, the outlier detection performed signals Class 1 as a cluster of outliers (24 out of 26). This indicates that the class does not necessarily capture a unique, separate technology, thus making the actual implementation of LCM for regulatory purposes more difficult. The same issue does not emerge for other classes.

Table 3. Descriptive statistics of the four classes: LCM.

Variable Mean (Std. Dev.)	**Full Sample**	**Class 1 – LCM**	**Class 2 – LCM**	**Class 3 – LCM**	**Class 4 – LCM**
Customers	171,012 (566,098)	91,473 (261,845)	142,837 (534,582)	257,653 (724,831)	141,862 (254,048)
volumes [million m^3]	270 (836)	140 (379)	231 (787)	397 (1,070)	223 (453)
customer density [*customers*/m]	0.088 (0.049)	0.083 (0.047)	0.087 (0.049)	0.089 (0.050)	0.094 (0.049)
output density [m^3/*customers*]	1,856 (737)	1,839 (684)	1,868 (686)	1,793 (703)	1,983 (1,126)
altitude [m]	185 (179)	197 (208)	193 (189)	169 (152)	169 (175)
opex [million €]	6.30 (16.1)	4.61 (15.3)	3.97 (11.70)	9.74 (21.20)	12.20 (20.00)
opex/customers [€/*customers*]	51.84 (31.74)	19.61 (11.89)	39.01 (16.19)	66.92 (25.97)	113.51 (37.08)
opex/volumes [€/m^3]	0.032 (0.022)	0.012 (0.009)	0.023 (0.012)	0.043 (0.022)	0.067 (0.023)
network [km]	1,897 (5,885)	1,417 (4,018)	1,554 (5,516)	2,830 (7,382)	1,579 (3,375)
area [km^2]	1,805 (6,532)	1,442 (4,350)	1,402 (5,979)	2,865 (8,468)	1,408 (3,374)
Observations	420	26	243	115	36

[17] Specifically, we applied non-parametric partial frontier efficiency analysis for outlier detection (order-*alpha*, see Tauschmann, 2012).

Finally, it is worth noting that, if we look at 12 DSOs classified as large by the NRA, both HPC and LCM results allocate DSOs to different clusters, suggesting that environmental variables are at least as important as size to define what are similar firms.[18]

In terms of efficiency analysis, according to the literature, the LCM is expected to improve on the regulator's ability to deal with observed and unobserved heterogeneity as compared to both size segmentation and HCP. To test this, we turn to the second step of the analysis.

5.3 DEA ESTIMATION

The efficiency scores for the OPEX DEA models estimated on the Full Sample and the 103 Sample are similar and suggest the presence of substantial inefficiency in the sector's network-related activities (**Table 4**). As the average DSO operates at 32.3% corrected efficiency for the Full Sample (32.9% for the 103 Sample), it is clear that these results are significantly lower than those obtained in the two previous studies regarding the Italian gas distribution sector. Erbetta and Rappuoli (2008), with smaller samples and less heterogeneous data, estimate an average VRS efficiency of 63.4%, while Goncharuk and lo Storto (2017) report an average VRS efficiency of 75.1%.[19]

Table 4. Descriptive statistics of VRS efficiency scores from OPEX model: Full Sample and 103 Sample.

	Mean	**Std. Dev.**	**Min**	**p25**	**p50**	**p75**	**Max**	**Obs.**
Full Sample								
Eff. Scores	0.368	0.213	0.060	0.204	0.314	0.479	1	420
Bias Corrected Eff. Scores	0.323	0.181	0.048	0.182	0.287	0.420	0.931	420
103 Sample								
Eff. Scores	0.377	0.227	0.060	0.204	0.317	0.487	1	412
Bias Corrected Eff. Scores	0.329	0.186	0.049	0.185	0.288	0.434	0.945	412

[18] For example, 5 firms are assigned to Class 1- HCP, 4 firms to Class 2-HCP, and 1 firm to Class 3-HCP (2 are outliers, not included in the HCP classification). Class 1 - HCP and Class 2 - HCP can be interpreted as firms serving 'rural' and 'semi-urban' areas. This is in line with expectations, as urban areas in the North and Centre of Italy are mostly served by local, municipal companies of medium size. The only "large" company in Class 3 – HCP served urban areas located in the South of the country. A more general comparison between HCP and LCM partitions can be found in the Supplementary Material.

[19] Nonetheless, it is worth noting that our results are not fully comparable with these studies, because they differ from ours in the sample size and in the choice of inputs and outputs. Erbetta and Rappuoli (2008)'s model uses our same outputs but total expenditures as the only input. Their sample includes 46 Italian DSOs in the pre-liberalization period (1994–1999). In their study of Italian and Ukrainian DSOs, Goncharuk and lo Storto (2017) rely on multiple inputs (material costs, employees and fixed assets) and outputs (volumes of natural gas and service area). Their sample includes 36 Italian and 30 Ukrainian companies observed in 2013. Notably, both studies refer to gas distribution as a whole, while ours focuses on network operations only. Nonetheless, our results differ from other DEA-based benchmarking analyses. Carrington et al. (2002) find average VRS efficiency of 87% for their sample of Australian and U.S. gas distribution operators; Zoric et al. (2009) average VRS efficiency of 71% in the UK, the Netherlands, and Slovenia; Ertürk and Türüt-Aşik (2011) average efficiency of 83% in Turkey; Tovar et al. (2015) 78% in Brazil. For further comparison with the literature we refer to the results (reported in the Supplementary Material) obtained using the DEA models that include *network* (as an additional input) and *area* (as an additional output).

Nevertheless, it seems unlikely that such a difference in performance between an average Italian DSO and the firm(s) on the DEA frontier is only due to inefficiency. More sensibly, it highlights that clear comparability issues arise when using the full sample without accounting for firms' heterogeneity. For example, as shown in **Table 5**, higher efficiency scores are found, on average, by using the regulator's size segmentation approach. Here we find an average bias corrected efficiency of 36.7%. However, small DSOs still exhibit a large variance in their performance, while medium-sized DSOs are associated with higher levels of efficiency than large ones. In this respect, two remarks are in order.

First, this result clearly supports our claim that size (the number of *customers*) cannot be the only variable affecting firms' efficiency. Moreover, our result is in line with the existing literature, which finds economies of scale for smaller DSOs, but none for large ones (Carrington et al., 2002; Farsi et al., 2007; Erbetta & Rappuoli, 2008). Second, that same result clearly challenges the regulatory decision to set for medium DSOs higher X-factors than for large ones. Indeed, our results suggest that lower efficiency improvements should be expected from medium-sized firms rather than from large ones. This is opposite to the current regulator's assumptions, which thus would appear less neutral (where neutrality implies absence of biased treatment across operators), in itself a desirable characteristic of regulatory decisions (Agrell & Brea-Solìs, 2017).

Table 5. Average VRS efficiency scores per classes (OPEX model) – size segmentation, HCP, and LCM.

	OPEX		**Obs.**
	Eff. Scores	Bias Corrected Eff. Scores	
Size segmentation			
Class 1 – Large	0.538	0.417	46
Class 2 – Medium	0.581	0.529	111
Class 3 – Small	0.335	0.281	263
Total	*0.422*	*0.367*	*420*
HCP			
Class 1 – HCP	0.527	0.468	77
Class 2 – HCP	0.426	0.377	275
Class 3 – HCP	0.599	0.514	60
Total	*0.470*	*0.414*	*412*
LCM			
Class 1 – LCM	0.820	0.753	26
Class 2 – LCM	0.502	0.445	243
Class 3 – LCM	0.714	0.676	115
Class 4 – LCM	0.810	0.747	36
Total	*0.606*	*0.553*	*420*

When considering average efficiency scores obtained with HCP, we observe that the average efficiency scores obtained for Class 1-HCP and Class 3-HCP (77 and 60 obs., respectively) are aligned or higher than those obtained for Class 1-Large and Class 2-Medium (46 and 111 obs., respectively) obtained with size segmentation. Also, the number of observations with very low efficiency scores in the largest size segmentation Class 3-Small (263 obs.) decreases when compared with scores in Class 2-HCP (275 obs.).[20] The latter is even more visible for the largest class obtained with the LCM (Class 2-LCM, 243 obs.). Furthermore, Class 2-LCM (243 obs.) and Class 3-LCM (115 obs.) have higher average efficiency scores than classes of similar or smaller size obtained with size segmentation or with HCP. Notably, when employing the LCM, the number of observations with very low efficiency scores (e.g., below 30%) decreases substantially, pointing to a key advantage of this approach (detailed descriptive statistics for the efficiency scores reported in **Table 5** are found in the Supplementary Material – see also Figure S1).

Moreover, the fact that large firms as per NRA definition are assigned to different clusters in both methods confirms again that size is not the only explanation for a DSO's efficiency. For example, we note that five of the large stationary firms are allocated to Class 3-LCM or Class 4-LCM, which exhibit a relatively high average efficiency score with respect to other classes. This would be in line with the current regulatory approach, whereby efficiency improvements for "large" distributors are relatively less demanding. However, a different approach would be envisaged for the "large" firms allocated to Class 2-LCM (relatively low average efficiency score). Because these are only four, it would be feasible for the regulator to individually analyze their peculiarities and motivate an expectation of higher efficiency improvements. Similar considerations can be made for the HCP partition.

In sum, the overall average bias-corrected efficiency increases to 41.4% with HCP, and to 55.3% with LCM.[21] This confirms that, by running a cluster analysis before benchmarking, NRAs would reduce the heterogeneity component otherwise erroneously attributed to inefficiency, and obtain more reliable and realistic measures of relative performance within classes.[22] Moreover, relatively to the Italian gas distribution, the LCM performs better than other approaches when the same input-output relationship is employed in the two steps (Llorca et al., 2014), also in the presence of contextual factors (Nieswand & Seifert, 2018). Another advantage of LCM with respect to HCP is that, by estimating a cost driver function, it captures better the differences in the technology employed by firms. Instead, HCP relies more on topological differences, which do not necessarily imply technological differences. This feature is especially relevant for the regulator's understanding of the drivers of firm efficiency, a field where the evidence provided by the existing literature is inconclusive.[23] For example, results are mixed in the two previous

[20] This would result in relatively easier efficiency targets for the firms in this class, reflecting a feasibility principle – improvements in cost efficiency take time to be implemented.

[21] These results are closer to those in Erbetta and Rappuoli (2008) and Goncharuk and lo Storto (2017), and are also robust to the inclusion of area as an additional output (see Supplementary Material) - average efficiency scores increase further compared to the OPEX model, although not substantially. Moreover, as expected, the models including network length as an input exhibit even higher average efficiency scores.

[22] Note that average efficiency scores were also higher than those obtained under size segmentation when HCP and LCM were applied using customers number as the only separating variable and assuming a three-class partition. Nonetheless, in both cases the average efficiency scores were lower than those obtained using multiple clustering variables. ~~More details on the different partitions can be found in the Supplementary Material.~~ We thank an anonymous referee for ~~pointing out this issue~~suggesting this additional analysis. Further details on can be found in Appendix B.

[23] Ertürk and Türüt-Aşik (2011) find that climate and customer density help explain differences in firms' efficiency. Differently, Carrington et al. (2002) find that climate and network age have no significant effect, but they identify that

benchmarking studies on Italian DSOs. Erbetta and Rappuoli (2008) highlight the disadvantages of low customer density and suggest mergers among small firms. Goncharuk and lo Storto (2017) find no clear relationship between efficiency and geographical extension of the service area or its population.

However, from the practitioners' point of view, it is important that also the application of the HCP leads to interesting results. In fact, the identification of classes via LCM in the first step is, in practice, a rather complex task. In this regard, our analysis suggests that the difficulty in addressing heterogeneity can be amplified both by practical limitations in the implementation of LCMs and by data availability. Specifically, we highlight the convergence problems in LCM's estimation. We ran several specifications, but most of them did not converge with more than three classes – in particular those that included input prices, which are to prefer for a proper specification of the cost function.[24] Moreover, the choice of the number of classes based on information criteria entailed some discretion: while the AIC criterion generally showed improvements as the number of classes increased, the BIC criterion exhibited a non-univocal pattern, thus making the choice less straightforward. Further reasons for dismissing a partition were a discrepancy between the sign of the estimated coefficients and common technological knowledge of the sector under study, or the presence of at least one class with very few observations (precluding a robust estimation in the second step, even with a parsimonious DEA model).

Overall, while our results clearly suggest moving away from simple size segmentation methods to better address firm heterogeneity, the choice between the alternative approaches (HCP and LCM) is less straightforward, especially in terms of practical implementation by the regulator. Indeed, both methods are statistically sophisticated, but also sensitive to outliers. Despite LCMs seem to provide deeper insights on the drivers of firm inefficiency, issues like data availability and recurrent convergence problems might favor the implementation of HCP methods.[25] Furthermore, the degree of discretionary judgement in the modeling decisions (e.g., model specification and choice of the partition) is slightly higher for LCMs than for HCP. In this respect, the HCP, with its lower modeling and analytical complexity, may feature as a more appealing option.

6. CONCLUSIONS AND POLICY IMPLICATIONS

This paper focuses on Italian gas distribution to evaluate the challenges posed by firms' heterogeneity in the application of benchmarking analysis for the regulation of network infrastructures. This is a particularly

small scale is a relevant source of inefficiency, in line with Zoric et al. (2009) that find that larger (and older) distributors perform better. Tovar et al. (2015) point out that customer and output density matter, while Farsi et al. (2007) find that network size and customer density are relevant efficiency drivers.

[24] Similar difficulties occurred in other applications (e.g., Agrell & Brea-Solìs, 2017) but not in others (e.g., Agrell et al., 2013 meet convergence problems after six classes, while Llorca et al., 2014 report none).

[25] For example, the use of LCMs may significantly restrict the number of specifications that can be employed, up to a point where it might "preclude the implementation of the theoretically appropriate model chosen by the regulator" (Agrell & Brea-Solìs, 2017, p. 367, footnote 15).

relevant issue not only in the Italian context, but also in other countries characterized by regulated sectors with a relatively high number of firms. In Europe, countries with a relatively high number of natural gas distribution companies include Germany, France, Poland, Romania and Portugal, while countries outside Europe include, for instance, Brazil, Colombia, Chile, Peru, and the US. Moreover, the ability of dealing with heterogeneity in an appropriate way is also crucial when performing international benchmarking analyses.

Differently from existing studies that use aggregated distribution costs, we employ unbundled, network-related costs, thus excluding expenditures for deregulated services (e.g., retail costs). We analyze and compare three approaches in terms of their ability to identify realistic and reliable best practices for DSOs and in terms of implementation complexity. All three involve a two-step analysis that identifies homogeneous classes of firms before estimating efficiency scores within classes through DEA, but they differ in the clustering methodology they use in the first step: size segmentation, hybrid clustering procedure, and latent class approach.

We find that all clustering solutions are superior to benchmarking on the full sample, but HCP and LCM perform better than size segmentation in providing realistic efficiency estimates. These results indicate that, for the identification of best practices within heterogeneous firms, benchmarking based on clustering is superior to one on the full sample. This superiority holds when focusing on a single activity and/or cost component (e.g., network operations) of the tariff setting procedure. At the same time, our results do not univocally lead to a strong preference for one or the other more advanced methodologies.

Our study has several policy implications for the regulatory and industry practice. First, our results advise against size segmentation methods for classification purposes and advocate in favor of clustering procedures that account for multiple sources of heterogeneity. In this respect, variables that are normally available to regulators with a high level of precision and that are stable over longer periods of time (e.g., customer density) are very useful.

Second, our study challenges the Italian NRA's assumption that expected productivity gains should linearly depend only on the DSOs' size. On the contrary, equitable and realistic efficiency goals should guide managerial decisions in network-related activities. In fact, a narrow focus on firm size might even prevent managers exploring different opportunities for efficiency gains. On a side note, given the ongoing restructuring process of the Italian gas distribution sector, we also argue that the clustering criteria should be at least periodically updated.

Third, when it comes to the choice of alternative clustering techniques, NRAs should consider that both HCP and LCM can adequately address heterogeneity through endogenous clusters' creation. The adoption of a benchmarking method that properly accounts for firm heterogeneity will reduce lengthy negotiations with regulated firms and favor regulatory stability. However, both methods require proper treatment of potential outliers and imply a non-negligible degree of subjective judgment on the part of the regulator. The latter is slightly higher for LCMs, where particular attention should also be paid to model specification: on the one hand, different specifications may lead to different (sometimes unrealistic) partitions; on the other hand, it may be difficult to identify at least one partition producing a workable number of observations per class. In these cases, the regulator would need to adopt one of the few viable specifications, thus making a discretionary choice that might be difficult to justify to firm management and industry stakeholders. Moreover, LCMs typically require large datasets. Even with a relatively high number of DSOs, a sufficiently long time series would still be necessary. This could be an issue for recent NRAs (that lack the necessary data), when accounting rules change, or when the industry undergoes structural modifications (which would undermine the stability of clusters' composition).

Nonetheless, in contexts where data availability is not an issue and the chosen model specification produces reasonable results, LCMs should be still preferred, as the method offers deeper insights on the drivers of firms' inefficiency. When this is not the case, the HCP becomes the most suitable option. The method is relatively simpler to implement, and the operating conditions of potential outlier firms could be separately addressed by the regulator. In this sense, it provides an appealing option for regulators in jurisdictions where less modeling complexity would facilitate communication with different stakeholders (consumer associations, regulated firms) and ensure higher trust and participation in the regulatory process.

To conclude, when selecting a clustering approach preliminary to benchmarking, all these factors should be carefully considered to balance implementation complexity and effective identification of firms' best practices. While our study offers a first contribution in this direction, further work should focus on solving these complexities and providing clear guidelines to NRAs that have to deal with heterogeneity in benchmarking analyses. To this end, further empirical investigation (and related micro-data collection) is needed on the technological characteristics that drive the variance of the efficiency of energy firms. This will be crucial to inform policy decisions that shape the performance of network system operators under the current technology transition.

REFERENCES

Afsharian, M., Ahn, H., & Thanassoulis, E. (2019). A frontier-based system of incentives for units in organisations with varying degrees of decentralisation. *Eur J Oper Res, 275*, 224-237.

Agrell, P.J., & Brea-Solìs, H. (2017). Capturing heterogeneity in electricity distribution operations: A critical review of latent class modelling. *Energy Policy, 104*, 361–372. https://doi.org/10.1016/j.enpol.2017.01.046

Agrell, P.J., Farsi, M., Filippini, M., & Koller, M. (2013). Unobserved heterogeneous effects in the cost efficiency analysis of electricity distribution systems. Economics Working Paper Series, WP No. 13/171, January 2013.

Alaeifar, M., Farsi, M., & Filippini, M. (2014). Scale economies and optimal size in the Swiss gas distribution sector. *Energy Policy, 65*, 86–93. https://doi.org/10.1016/j.enpol.2013.09.038

ARERA (2000). Definizione di criteri per la determinazione delle tariffe per le attività di distribuzione del gas e di fornitura ai clienti del mercato vincolato. Deliberazione n. 237/00. Available (in Italian) from: www.arera.it.

ARERA (2004). Definizione di criteri per la determinazione delle tariffe per l'attività di distribuzione di gas naturale. Deliberazione n. 170/04. Available (in Italian) from: www.arera.it.

ARERA (2008). Testo unico della regolazione della qualità e delle tariffe dei servizi di distribuzione e misura del gas per il periodo di regolazione 2009-2012 (TUDG). Deliberazione n. ARG/gas 159/08. Available (in Italian) from: www.arera.it.

ARERA (2012). Relazione annuale. Available (in Italian) from: www.arera.it.

ARERA (2013). Testo Unico delle disposizioni della regolazione della qualità e delle tariffe dei servizi di distribuzione e misura del gas per il periodo di regolazione 2014-2019 (TUDG). Deliberazione n. 573/2013/R/gas. Available (in Italian) from: www.arera.it.

Badunenko, O., & Mozharovskyi, P. (2016). Nonparametric frontier analysis using Stata. *The Stata J, 16(3)*, 550-589.

Banker, R.D., & Natarajan, R. (2008). Evaluating contextual variables affecting productivity using Data Envelopment Analysis. *Operations Research, 56(1)*, 48-58. https://doi.org/10.1287/opre.1070.0460

Bjørndal, E., Bjørndal, M., Cullmann, A., & Nieswand, M. (2018). Finding the right yardstick: regulation of electricity networks under heterogeneous environments. *Eur J Oper Res, 265*, 710-722.

Cagliano, R., Caniato, F., & Spina, G. (2003). E-business strategy - How companies are shaping their supply chain through the Internet. *International Journal of Operations & Production Management, 23(10)*, 1142-1162. https://doi.org/10.1108/01443570310496607

Cambini, C., Croce, A., & Fumagalli, E. (2014). Output-based incentive regulation in electricity distribution: Evidence from Italy. *Energy Econ, 45*, 205–216. https://doi.org/10.1016/j.eneco.2014.07.002

Cambini, C., Fumagalli, E., & Rondi, L. (2016). Incentives to quality and investment: evidence from electricity distribution in Italy. *J Regul Econ, 49*, 1–32. https://doi.org/10.1007/s11149-015-9287-x

Carrington, R., Coelli, T.J., & Groom, E. (2002). International benchmarking for monopoly price regulation: The case of Australian gas distribution. *J Regul Econ, 21*, 191–216. https://doi.org/10.1023/A:1014391824113

Coelli, T.J., Rao, P.D.S., O'Donnell, C.J., & Battese, G.E. (2005). *An introduction to efficiency and productivity* *analysis*. Springer, New York, USA, 2nd Edition.

Dai, X., & Kuosmanen, T. (2014). Best-practice benchmarking using clustering methods: Application to energy regulation. *Omega, 42*, 179–188. https://doi.org/10.1016/j.omega.2013.05.007

Daraio, C., Diana, M., Di Costa, F., Leporelli, C., Matteucci, G., & Nastasi, A. (2016). Efficiency and effectiveness in the urban public transport sector: a critical review with directions for future research. *Eur J Oper Res, 248*, 1-20.

Devicienti, F., Manello, A., & Vannoni, D. (2017). Technical efficiency, unions and decentralized labor contracts. *Eur J Oper Res, 260*, 1129-1141.

Dolnicar, S., Grun, B., & Leisch, F. (2013). Required sample sizes for data-driven market segmentation analyses in tourism. *Journal of Travel Research, 53(3)*, 296-306. https://doi.org/10.1177/0047287513496475.

Erbetta, F., & Rappuoli, L. (2008). Optimal scale in the Italian gas distribution industry using data envelopment analysis. *Omega, 36*, 325–336. https://doi.org/10.1016/j.omega.2006.01.003

Ertürk, M., & Türüt-Aşik, S. (2011). Efficiency analysis of Turkish natural gas distribution companies by using data envelopment analysis method. *Energy Policy, 39*, 1426–1438. https://doi.org/10.1016/j.enpol.2010.12.014

Everitt, B.S., Landau, S., Leese, M., & Stahl, D. (2011). *Cluster Analysis, Quality and Quantity*. https://doi.org/10.1007/BF00154794

Farrell, M.J. (1957). The measurement of productive efficiency. *J Royal Statistical Society, Series A, 120(3)*, 253–90. https://doi.org/10.2307/2343100

Farsi, M., Filippini, M., & Kuenzle, M. (2007). Cost efficiency in the Swiss gas distribution sector. *Energy Econ, 29*, 64–78. https://doi.org/10.1016/j.eneco.2006.04.006

Formann, A.K. (1984). Latent class analysis: introduction to theory and application. Weinheim: Beltz.

Giannakis, D., Jamasb, T., & Pollitt, M.G. (2005). Benchmarking and incentive regulation of quality of service: An application to the UK electricity distribution networks. *Energy Policy, 33*, 2256–2271. https://doi.org/10.1016/j.enpol.2004.04.021

Goncharuk, A.G., & lo Storto, C. (2017). Challenges and policy implications of gas reform in Italy and Ukraine: Evidence from a benchmarking analysis. *Energy Policy, 101*, 456–466. https://doi.org/10.1016/j.enpol.2016.10.037

Greene, W. (2005). Reconsidering heterogeneity in panel data estimators of the stochastic frontier model. *J Econometrics, 126(2)*, 269-303. https://doi.org/10.1016/j.jeconom.2004.05.003

Haney, A.B., & Pollitt, M.G. (2009). Efficiency analysis of energy networks: An international survey of regulators. *Energy Policy, 37*, 5814–5830. https://doi.org/10.1016/j.enpol.2009.08.047

Jamasb, T., & Pollitt, M.G. (2003). International benchmarking and regulation: An application to European electricity distribution utilities. *Energy Policy, 31*, 1609–1622. https://doi.org/10.1016/S0301-4215(02)00226-4

Jamasb, T., & Pollitt, M. (2001) Benchmarking and regulation: International electricity experience. *Util Policy, 9*, 107–130. https://doi.org/10.1016/S0957-1787(01)00010-8

Johnson, A., & Kuosmanen, T. (2011). One-stage estimation of the effects of operational conditions and practices on productive performance: asymptotically normal and efficient, root-n consistent StoNEZD method. *J Prod Anal, 36(2)*, 219-230. https://doi.org/10.1007/s11123-011-0231-5

Lazarsfeld, P.F., & Henry, N.W. (1968). *Latent Structure Analysis*. Houghton Mifflin, Boston.

Llorca, M., Orea, L., & Pollitt, M.G. (2014). Using the latent class approach to cluster firms in benchmarking: An application to the US electricity transmission industry. *Oper Res Perspect, 1*, 6–17. https://doi.org/10.1016/j.orp.2014.03.002

López, J.J., Aguado, J.A., Martín, F., Muñoz, F., Rodríguez, A., & Ruiz, J.E. (2011). Hopfield-K-Means clustering algorithm: A proposal for the segmentation of electricity customers. *Electr Power Syst Res, 81*, 716–724. https://doi.org/10.1016/j.epsr.2010.10.036

Manello, A. (2017). Productivity growth, environmental regulation and win-win opportunities: the case of chemical industry in Italy and Germany. *Eur J Oper Res, 262*, 733-743.

Milligan, G.W., & Cooper, M.C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika, 50*, 159–179. https://doi.org/10.1007/BF02294245

Nieswand, M., & Seifert, S. (2018). Environmental factors in frontier estimation – A Monte Carlo analysis. *Eur J Oper Res, 265*, 133–148. https://doi.org/10.1016/j.ejor.2017.07.047

Orea, L., & Jamasb, T. (2017). Regulating heterogeneous utilities: A new latent class approach with application to the Norwegian Electricity Distribution Networks. *Energy J, 38*, 101–127. https://doi.org/10.5547/01956574.38.4.lore

Shephard, R.W. (1970). *Theory of cost and production function*. Princeton University Press, Princeton.

Silva, E., Macedo, P., & Soares, I. (2019). Maximum entropy: a stochastic frontier approach for electricity distribution regulation. *J Regul Econ., 55(3)*, 237-257. https://doi.org/10.1007/s11149-019-09383-y

Simar, L., & Wilson, P.W. (2000). Statistical inference in nonparametric frontier models: the state of the art. *J Prod Anal, 13(1)*, 49–78. https://doi.org/10.1023/A:1007864806704

Simar, L., & Wilson, P.W. (2007). Estimation and inference in two-stage, semi-parametric models of production processes. *Journal of Econometrics, 136(1)*, 31-64. https://doi.org/10.1016/j.jeconom.2005.07.009

Simar, L., & Wilson, P.W. (2011). Two-stage DEA: Caveat emptor. *J Prod Anal, 36(2)*, 205-218. https://doi.org/10.1007/s11123-011-0230-6

Tauchmann, H. (2012). Partial frontier efficiency analysis. *The Stata Journal, 12(3)*, 461-478.

Thanassoulis, E. (1996). A Data Envelopment Analysis approach to clustering operating units for resource allocation purposes. *Omega, 24(4)*, 463-476.

Tovar, B., Ramos-Real, F.J., & Fagundes de Almeida, E.L. (2015). Efficiency and performance in gas distribution. Evidence from Brazil. *Appl Econ, 47*, 5390–5406. https://doi.org/10.1080/00036846.2015.1047093

Zorić, J., Hrovatin, N., & Scarsi, G. (2009). Gas distribution benchmarking of utilities from Slovenia, the Netherlands and the UK: An application of data envelopment analysis. *South East Eur J Econ Bus, 4*, 113–124. https://doi.org/10.2478/v10033-009-0008-1

APPENDIX A.

Table A.1 reports the standard operational unit cost and the expected efficiency gains (X factor) in the Italian regulation over the period 2000-2020.

Table A.1. Summary of regulatory deliberations (2000-2020).

Deliberation Nr.	Tariff Period	Years	Standard operational unit cost [€/customer]	X factor
237-00	I	2000-03	n.a.	
170-04	II	2004-08	122.13	5%
159-08	III	2009-12	Between 56.46 for small firms with low customer density, and 39.30 for large firms with high customer density	Small 5.4%; Medium 4.6%; Large 3.2%
367-14 & 775-16	IV	2014-19 (standard operational unit costs are given for the year 2017 as values changed over the period)	Between 51.35 for small firms with low customer density, and 24.40 for large firms with high customer density	Small 2.5%; Medium 2.5%; Large 1.7%
570-19	V	2020-25	Between 43.59 for small firms with low customer density and 26.55 for large firms with high customer density (if public tender not completed), and between 33.68 for smaller concessions with low customer density and 26.55 for larger concessions with high customer density (if public tender completed)	Small 6.59%; Medium 4.79%; Large 3.53%

Table A.2 shows the results obtained when applying the Calinski-Harabasz rule and the Duda-Hart rule, to select the optimal number of clusters in the HCP. The selected number of clusters is three, as it corresponds to the highest values of both rules' statistics.

Table A.2. Stopping rules. Calinski-Harabasz rule (panel A) and Duda-Hart rule (panel B).

Number of Clusters	Calinski-Harabasz rule	Duda-Hart rule	
	Pseudo-F	Je(2)/Je(1)	Pseudo T-squared
1	-	0.77	122.55
2	122.55	0.72	125.39
3	125.13	0.80	69.81
4	108.52	0.77	68.51
5	101.38	0.64	49.93
6	114.56	0.47	77.99
7	123.32	0.51	64.31

Table A.3 reports the Akaike's (AIC) and the Bayesian (BIC) information criteria for the identification of the more appropriate number of clusters in the LCM. As both AIC and BIC show the largest improvement when moving from three to four classes, the selected number is four.

Table A.3. Choice of the number of classes: AIC and BIC criteria.

Number of classes	**AIC**	**BIC**
1	675.28	695.48
2	668.15	720.68
3	637.89	718.69
4	599.01	708.10
5	593.60	730.97

Table A.4 shows the LCM estimation results with four classes. Results indicate that as output (*customers*) rise, *opex* increases. As expected, also *volumes* of gas served increase *opex* in all classes but for Class 3. As for the latter, although the sign is negative, the same variable does not have a significant effect on *opex*. *Customer density* was expected to have a negative, significant effect on all classes. This appears to be true only for Class 3. While the same variable has no significant effect in Class 1 and in Class 2, the effect on *opex* is significant and positive in Class 4. Notably, Class 4 has the highest *customer density* across classes, pointing to increasing difficulties encountered when serving extremely dense areas. Finally, *altitude* was expected to have a positive, significant effect on all classes. This is confirmed, however, only for relatively lower altitudes (for Class 3 and Class 4). Note that, the contextual variable *output density* was dropped because of collinearity.

Table A.4. LCM estimation results with four classes.

	Class 1	**Class 2**	**Class 3**	**Class 4**
	coefficient (standard error)	Coefficient (standard error)	coefficient (standard error)	coefficient (standard error)
log (*customers*)	0.866*** (0.135)	0.557*** (0.077)	0.902*** (0.38)	0.506*** (0.044)
log (*volumes*)	0.435*** (0.139)	0.365*** (0.074)	- 0.063 (0.039)	0.410*** (0.045)
customer density	- 0.816 (1.035)	- 0.899 (0.628)	- 1.309*** (0.399)	1.051* (0.608)
Altitude	0.000 (0.000)	0.000 (0.000)	0.001*** (0.000)	0.000* (0.000)
const.	- 3.353*** (1.204)	1.817*** (0.601)	6.320*** (0.324)	2.395*** (0.440)
Sigma	0.204*** (0.039)	0.361*** (0.026)	0.102*** (0.015)	0.092*** (0.014)
prior class probability	0.095 (0.025)	0.603 (0.041)	0.219 (0.034)	0.083 (0.016)

[a] Note: ***, **, *: significant at 1%, 5% and 10%, respectively.

APPENDIX B.

To directly compare the results of the size segmentation approach with the HCP and LCM methods, the latter two were applied using a single separating variable (*size*) and assuming a pre-fixed number of classes (3). The resulting allocation is as follows:

- Size-segmentation: 11% of the observations are Large, 27% are Medium, and 62% are Small;
- HCP: 8% of the observations are in Class 1-HCP, 88% are in Class2-HCP, and 5% are in Class3-HCP;
- LCM: 30% of the observations are in Class 1-LCM, 49% are in Class2-HCP, and 21% are in Class3-HCP.

As illustrated in **Table B.1,** the HCP partition tends to aggregate most DSOs in one class, separating out DSOs characterized by bigger size. On the contrary, no evident overlap emerges between size segmentation and LCM, where each class includes DSOs of all sizes.

Table B.1. Direct comparison of size-segmentation with HCP and LCM partitions assuming one separating variable (*size*) and three classes.

	Class 1-HCP	**Class 2-HCP**	**Class 3-HCP**	**Class 1-LCM**	**Class2-LCM**	**Class 3-LCM**
Large	50%	0%	50%	29%	59%	12%
Medium	7%	93%	0%	29%	52%	19%
Small	0%	100%	0%	32%	47%	21%