The Returns to Scale Assumption in Incentive Rate Regulation

Rajiv Banker
Fox School of Business and Management, Temple University,
1801 N Broad St, Philadelphia, PA 19122, USA
banker@temple.edu

Daqun Zhang
College of Business, Texas A&M University – Corpus Christi,
6300 Ocean Dr, Corpus Christi, TX 78412, USA
david.zhang@tamucc.edu

Last revised: March 29, 2016
The Returns to Scale Assumption in Incentive Rate Regulation

Abstract

This paper challenges a common (though not universal) practice of maintaining the assumption of constant or non-decreasing returns to scale (CRS or NDRS) for benchmarking of local monopolies in incentive rate regulation. We consider the situation when firms get larger by acquiring customers that are costlier to serve, but detailed data on different types of customers are not available to regulators. If the output is specified only in terms of total number of customers or total sales (in monetary or quantity terms) without explicitly capturing the impact of different types of customers who may have different marginal costs, then we show that the estimated production function will appear to exhibit variable returns to scale (VRS) with a region of increasing returns followed by a region of decreasing returns to scale even when the true production technology has CRS or NDRS. This apparent region of decreasing returns is only an illusion caused by the mis-specification of the variables in the production function. We demonstrate the magnitude of the distortion in the efficiency of individual firms with extensive Monte Carlo simulation experiments. Our simulation results indicate that in most scenarios the VRS model performs much better than the CRS or NDRS models or a NDRS model that corrects for customer mix only in the second stage. These insights also caution scholars studying efficiency of different organizations in a broad range of contexts beyond regulation. When detailed data on all inputs and outputs are not used, the VRS production model needs to be employed without imposing any restrictions on returns to scale such as CRS or NDRS.

Keyword

Incentive Regulation, Production Function, DEA, Returns to Scale, Electricity Distribution
1. Introduction

Data Envelopment Analysis (DEA) is an excellent example of an operations research methodology that has enjoyed a wide range of applications in practice. Applications of DEA for benchmarking and incentive rate regulation are now common in several utility industries in many countries around the world that affects billions of dollars annually. Regulators responsible for electric power around the world are increasingly relying on estimated production or cost frontiers to benchmark firms and structure efficient incentive contracts for price/revenue cap regulation; the most common estimation method has been DEA (e.g. Jamasb and Pollitt 2001; Haney and Pollitt 2009). The principal advantage cited for this estimated frontier approach over the traditional reference unit method is that detailed engineering and cost data are not required for an individual firm’s operations. In this paper, we document a serious flaw in this simplified approach when returns to scale (RTS) assumptions are imposed in the estimation of the production function.

Economic theory suggests that increasing returns to scale must prevail in the electricity distribution industry to justify local monopolies. Consequently, several regulators have assumed constant or increasing returns to scale (CRS or IRS) in specifying the estimated production function for benchmarking regulated firms. In this paper, we document that such an assumption is invalid if the regulator has not.

---

1 For instance, OFGEM (2012) reports that electricity customers in the U.K. pay on average £ 4.4 billion annually for electricity distribution, which accounts for 17% of an average domestic customer’s bill. The monthly average exchange rate in December 2012 was approximately GB£ 1 = US$ 1.6.

2 For example, regulators of electricity transmission or distribution in Austria, Belgium, Brazil, Finland, Germany, Netherlands and Norway at various times have imposed constant or non-decreasing returns to scale assumption (CREG 2011; ACCC 2012; Nillesen and Pollitt 2007; Schweinsberg et al. 2011; Migueis et al. 2012; Bogetoft and Otto 2010; ANEEL 2010).
specified outputs and inputs in sufficient detail to capture the differences in the marginal costs to serve different types of customers. Estimated frontiers based only on aggregated outputs distort the true efficiency of individual firms, because of model misspecification errors if CRS or IRS assumptions are imposed. Our Monte Carlo experiments document that empirical models with a more flexible assumption of variable returns to scale (VRS) generate more precise estimators of the underlying true production function than models with constant or non-decreasing returns to scale (NDRS) assumption even when the true production function exhibits increasing returns to scale.

Since the late 1980s, regulation of the electric utility industry has been reformed in many developed and developing countries. Traditional cost recovery regulation has been replaced in many countries by incentive regulation in order to improve the operating efficiency of electric utilities and maximize social welfare (Jamasb and Pollitt 2001). To motivate firms to reduce costs, regulators set prices or revenues based on the estimated costs of best practice firms. Consequently, frontier-based benchmarking methods, including Data Envelopment Analysis (DEA), Corrected Ordinary Least Squares (COLS) and Stochastic Frontier Analysis (SFA), are now widely used in the practice of electricity rate regulation worldwide.

DEA is a nonparametric mathematical programming method (Charnes et al 1978; Banker et al. 1984; Banker 1993), while both COLS and SFA are parametric econometric method (Aigner et al 1977). Compared to COLS and SFA, DEA only requires general assumptions on the production possibility set, and is more robust to
functional specification errors. Banker (1993) shows that DEA estimators of the best practice production function exhibit a desirable asymptotic property of consistency and are maximum likelihood estimators if the deviation of actual output from the frontier comes from a distribution with a monotone decreasing probability density function -- assumptions that are much weaker than those required for similar properties for the COLS or SFA estimators. Bogetoft (1994) shows that a variety of DEA frontiers contain all relevant information about managerial effort for optimal contracting under different distribution conditions. Ruggiero (1999) shows that cross-sectional stochastic frontier models do not decompose noise and inefficiency accurately and the simulation results suggests that deterministic approach outperform stochastic frontier models in nearly all experimental settings. Ondrich and Ruggiero (2001) further show that the COLS deterministic model performs as well as the stochastic frontier model because stochastic frontier models cannot provide a reliable absolute measure of inefficiency. Because of the nice features of DEA, many regulators for the electricity industry have shown a strong preference for DEA over COLS or SFA in practice (Haney and Pollitt 2009; Bogetoft and Otto 2010).

When using DEA models to estimate the production function, many regulators have appealed to theoretical considerations to justify a natural monopoly and invoked the non-decreasing returns to scale (NDRS) assumption for the regulation of electricity distribution firms (Schweinsberg et al. 2011; Nillesen and Pollitt 2004; Agrell and Bogetoft 2003, 2004, 2007; CEPA 2003; Honkapuro 2008). In this paper we show that this theoretical argument does not hold unless all outputs and inputs are
included in the model and measured without error, an assumption that is unlikely to hold in practice. As a result, severely distorted estimation of the production frontier may occur when the assumption of CRS or NDRS is maintained, and even threaten the long-term financial viability of some large electricity firms who are compensated based on the distorted estimates.

A common situation where this occurs depends on how some firms grow. For instance, if the regulated price is not based on the cost of serving different customers, electricity distribution firms may first install electric lines in those locations where they can serve customers who are easiest to reach. As firms grow larger, they may need to expand their network gradually to reach customers that are costlier to serve. Thus, cost per customer may appear to increase rather than decrease when the firm grows. As a result, CRS or NDRS DEA models with only the aggregate number of customers as output may yield very low efficiency scores for electricity distribution firms that have grown by reaching customers that are hard to serve. In such a case, the low efficiency scores should not be attributed to managerial slack, but to the failure of simple empirical models to capture the heterogeneity of customers.

The theoretical production function for electricity distribution firms requires rather complex mathematical representations regarding how multiple inputs are transformed into multiple outputs, given a variety of contextual variables and input prices. However, it is not feasible to identify and measure all relevant outputs and characteristics of customers in practice. By necessity, regulators often represent costs of electricity firms as a parsimonious function of a few outputs, such as total number
of customers, total electricity sold and length of network. Regulators use such simple model specifications also because they have only a limited number of observations, and more complex specifications cannot be reliably estimated based on only a few degrees of freedom.

Our Monte Carlo experiments to examine this situation show that if the output is specified only in terms of total number of customers or total sales without explicitly capturing the impact of different types of customers in the production function, the estimated production function will not comply with the NDRS assumption. In fact, a more general specification of the production function that allows VRS performs much better in estimating the true production function even when the true data generating process (DGP) shows IRS. A simple second stage analysis that regresses estimated efficiency scores on customer mix to control for the impact of customer heterogeneity also cannot correct the model specification bias if the CRS or NDRS assumption is imposed in estimating the production function in the first stage.

The remainder of this paper is organized as follows. In Section 2, we review regulatory practice for specifying returns to scale in estimating the production function and also the mixed empirical evidence reported in prior academic research literature. In Section 3, we present a simple theoretical model to motivate why the empirical production or cost function may display decreasing returns to scale even when the true production function has only increasing returns to scale. In Section 4, we conduct Monte Carlo simulation experiments to evaluate the performance of estimation methods with different assumptions of returns to scale. In Section 5, we
illustrate these insights in the context of both the U.S. and Brazilian electricity distribution industries. We conclude with a discussion of our principal findings and implications for practice in Section 6.

2. Literature Review

To justify the existence of regional monopolies, economic theory suggests that constant or increasing returns to scale prevail in such industries that lead to lower average costs (scale economies) for larger firms (Joskow 1997, 2006).\(^3\) If the cost structure has large proportion of fixed cost and small marginal costs, then average costs should decrease as a firm grows larger and scale economies prevail to justify a monopoly serving the entire regional market. If, instead, decreasing returns to scale prevail, their average costs can be reduced and social welfare increased if the regulator breaks up the monopoly into smaller firms that operate at the most productive scale size (Banker 1984).

Appealing to these intuitive arguments from economic theory and arguing that the cost structure in the electricity transmission and distribution industries is consistent with increasing returns to scale, many regulators and research studies have employed Data Envelopment Analysis (DEA) models with the additional imposition of a CRS or NDRS assumption for the empirical estimation of the cost or production function (CREG 2011; ACCC 2012; Nillesen and Pollitt 2007; Schweinsberg et al. 2011; Migueis et al. 2012; Bogetoft and Otto 2010; ANEEL 2010). They argue against the more general VRS assumption because it allows decreasing returns to scale for large

\(^3\) Alternative conditions, such as the existence of scope economies, may also exist to justify natural monopolies.
firms, which is apparently inconsistent with economic theory (CEPA 2003; Agrell and Bogetoft 2007). For example, the Dutch Electricity Regulatory Service (DTe) used a CRS DEA model, assuming that electricity distributors are able to reach the optimal scale by merging or de-merging (Nillesen and Pollitt 2004). Agrell and Bogetoft (2004) recommend that Norwegian and Swedish regulators should employ a NDRS DEA model that allows smaller distributors with limited opportunities of merger or network growth to continue to operate at small scale, but requires larger distributors to operate at most productive scale size that may be smaller than their current size if decreasing returns to scale exist. Similarly, CEPA (2003) recommends that the British regulator, Ofgem, “should not explicitly allow for diseconomies of scale … large firms should be encouraged to mimic more efficient smaller firms by disaggregating their activities appropriately (p.62).”

In a consulting report for the German electricity regulator, Agrell and Bogetoft (2007) suggest that a CRS specification is “uncontroversial” and robust for the purpose of long term cost efficiency, because no concession areas are predefined in Germany. In contrast, based on the same logic, the Finnish regulator used the VRS model in order to compensate for the possible scale inefficiency of small and large companies, because companies cannot choose their scale of operation (Honkapuro 2008). The use of all of VRS, CRS or NDRS models is thus supported in the literature depending on the flexibility in adjusting the size of the firms. The conclusion seems to be that if firms cannot change their scale of operation, then the VRS model seems to be preferred, if firms can adjust their size, then the CRS model seems to be preferred,
and if there exist small markets where only a small firm can operate, then the NDRS model seems to be recommended. In this paper, we show that the logic behind this widely-accepted conclusion is not valid if all outputs and inputs are not specified correctly.

Mirroring the differences in the assumption about returns to scale in regulatory practice, prior academic research examining the electricity transmission and distribution industries provides mixed empirical results for returns to scale. By necessity, all of these empirical studies specify relatively simple and parsimonious set of outputs to describe the production or cost function. Neuberg (1977) estimates Cobb-Douglas cost functions for U.S. electricity distributors and finds that increasing returns to scale do not prevail for the entire output range. We draw on the innovative analysis by Roberts (1986) who develops three different measures to capture network density and size to analyze the cost structure of firms that have geographically dispersed customers. His empirical results show that only increasing output (electricity sold) to existing customers leads to economies of scale. There are no significant economies from increased customer density and increased geographic size of the service area because cost reduction due to an increase in output is offset by an increase in average costs due to an increase in the number of customers. This suggests that mergers and acquisitions by combining service territories do not necessarily lead to economies of scale.

Several studies report regions of non-increasing returns to scale for large firms.

---

4 Roberts (1986) states that "increased demand by existing customers, demands by new customers within the firm’s service area, or an expansion of the service area can all lead to increased output but each can have a different impact on unit cost and thus lead to a different measure of scale effects (p.378)."
Nelson and Primeaux (1988) study municipally owned electric utilities in the U.S. and find that scale economies holding output per customer fixed have been exhausted by the larger firms in the sample. Kaserman and Mayo (1991) study privately owned electric utilities in the U.S. and conclude that “for the output of each stage in isolation, … economies of scale are exhausted well within the range of representative outputs (p.497)”. In particular, the estimated economies of scale for distribution stage get exhausted around 5,000 GWH, which is less than the sample mean of about 7,000 GWH.

In addition to the studies that use data for electric utilities in the U.S., a number of papers that use data from other countries also document the empirical evidence of exhausted economies of scale for large distribution firms. Salvanes and Tjotta (1994) study the Norwegian electricity distribution industry and find that economies of scale are not present but economies of output density are. Hjalmarsson and Veiderpass (1992) find significant differences in scale efficiencies between rural and urban electricity retail distributors in Sweden. Filippini (1998) finds that scale economies exist only for small and medium-sized electricity distribution utilities in Switzerland, but economies of density prevail for most output levels. Pacudan and Guzman (2002) find that large distribution firms do not benefit from economies of scales in Philippines. The six largest electricity distribution firms out of fifteen in the sample are found to operate in the region of decreasing returns to scale. Jara-Díaz et al (2004) document that “returns to scale get exhausted after a proportional increase of 7% in production at the mean (p.1006)”, though these Spanish electricity firms exhibit
slightly increasing returns to scale at the mean. Arocena (2008) documents evidence of decreasing returns to scale for the largest electricity utilities in Spain when current mix is kept constant. Arcos and Toledo (2009) find that the estimated total cost function is convex with respect to electricity sold, suggesting that diseconomies of scale may exist in Spanish electric utility industry.

Furthermore, a number of papers investigate the average cost function and the minimum efficient scale (MES) in electricity industry (Neuberg 1977; Yatchew 2000; Giles and Wyatt 1993; Salvanes and Tjotta 1994; Filippini 1996, 1997; Kwoka 2005). These studies document substantial evidence that the transmission and distribution companies have a U-shaped average cost function and MES can be reached even at small or medium size. Neuberg (1977) shows that the long run average cost curve of U.S. municipally owned distributors appears to be U-shaped rather L-shaped and the optimum distribution unit size may range from 85,000 to 288,100 customers. Yatchew (2000) estimates the average cost function (average cost per customer) of approximately 300 municipal electric distribution utilities in Canada. He finds the MES is around 20,000 customers, while larger firms exhibit constant or decreasing returns to scale. Therefore, he concludes that horizontal mergers between distributors are not likely to produce substantial scale economies in the operation of their business.

any output in the range 500-3500 gwh is essentially consistent with minimum AC”. Correspondingly, the implied MES of these distribution firms is around 30,000 customers (Yatchew 2000). Similarly, Salvanes and Tjotta (1994) analyze 100 Norwegian distributors for whom annual electricity sold ranges from about 11 Gwh to 7500 Gwh. They find that “… optimal size comprises plants serving about 20,000 customers and is relatively independent of the level of Gwh produced (p.35).” Although Filippini (1996, 1997) finds increasing returns to scale throughout his sample of 39 Swiss distributors, the largest distributors have electricity sold of only about 300 Gwh, which are smaller than the MES in Giles and Wyatt (1993). Therefore, Yatchew (2000) concludes that Filippini’s findings of increasing returns to scale throughout his sample may not be inconsistent with other studies. Kwoka (2005) estimates the average cost function of 543 electricity utilities in the U.S. in 1989 and finds the average cost curve is U-shaped with respect to electricity sold when holding usage per customer and density (customers per mile) constant. Kelly (2001) criticizes the naïve reasoning that electricity distributors should operate approximately at the estimated MES, because it implicitly assumes that services provided by the different electric distribution systems are homogenous, and can attribute the cost difference between electricity distributors only to their physical size.

Overall, the empirical evidence suggests that the assumption of increasing (or non-decreasing) returns to scale may not be empirically valid for large electricity firms when output is measured in aggregate variables such as number of customers or quantity of electric power sold. We posit here that this empirical finding arises
because electricity firms operate in different service areas and marginal costs for customers in different market segments are different. Simple output measures such as number of customers and electric power sold are not adequate to capture the complexities of how costs arise for power transmission and distribution.

In the next section, we show that this empirical evidence of apparent diseconomies of scale can occur when, in fact, increasing returns to scale prevail, but firms acquire costlier customers to grow large and a relatively parsimonious production or cost function with only a few input and output variables is employed for estimation.

3. Model Misspecification and Returns to Scale

3.1 Specification of Production Function

The production function for electricity distribution describes how a firm transforms multiple inputs into multiple outputs, given a variety of contextual variables and input prices. Different outputs with different complexity may require different quantities of the various inputs, which may also vary depending on prices and contextual conditions as well as levels of other outputs and capacities. Hence, the specification of the true production function may require more complex mathematical representations than the simple direct functional relationship between total operating costs and a few outputs specified by many regulators.

Examples of complexities in production relationships abound in electricity distribution. Some electricity firms serve customers dispersed over large geographical regions, including rural and semi-urban areas. These areas often have irregular
topography, roads in poor conditions, and other adverse conditions that require special vehicles and tools. To maintain the distribution network or to meet and deal with customer demands and emergencies within acceptable time norms, the distribution firms have to hire more teams that are scattered throughout the service area. The distances and difficulties involved for both maintenance and commercial service operations make such customers much more expensive to serve.

There are problems of a different nature in the major metropolitan areas where customers require higher quality of electricity delivery service, such as lower frequency of interruptions and lower average duration of an interruption. To achieve higher levels of quality, distributors also need to have more teams located in different parts of a city due to traffic congestion and related transportation difficulties that increase the average time of travel. Unlike in rural areas, difficulties in parking the service vehicles, especially in high traffic flow areas or near shopping centers, often requires scheduled intervention by local transit authorities.

Maintenance tasks on networks in large cities are also undermined by the high loading of the poles of the concession area with fully utilized networks (including telephone services and cable television). Underground networks are common in commercial and densely populated areas. While maintenance is less frequent for this type of network, the cost per intervention is much higher. Therefore, it is much more expensive to carry out maintenance operations for an underground network than an overhead network, although the frequency of problems is lower for an underground network. Additionally, when electricity is distributed to many areas with fast
expanding “informal urban settlements” (slums), average costs may actually increase rather than decrease with size and density because of the poor infrastructure, rugged topography and, in many cases, high rates of violence encountered in the slums.

However, it is rarely feasible to identify and measure all such relevant outputs and location differences in practice. Instead, prior empirical research and current regulatory practice often represents the cost function as a simple and parsimonious function. For example, regulators in countries, such as UK, Brazil and Finland only use operating expenses (OPEX) as the single input, and exclude capital expenditure (CAPEX) from benchmarking because CAPEX is considered to be lumpy and not controllable in the short term (Jamasb et al. 2004). Some other regulators, such as in the Netherlands and Norway use either total expenses (TOTEX) or both OPEX and CAPEX separately as input measures to allow potential tradeoffs between these two types of costs (Jamasb and Pollitt, 2003; Jamasb et al., 2004; Agrell et al., 2005). Surprisingly, OPEX is typically not analyzed into detailed cost components to capture how different costs are affected differently by different outputs (Banker and Johnston 1993). The total number of customers, total electricity delivered and service area (or network size/length) are the three most frequently used measures of outputs (Cullmann and Hirschhausen, 2008; Zhang and Bartels, 1998; Bagdadioglu et al., 1996; Berg et al. 2005; Bó and Rossi, 2007; Estache et al., 2004; Pacudan and Guzman, 2002; Perez-Reyes and Tova, 2009; Jamasb and Pollitt, 2001). Regulators

---

5 Some studies of the electricity transmission industry use peak load as an additional output measure (Neuberg, 1977; Salvanes and Tjotta, 1994; Burns and Weyman-Jones, 1996). Several studies divide number of customers and electricity transmitted into two categories: high and low voltage customers (or connections), and high and low voltage electricity transmitted (Agrell et al. 2005; Hjalmarsson and Veiderpass 1992).
and researchers often use such a simple model specification because of only a few degrees of freedom with limited number of observations. A complex specification of the empirical model requires many parameters to be estimated, and thus it may not be possible to reliably estimate a complex model.

3.2 Imposing Structure on the Empirical Model

An electricity firm needs large capital investments in its infrastructure such as distribution lines, substations and transformers, to deliver electric energy to its customers, while the marginal costs of serving an extra customer are relatively smaller. Such a cost structure with large fixed costs and small marginal costs leads to a decrease in average costs when a distribution firm serves more customers within its existing service area. This intuition can be shown in a simple analytical model. Suppose the cost function of a distribution firm is defined by,

\[ C(y) = F + ky, \]

where \( y \) is the number of customers served; \( F \) is the fixed costs related to various capacity resources; and \( k \) is the marginal cost of serving customers. Given that fixed costs exist \( (F > 0) \), conventional intuition suggests that the average cost, \( C(y)/y = F/y + k \), goes down when the number of customers increases because fixed costs \( F \) are spread over a larger customer base. Appealing to this intuition, many

---

6 To model the impact of contextual factors on costs of electric distribution firms, prior studies also include a number of contextual variables in a second stage analysis such as customer density measures, output density measures and some weather factors (Estache et al., 2004; Hattori, 2002; Berg et al., 2005; Agrell et al., 2005). To justify the two stage analysis, several assumptions are required that may be violated in this case (Banker and Natarajan 2008). This includes the assumption that the impact of contextual variables on costs is functionally separable from the outputs. While this separability assumption makes the estimation convenient, it may not hold in the regulatory context for measures of density and complexity.

7 For example, with only 14 electricity distribution network operators, the UK regulator uses a single composite measure of output, placing 50% weight on the number of customers, 25% on electricity distributed and 25% on network length (CEPA 2003; Pollitt 2005).
regulators prefer to impose a structure of constant returns to scale or non-decreasing returns to scale on the empirical production function as discussed in previous section. However, this requires a key assumption that the marginal cost of serving all customers is the same, which is unrealistic in practice.

To consider heterogeneity of customers in a simple and transparent model, we assume there are only two types of customers in the market: regular customers and complex customers. Then the true cost function is given by

\[ C = F + k_1 y_1 + k_2 y_2, \text{ with } k_2 > k_1 > 0, F > 0. \]

(1)

Here \( y_1 \) and \( y_2 \) are the number of regular and complex customers respectively. The complex customers are more costly to serve as embodied in \( k_2 > k_1 \). The marginal costs \( k_1, k_2 \) are constant, but due to the fixed costs \( F > 0 \) being spread over a larger base as the firm grows larger, economies of scale exist.

Because of information asymmetry, the distribution firm cannot identify the types of customers in advance, but has information about the proportion of complex and regular customers in the regional market. If the electricity price set by regulators does not recognize the different costs of serving different types of customers, then the electricity distribution firm has the incentive to serve the customers in locations that are easiest to reach. In other words, the distribution firm will choose the areas with higher proportion of regular customers to maximize their profits. In fact, the evolution of the U.S. electricity industry reflects this pattern. In the early days of electrification, investor-owned utilities in the U.S. preferred to operate mainly in

---

8 For simplicity in this illustration, we assume that each customer consumes the same amount of electric power so that the sales revenue per customer is a constant.
urban areas, because the average cost of serving customers in densely populated areas is lower and wealthy urban customers tend to use more electric power. As a result, many consumer-owned electric cooperatives had to be established in the 1930s to provide electric service to remote rural areas (RAP 2011). These cooperatives still serve approximately 12.8% of the U.S. electric customers, mainly in rural areas.9

If the firm grows larger, perhaps because of regulatory or national policy directive to provide electricity to all, including the less advantaged segments of society, the proportion of customers that are costly to serve is likely to increase. In other words, growth in size is often associated with a change in the customer mix that leads to an increase in average costs. To examine the impact of model misspecification due to the use of simplified aggregated output measures, we consider the problem if the regulator measures output as simply the total number of customers \( y = y_1 + y_2 \). Then the average cost per customer is given by

\[
C(y)/y = (F/y) + k_1(y_1/y) + k_2(y_2/y)
= (F/y) + k_1 + (k_2 - k_1)\phi
\]

where \( \phi = y_2 / y \) is the proportion of customers who are more costly to serve. If all customers were homogeneous and \( k_2 = k_1 \), then \( C(y)/y \) would decrease with \( y \). However, if heterogeneity of customers makes the marginal cost difference \( k_2 - k_1 \) large and the proportion \( \phi \) of complex customers increases (\( \partial \phi / \partial y > 0 \)) as the firm grows larger, then the conventional intuition need not hold. This is seen by differentiating equation (2) with respect to \( y \):

---

9 Similar evolution patterns appear also in other network industries such as cable television and telecommunication.
$$\frac{d(C/y)}{dy} = -(F/y^2) + (k_2 - k_1) \frac{\partial \phi}{\partial y}$$  \hspace{1cm} (3)

If the second term in equation (3) above is sufficiently large and positive to offset the negative first term, then average costs will appear to increase with size, consistent with much of the prior empirical research but contrary to conventional intuition. This apparent empirical pattern may be interpreted as evidence that decreasing returns to scale prevail, but such an inference would be incorrect.

The apparent increase in average costs for larger firms is only an illusion in this setting. The true technology has an IRS cost structure as embodied in equation (1). The error arises because the output is incorrectly measured by the regulator as only the total number of customers instead of employing detailed information on customers with different marginal costs. The problem would disappear if heterogeneous customer groups are identified separately. We chose to present this analysis with only two customer groups in order to make the intuition transparent. However, in reality, we have many different types of customers demanding many different levels of quality of service, and we have many different types of locations to serve that have different marginal costs. All such differences would need to be recognized in the specification of the production (or cost) function to ensure that it displays IRS. But, in practice, such details are rarely available, and most regulators and research studies have specified highly aggregated output measures. Consequently, it is inevitable that empirical analysis finds evidence of both IRS and DRS. Under these conditions, it is clear that imposing the IRS (or NDRS) assumption is not valid, and it seems likely that the VRS model would provide more accurate estimates of efficiency if only the
aggregate output measures can be used. In the next section, we explore these issues using simulation studies.

Is it likely that the proportion of customers who are more costly to serve will increase as the firm grows larger? Economic theory indicates that if the prices are set the same for all customers without differentiating the prices to reflect differences in the costs to serve different customers then a firm will likely choose the customers that are less costly to serve. Therefore, while it may be true that the full and complete specification of the production relationship between all inputs, outputs and contextual variables may exhibit non-decreasing returns to scale (NDRS), it is not appropriate to impose the NDRS assumption on the much simpler abstraction employed for empirical purposes. In fact, the simple empirical abstraction must be estimated imposing minimum structure and in the most flexible functional form.\textsuperscript{10} For DEA models, it requires the use of the variable returns to scale (VRS) model, such as that provided by Banker et al. (1984), to ensure more accurate prediction of costs which is essential for incentive regulation.

4. Simulation Study

The conceptual analysis in the previous section showed that the estimated

\textsuperscript{10} The specification error from imposing NDRS assumption instead of specifying a more general VRS function can also be seen to be related to omitted variable bias. Consider a true cost function given by

\[ \ln c = \alpha_0 + \alpha_1 \ln y_1 + \alpha_2 \ln y_2 + \epsilon, \]

where \( y_1 \) and \( y_2 \) are the two output variables, \( \alpha_1, \alpha_2 > 0, \alpha_1 + \alpha_2 < 1 \), and \( \epsilon \) represents the error term in the estimation. The returns to scale are given by \( \alpha_1 + \alpha_2 \) which is less than 1, signifying increasing returns to scale. Suppose the outputs \( y_1 \) and \( y_2 \) are related as, \( \ln y_2 = \gamma \ln y_1 + \eta \), where \( \eta \) is random noise and \( \gamma > 1 \). If the empirical specification omits the variable \( y_2 \) and instead assumes a simple relationship

\[ \ln c = \beta_0 + \beta_1 \ln y_1, \]

then the expectation of the estimated coefficient \( \hat{\beta} \) is \( \alpha_0 + \gamma \alpha_2 \). Thus, if \( \gamma \) is sufficient large (>\( (1-\alpha_2)/\alpha_1 \)), \( E(\hat{\beta}) \) can be greater than one signifying decreasing returns to scale. In other words, while the true production function exhibits NDRS, we cannot assume the same for the simpler empirical specification because of such biases resulting from omitted output variables.
production function will display VRS even when the true function has IRS if the regulator specifies only aggregate output measures. In this section, we employ a simulation study to assess the magnitude of distortion in estimating the efficiency of different firms due to this misspecification. In addition, we evaluate whether the attempt by some regulators to correct for the misspecification in a second stage regression is valid and effective.

4.1 Experimental Design

For the data generating process (DGP), we assume that there are two types of customers: regular customers with low marginal costs and complex customers with high marginal costs. The true cost function is:

$$\ln(c) = \ln(k_1y_1^{\alpha_1} + k_2y_2^{\alpha_2}) + \varepsilon \tag{4}$$

where $y_1$ and $y_2$ are the two outputs representing regular and complex customers respectively; $c$ represents the costs of electricity distribution firms, such as operating expenses or total expenses; both $\alpha_1$ and $\alpha_2$ are less than 1, implying non-decreasing returns to scale. $k_2$ is greater than $k_1$, implying that the marginal costs of complex customers are likely higher than that of regular customers. $\varepsilon = \ln(\theta) \geq 0$ reflects the logarithm of true inefficiency, which is defined by,

$$\ln(\theta) = \ln(c) - \ln(k_1y_1^{\alpha_1} + k_2y_2^{\alpha_2}) \tag{5}$$

We assume $\varepsilon$ is randomly generated from a half-Normal distribution $|N(0, \sigma^2)|$.

If the regulator specifies the output only as aggregate number of customers and the empirical estimation does not discriminate between $y_1$ and $y_2$ then costs are related only to the total number of customers, $y = y_1 + y_2$. We estimate the cost function
using the following two parametric COLS models and four different DEA models, for which the input is the observed \( c \), and the output is the aggregate output \( y \): Cobb-Douglas COLS model, Translog COLS model, BCC VRS model, CCR CRS model, NDRS DEA model and two stage NDRS DEA model. For the two stage NDRS DEA model, we use \( \hat{\phi}_i = y_{2i} / y_i \) as a contextual variable to adjust inefficiency scores in the second stage regression analysis.$^{11}$

Next, to test the performance of different models with detailed information on customer types, we employ both the number of regular customers \( y_1 \) and the number of complex customers \( y_2 \) in the estimation model as outputs, and observed costs as input. We estimate the following COLS and DEA models to estimate the cost function: Cobb-Douglas COLS model, Translog COLS model, BCC VRS model, CCR CRS model, and NDRS DEA model.$^{12}$

### 4.2. Monte Carlo Simulation

We conduct 2,000 Monte Carlo experiments. For each trial, we draw a sample of size \( N \), where the sample size \( N \) takes any integer value over the interval \([50, 200]\) with equal probability. Next, we generate \( \alpha_1 \) and \( \alpha_2 \) from independent uniform distributions on the interval \([0.85, 0.95]\). We set \( k_1 = 1 \) and generate \( k_2 \) from an independent uniform distribution on the interval \([2, 12]\). We draw the total output

---

$^{11}$ In reality, the regulators may not know the exact proportion of complex customers but may have several contextual variables instead that can help them infer the proportion of complex customer. For simplicity, in this simulation experiment, we use \( y_{2i} / y_i \) directly assuming the regulators can exactly measure the proportion of regular customers without any errors-in-variables issue. This is the best scenario favoring the regulators’ two stage model, which is equivalent in our model to the assumption that direct and perfect information on customer types is available for regulator’s second stage analysis.

$^{12}$ We consider the COLS method for parametric estimation because it is more frequently used than other maximum likelihood methods. Results with standard composed error models are very similar.
\( y_i, \, i = 1, \ldots, N \) for \( N \) decision making units (DMUs) randomly from a uniform distribution on the interval \([4, 20] \). To generate the proportion of complex customer, \( \phi_i \) for each firm \( i \) in the sample, we first draw a random number \( (x_i) \) from a normal distribution \( N(0,1) \). We insert the random number \( (x_i) \) into the cumulative distribution function of logistic distribution to calculate the proportion \( (\phi_i) \) of \( y_{2i} \):

\[
\phi_i = \frac{0.2}{1 + e^{x_i - \tau y_i}}, \, i = 1, \ldots, N \tag{6}
\]

Here \( x_i \) is random noise, representing the effect of uncertain heterogeneity of customers. \( \tau \) is a parameter, representing growth rate of complex customers with respect to firm size. Therefore, the proportion of complex customers, \( \phi_i \), is determined by the realization of both firm size (total number of customers \( y_i \)) and unobserved contextual factors (noise \( x_i \)). We multiply the logistic distribution function by 0.2 so that the proportion of complex customers is no more than 20\%. We assume only a small proportion of complex customers exists for all firms in order not to load the specification in favor of finding cost distortions. The derivative of \( \phi_i \) with respect to \( y_i \) is

\[
\frac{\partial \phi_i}{\partial y_i} = 0.02 * \tau * \left( \frac{1}{1 + e^{x_i - \tau y_i}} \right)^2 e^{x_i - \tau y_i} > 0, \tag{7}
\]

Therefore, the expected proportion of complex customers is higher when an electricity distributor is larger (has more customers).

We next generate \( y_{1i} \) and \( y_{2i} \) as

\[
y_{1i} = (1 - \phi_i) y_i \text{ and } y_{2i} = \phi_i y_i, \, i = 1, \ldots, N. \tag{8}
\]

Each pair \( (y_{1i}, y_{2i}) \) represents the two outputs for observation \( i \). We generate \( \varepsilon_i \) for
each observation \( i = 1, \ldots, N \), from independent half-normal distribution \( |N(0, \sigma^2)| \).

The variance \( \sigma^2 \) of the logarithm of the random inefficiency term \( \epsilon_i \) is generated from an independent uniform distribution over the interval \([0, 0.1998]\) so that the range of mean efficiency, given by \( E(1/\theta) = \exp(-\sigma_u \sqrt{2/\pi}) \), is between 0.7 and 1 (Banker et al. 1994).

We obtain a sample of \( N \) observations by inserting the realizations of \( y_{i1}, y_{i2} \), and \( \epsilon_i \) into the following equation to calculate the observed operating costs \( c_i \),

\[
c_i = (k_1 y_{i1}^{\alpha_1} + k_2 y_{i2}^{\alpha_2}) \cdot e^{\epsilon_i}.
\]  

(9)

Using the observed pair of costs and aggregate output \( (c_i, y_i) \), we estimate the two COLS and four DEA models listed earlier to get the estimated value \( (\hat{\theta}_i) \) for technical inefficiency. Then we obtain the predicted operating cost by projecting the observations to the cost frontier, \( \hat{c}_i^k = c_i^k / \hat{\theta}_i^k \), where \( c_i^k \) and \( \hat{c}_i^k \) are observed total cost and estimated efficient cost of DMU \( i \) at iteration \( k \). We calculate the mean absolute error \( (MAE) \) as the average of the absolute difference between the predicted and true operating costs for different models. We also calculate root mean squared error \( (RMSE) \) between predicted and true operating costs for different models.

After conducting 2,000 Monte Carlo trials, we calculate the overall averages of \( MAE \) and \( RMSE \) for the 2,000 trials for each model:

\[
MAE = \frac{1}{2000} \sum_{k=1}^{2000} \left( \frac{1}{N} \sum_{i=1}^{N} |c_i^k - \hat{c}_i^k| \right)
\]  

(10)

\[
RMSE = \frac{1}{2000} \sum_{k=1}^{2000} \left( \sqrt{\frac{1}{N} \sum_{i=1}^{N} (c_i^k - \hat{c}_i^k)^2} \right)
\]  

(11)

In a similar manner, we also compare the performance of various COLS and DEA
models when detailed information about different types of customers is employed by the regulator. That is, after the generation of random samples, we directly use the observed dataset \((c_i, y_{it}, y_{it})\) , \(i = 1, \ldots, N\), and estimate the two COLS and four DEA models to obtain the corresponding predicted costs. To examine the robustness of our simulation results, we conduct Monte Carlo experiments for four different values of the parameter \((\tau = 0.1, 0.2, 0.3, 0.4)\) separately, which corresponds to average proportion of complex customers = 0.108, 0.148, 0.171, and 0.183.

4.3. Monte Carlo Simulation Results

The simulation results of the 2,000 Monte Carlo experiments are presented in Table 1 below. Overall, the BCC model that allows variable returns to scale performs the best with the lowest MAE and RMSE with a substantial advantage over all COLS and other DEA models when only the total number of customers is employed as the output measure. The two stage NDRS model improves the estimation performance only slightly over the basic NDRS model, and it also performs much worse than the VRS model. Both Cobb-Douglas COLS and Translog COLS perform slightly better than CCR, NDRS and two stage NDRS models, because like the BCC model they do not require constant or non-decreasing returns to scale assumptions. However, Cobb-Douglas COLS and Translog COLS perform much worse than the BCC model in our Monte Carlo experiments because of functional misspecification errors.

As expected, when the number of complex and the number of regular customers are available, the NDRS DEA model outperforms all other models, because it assumes the correct underlying model specification. Very interestingly, the BCC model
performs very well in terms of MAE but worse in terms of RMSE, indicating that except for a few outliers\textsuperscript{13} the BCC model provides robust estimation of the cost function even though it does not use the information on IRS technology even when detailed information about the outputs is available. The two stage NDRS model commonly used by regulators performs much worse than other models in all settings because it is not the right way to capture the information on outputs with different marginal costs. Untabulated simulation results with more outputs and inputs and different functional specification indicate that these insights are robust.

Overall, these simulation results clearly indicate that we have a much better fit of data and robust estimation of the cost function using the flexible BCC model specification when the empirical model is a simple abstraction of the true cost model, such as when detailed information on outputs is not available. In particular, the MAE and RMSE of the NDRS model and the two stage NDRS model are much higher than for the BCC model. This implies that those regulators employing the NDRS model may be committing large errors in benchmarking costs.

<table>
<thead>
<tr>
<th>Customer Mix</th>
<th>Models</th>
<th>Only $y$ is known</th>
<th>Both $y_1$ and $y_2$ are known</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\tau = 0.1$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\text{mean}(\phi) = 0.108$</td>
<td>Cobb-Douglas COLS</td>
<td>6.147</td>
<td>5.127</td>
</tr>
<tr>
<td></td>
<td>Translog COLS</td>
<td>6.075</td>
<td>5.080</td>
</tr>
<tr>
<td></td>
<td>CCR CRS DEA</td>
<td>6.485</td>
<td>5.362</td>
</tr>
<tr>
<td></td>
<td>BCC VRS DEA</td>
<td>5.016</td>
<td>4.003</td>
</tr>
<tr>
<td></td>
<td>NDRS DEA</td>
<td>6.422</td>
<td>5.225</td>
</tr>
<tr>
<td></td>
<td>NDRS-two stage</td>
<td>5.024</td>
<td>4.617</td>
</tr>
<tr>
<td>$\tau = 0.2$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cobb-Douglas COLS</td>
<td>5.962</td>
<td>5.294</td>
</tr>
</tbody>
</table>

\textsuperscript{13} Analysis of these outliers revealed that they are likely to be large firms with low true efficiency. This suggests that regulators face a different challenge in benchmarking large firms.
Meanwhile, we investigate how the model specification affects empirical results of test of returns to scale, and use two test statistics to examine the null hypothesis of existence of non-decreasing returns to scale (Banker 1993, 1996; Banker and Natarajan 2011). We first construct the following test statistics, assuming the logarithm of the true inefficiency $\theta_j$ is distributed as exponential over the interval $[0, \omega)$,

$$\sum_{j=1}^{N} \ln \theta_j^{\text{NDRS}} / \sum_{j=1}^{N} \ln \theta_j^{\text{VRS}},$$

where $\theta_j^{\text{NDRS}}$ and $\theta_j^{\text{VRS}}$ are the estimated inefficiencies of observation $j$ solving the NDRS and VRS DEA models. As the above test statistics is always greater than or equal to 1 by construction, this test statistics has the half-$F$ distribution, $|F_{2N,2N}|$, with $2N, 2N$ degrees of freedom over the range $[1, \omega)$. The half-$F$ distribution is the $F$-

<table>
<thead>
<tr>
<th>$\tau = 0.3$</th>
<th>$\tau = 0.4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\text{mean}(\phi) = 0.183$</td>
<td>$\text{mean}(\phi) = 0.171$</td>
</tr>
<tr>
<td>$\text{Translog COLS}$</td>
<td>$\text{Translog COLS}$</td>
</tr>
<tr>
<td>5.143</td>
<td>4.161</td>
</tr>
<tr>
<td>$\text{CCR CRS DEA}$</td>
<td>$\text{CCR CRS DEA}$</td>
</tr>
<tr>
<td>5.457</td>
<td>4.160</td>
</tr>
<tr>
<td>$\text{BCC VRS DEA}$</td>
<td>$\text{BCC VRS DEA}$</td>
</tr>
<tr>
<td>5.143</td>
<td>4.747</td>
</tr>
<tr>
<td>$\text{NDRS DEA}$</td>
<td>$\text{NDRS DEA}$</td>
</tr>
<tr>
<td>5.134</td>
<td>4.717</td>
</tr>
<tr>
<td>$\text{NDRS-two stage}$</td>
<td>$\text{NDRS-two stage}$</td>
</tr>
<tr>
<td>4.624</td>
<td>4.374</td>
</tr>
</tbody>
</table>
distribution truncated below at 1, the median of the $F$ distribution when the two
degrees of freedom are equal. Next, if the logarithm of the true inefficiency $\theta_j$ is
distributed as half-normal over the interval $[0, \infty)$, then the test statistic is constructed
as,

$$\sum_{j=1}^{N} \left( \ln \theta_{j, \text{NDRS}} \right)^2 / \sum_{j=1}^{N} \left( \ln \theta_{j, \text{VRS}} \right)^2 .$$

This test statistic has the half-$F$ distribution $|F_{N,N}|$, with $N$, $N$ degrees of freedom over
the range $[1, \infty)$.

In addition, we employ the Kolmogrov-Smirnov test statistic to examine the null
hypothesis when no parametric assumptions are maintained about the probability
distribution (Banker 1993, Banker and Natarajan 2011). Table 2 shows the
percentages of Monte Carlo experiments that the null hypothesis of NDRS is rejected
at significance level of 1%, 5% or 10%. When the aggregate $y$ is used as single output
and the average percentage of complex customer is 10.8% ($\tau = 0.1$), we can reject
NDRS at 1% level in 5.2% of 2000 Monte Carlo experiments, using the test statistics
$\sum_{j=1}^{N} \ln \theta_{j, \text{NDRS}} / \sum_{j=1}^{N} \ln \theta_{j, \text{VRS}}$. As the proportion of complex customers increases ($\tau = 0.2,
0.3$ or $0.4$), NDRS is rejected in more than 50% of Monte Carlo experiments at 1%
level. That is, the estimated production function apparently exhibits a region of
decreasing returns to scale when the total number of customers, $y$, is used as a single
output. Not surprisingly, if the number of regular customers and number of complex
customers are used as two separate outputs, NDRS is rejected in only a few Monte
Carlo experiments. In many settings, we cannot reject NDRS in all 2000 experiments,
using the three test statistics. This suggests that the estimated production function
indeed reflects the underlying true production technology when the all relevant outputs are correctly identified in the empirical model.

Table 2: Percentages of Rejections of Non-Decreasing Returns to Scale in 2000 Monte Carlo Experiments

<table>
<thead>
<tr>
<th>Customer Mix</th>
<th>Test Statistics</th>
<th>Only y is known</th>
<th>Both y1 and y2 are known</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1%</td>
<td>5%</td>
</tr>
<tr>
<td>$\tau = 0.1$, mean($\phi$) = 0.108</td>
<td>$\frac{\sum_{j=1}^{N} \ln \theta_j^{NDRS}}{\sum_{j=1}^{N} \ln \theta_j^{VRS}}$</td>
<td>5.2%</td>
<td>16.1%</td>
</tr>
<tr>
<td></td>
<td>$\frac{\sum_{j=1}^{N} (\ln \theta_j^{NDRS})^2}{\sum_{j=1}^{N} (\ln \theta_j^{VRS})^2}$</td>
<td>9.2%</td>
<td>21.2%</td>
</tr>
<tr>
<td></td>
<td>Kolmogrov-Smirnov test</td>
<td>28.6%</td>
<td>39.6%</td>
</tr>
<tr>
<td></td>
<td>Translog test</td>
<td>23.4%</td>
<td>42.8%</td>
</tr>
<tr>
<td>$\tau = 0.2$, mean($\phi$) = 0.148</td>
<td>$\frac{\sum_{j=1}^{N} \ln \theta_j^{NDRS}}{\sum_{j=1}^{N} \ln \theta_j^{VRS}}$</td>
<td>50.3%</td>
<td>67.5%</td>
</tr>
<tr>
<td></td>
<td>$\frac{\sum_{j=1}^{N} (\ln \theta_j^{NDRS})^2}{\sum_{j=1}^{N} (\ln \theta_j^{VRS})^2}$</td>
<td>56.8%</td>
<td>70.8%</td>
</tr>
<tr>
<td></td>
<td>Kolmogrov-Smirnov test</td>
<td>75.2%</td>
<td>82.1%</td>
</tr>
<tr>
<td></td>
<td>Translog test</td>
<td>18.6%</td>
<td>31.5%</td>
</tr>
<tr>
<td>$\tau = 0.3$, mean($\phi$) = 0.171</td>
<td>$\frac{\sum_{j=1}^{N} \ln \theta_j^{NDRS}}{\sum_{j=1}^{N} \ln \theta_j^{VRS}}$</td>
<td>65.7%</td>
<td>76.7%</td>
</tr>
<tr>
<td></td>
<td>$\frac{\sum_{j=1}^{N} (\ln \theta_j^{NDRS})^2}{\sum_{j=1}^{N} (\ln \theta_j^{VRS})^2}$</td>
<td>67.4%</td>
<td>77.0%</td>
</tr>
<tr>
<td></td>
<td>Kolmogrov-Smirnov test</td>
<td>80.8%</td>
<td>85.9%</td>
</tr>
<tr>
<td></td>
<td>Translog test</td>
<td>14.3%</td>
<td>24.5%</td>
</tr>
<tr>
<td>$\tau = 0.4$, mean($\phi$) = 0.183</td>
<td>$\frac{\sum_{j=1}^{N} \ln \theta_j^{NDRS}}{\sum_{j=1}^{N} \ln \theta_j^{VRS}}$</td>
<td>55.6%</td>
<td>67.4%</td>
</tr>
<tr>
<td></td>
<td>$\frac{\sum_{j=1}^{N} (\ln \theta_j^{NDRS})^2}{\sum_{j=1}^{N} (\ln \theta_j^{VRS})^2}$</td>
<td>57.1%</td>
<td>67.3%</td>
</tr>
<tr>
<td></td>
<td>Kolmogrov-Smirnov test</td>
<td>71.5%</td>
<td>76.7%</td>
</tr>
</tbody>
</table>
In summary, when a simple function is used to represent a much more complex true relationship, the best fit is obtained by estimating the most general or flexible functional form instead of imposing theoretical structure that cannot be sustained due to specification errors. In the DEA methodology, this means that we should use the variable returns to scale (VRS) model instead of the NDRS model to estimate the simplified distribution cost model.

5. Empirical Analysis

In this section, we empirically examine the returns to scale assumption using two separate research samples from the electricity distribution industry in the U.S. and Brazil. The U.S. electricity industry is subject to a cost of service regulation,\(^\text{14}\) while the Brazilian regulator, ANEEL has switched from the cost of service regime to incentive regulation, relying on DEA models to estimate efficient costs of electricity distribution firms to set price/revenue cap (McDermott 2012; ANEEL 2010). The research design enables us to provide empirical comparative evidence on returns to scale assumption for distribution firms subject to different types of regulation.

5.1. Empirical Results for Electricity Distribution Firms in the U.S.

We first investigate a sample of the U.S. electricity distribution firms drawn from Federal Energy Regulatory Commission (FERC) Form-1 database. FERC requires

\(^{14}\) Under the cost of service regime, electric utilities are fully reimbursed for their operational expenditures plus a guaranteed rate of return on the capital investments (RAP 2011; McDermont 2012)
major electric utilities to file their financial and operational information annually for public use.\textsuperscript{15} A firm is classified as “major” if its the annual sales in each of prior three years exceeds one of following thresholds: (1) 1,000,000 megawatt hours (MWh) of total annual sales; (2) 100 MWh of annual sales for resale; (3) 500 MWh of annual power exchanges delivered or annual wheeling for others.

To assess returns to scale for the distribution industry, we focus only on pure electricity distribution firms that do not produce electric power in their own generation plants. Our sample period is from 1994 to 2012.\textsuperscript{16} The research sample comprises 76 electricity distribution firms and 484 firm-year observations. The panel is not balanced, and the number of electricity distribution firms increases over time because some states such as New York, Pennsylvania and Massachusetts restructured the industry and required vertically integrated electric utilities to divest their generation assets and become wire-only distribution firms (RAP 2011; McDermott 2012).

As in many prior studies, we use the simple aggregate number of customers (Customer) and total electric power delivered (Energy) as two outputs.\textsuperscript{17} We use three categories of costs as single input separately, which include 1. Distribution expenses (DistExp), referring to operational and maintenance expenses of distribution lines; 2. Customer service expenses (CustExp), referring to customer accounts, customer service and informational expenses; 3. Total expenses (TotExp), referring to both distribution and customer service costs.

\textsuperscript{15} The dataset is publicly available at http://www.ferc.gov/docs-filing/forms/form-1/data.asp.
\textsuperscript{16} The earliest year of current FERC Form 1 database is 1994.
\textsuperscript{17} Network length is not included as an output because it is not available in FERC Form 1 database.
We use the three test statistics for returns to scale as in Section 4.3 to examine the null hypothesis that the U.S. electricity distribution firms exhibit constant or non-decreasing returns to scale (Banker 1993, 1996; Banker and Natarajan 2011). That is, we examine the null hypothesis that the inefficiency scores estimated by CRS (or NDRS) DEA model and those estimated by VRS DEA model are drawn from the same distribution. If the null hypothesis is rejected, it means that the U.S. electricity distribution firms do not exhibit CRS (or NDRS).

The summary statistics of inputs and outputs are shown in Table 3. The U.S. electricity distribution firms on average served 459,346 customers and delivered 10,564,921 megawatt hours of electric power annually. The average of annual operating and maintenance costs of distribution lines are approximately $45.4 million, and the average of customer service costs is $42.3 million.

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>25&lt;sup&gt;th&lt;/sup&gt;</th>
<th>Median</th>
<th>75&lt;sup&gt;th&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>DistExp</td>
<td>45,357,332</td>
<td>72,363,554</td>
<td>2,691,617</td>
<td>6,841,974.5</td>
<td>60,825,195</td>
</tr>
<tr>
<td>CustExp</td>
<td>42,325,987</td>
<td>73,670,614</td>
<td>1,421,676.5</td>
<td>5,646,634.5</td>
<td>47,472,800</td>
</tr>
<tr>
<td>TotExp</td>
<td>87,683,319</td>
<td>140,924,654</td>
<td>4,170,101</td>
<td>11,878,322</td>
<td>107,132,717</td>
</tr>
<tr>
<td>Customer</td>
<td>459,345.8</td>
<td>734,540.9</td>
<td>28,868.5</td>
<td>72,397.5</td>
<td>611,990.5</td>
</tr>
<tr>
<td>Energy</td>
<td>10,564,921</td>
<td>17,964,683</td>
<td>585,271</td>
<td>1,847,438</td>
<td>13,962,112</td>
</tr>
</tbody>
</table>

We run CRS, VRS and NDRS DEA models to estimate the DEA inefficiency scores, and then conduct the DEA-based returns to scale tests for each model specification separately. The test results in Table 4 reject both CRS and NDRS...
assumptions for U.S. electricity distribution firms when compared with the more general VRS assumption at the 1% level for all alternative test statistics.

We further investigate the nature of returns to scale for the U.S. electricity distribution firms employing the translog function as in the prior literature (e.g. Akkemik 2009; Huang, Chen and Yang, 2010; Bó and Rossi, 2007; Hattori, 2002; Burns and Weyman-Jones, 1996; Filippini, 1996, 1998; Scully, 1998). The translog cost function is specified as

$$
\ln(Cost_{jt}) = \beta_0 + \beta_1 \ln(Energy_{jt}) + 1/2 \beta_2 [\ln(Energy_{jt})]^2 + \beta_3 \ln(Customer_{jt}) + 1/2 \beta_4 [\ln(Customer_{jt})]^2 + \beta_5 \ln(Energy_{jt}) \cdot \ln(Customer_{jt}) + \epsilon_{jt}
$$

(12)

where $Cost_{jt}$ represents the reported distribution expenses(DistExp), customer expenses(CustExp), or total expenses (TotExp) of an electricity distribution firm $j$ in year $t$; $Energy$ represents electric power delivered to customers, and $Customer$ refers to the total number of customers. The elasticity of costs with respect to outputs is

<table>
<thead>
<tr>
<th>Table 4: Results of DEA Tests of Returns to Scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test Statistic</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Hypothesis 1 (CRS vs VRS): the technology exhibits constant returns to scale</td>
</tr>
<tr>
<td>$1/N \sum_{j=1}^{N} \ln \theta_j^{VRS}$</td>
</tr>
<tr>
<td>$1/N \sum_{j=1}^{N} (\ln \theta_j^{VRS})^2$</td>
</tr>
<tr>
<td>Kolmogrov-Smirnov test</td>
</tr>
<tr>
<td>Hypothesis 2 (NDRS vs VRS): the technology exhibits no-decreasing returns to scale</td>
</tr>
<tr>
<td>$1/N \sum_{j=1}^{N} \ln \theta_j^{NDRS}$</td>
</tr>
<tr>
<td>$1/N \sum_{j=1}^{N} (\ln \theta_j^{NDRS})^2$</td>
</tr>
<tr>
<td>Kolmogrov-Smirnov test</td>
</tr>
</tbody>
</table>
The inverse of cost elasticity is a measure of economies of scale at observation \( j \) in year \( t \):

\[
e(y_{jt}) = \frac{\partial \ln(Cost_{jt})}{\partial \ln(Energy_{jt})} + \frac{\partial \ln(Cost_{jt})}{\partial \ln(Customer_{jt})} = \beta_1 + \beta_2 \ln(Energy_{jt}) + \beta_3 + \beta_4 \ln(Customer_{jt}) + \beta_5 (\ln(Energy_{jt}) + \ln(Customer_{jt}))
\]

(13)

We use pooled OLS regression method to estimate equation (12). The estimation results are shown in Table 5.

**Table 5: Estimation Results of Translog Cost Function for Distribution Firms in the U.S.**

<table>
<thead>
<tr>
<th>VARIABLES</th>
<th>(1) Log(Dist Exp)</th>
<th>(2) Log(Cust Exp)</th>
<th>(3) Log(Tot Exp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Log(Energy)</td>
<td>0.744</td>
<td>3.338***</td>
<td>2.040***</td>
</tr>
<tr>
<td></td>
<td>(1.464)</td>
<td>(7.033)</td>
<td>(5.126)</td>
</tr>
<tr>
<td>Log(Customer)</td>
<td>-0.476</td>
<td>-1.984***</td>
<td>-1.327***</td>
</tr>
<tr>
<td></td>
<td>(-0.997)</td>
<td>(-4.689)</td>
<td>(-3.604)</td>
</tr>
<tr>
<td>( \frac{1}{2} )Log(Energy)(^2)</td>
<td>-0.0129</td>
<td>-0.467***</td>
<td>-0.207***</td>
</tr>
<tr>
<td></td>
<td>(-0.162)</td>
<td>(-5.849)</td>
<td>(-3.788)</td>
</tr>
<tr>
<td>( \frac{1}{2} )Log(Customer)(^2)</td>
<td>0.177***</td>
<td>-0.0896</td>
<td>0.0973***</td>
</tr>
<tr>
<td></td>
<td>(3.233)</td>
<td>(-1.295)</td>
<td>(3.092)</td>
</tr>
<tr>
<td>Log(Energy)*</td>
<td>-0.0446</td>
<td>0.289***</td>
<td>0.0835**</td>
</tr>
<tr>
<td></td>
<td>(-0.715)</td>
<td>(4.220)</td>
<td>(2.155)</td>
</tr>
<tr>
<td>Log(Customer)*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>7.762***</td>
<td>-3.112**</td>
<td>3.654***</td>
</tr>
<tr>
<td></td>
<td>(6.473)</td>
<td>(-2.287)</td>
<td>(3.533)</td>
</tr>
<tr>
<td>Observations</td>
<td>484</td>
<td>484</td>
<td>484</td>
</tr>
<tr>
<td>R-squared</td>
<td>0.944</td>
<td>0.956</td>
<td>0.963</td>
</tr>
</tbody>
</table>

Robust t-statistics in parentheses *** p<0.01, ** p<0.05, * p<0.1

Table 6 shows the results of tests of returns to scale based on equation (13). We find that if distribution expenses is used as the input, the economies of scale measure
of 206 out of 484 firm-year observations is greater than 1, and significant at 1% for 175 of those observations. If customer expenses is used as the input, there are 291 firm-year observations exhibiting significant diseconomies of scale. If total expenses is used as the input, there are 187 firm-year observations exhibiting significant diseconomies of scale.

Table 6: Results of Tests of Returns to Scale at Each Observation

<table>
<thead>
<tr>
<th></th>
<th>DistExp</th>
<th>CustExp</th>
<th>TotExp</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CRS is rejected</td>
<td>CRS is not rejected</td>
<td>CRS is rejected</td>
</tr>
<tr>
<td>Increasing</td>
<td>270</td>
<td>8</td>
<td>53</td>
</tr>
<tr>
<td>Returns to Scale</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Decreasing</td>
<td>175</td>
<td>31</td>
<td>291</td>
</tr>
<tr>
<td>Returns to Scale</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The test size is 1% level.

In summary, the empirical evidence with both non-parametric and parametric methods reveals a region of apparent DRS for the large electricity distributors in the U.S. These results are consistent with the empirical results in the prior literature that report economies of scale are apparently exhausted for large electricity firms (Neuberg 1977; Nelson and Primeaux 1988; Salvanes and Tjotta 1994; Filippini 1998; Arocena 2008).

5.2 Empirical Results for Electricity Distribution Firms in Brazil

More recently, in his submitted testimony, Banker (2011) considers the case of Brazil because its regulator uses a benchmarking model assuming NDRS for price cap
The National Agency for Electric Energy (ANEEL) in Brazil launched the third cycle of periodic tariff review of electricity distribution utilities in 2009 (ANEEL 2010, 2011). One important innovation in ANEEL's proposal, following the example of several European regulators, is its use of a DEA model to provide the benchmark that forms the basis of incentive regulation. However, a problem with any new methodology is that all the right answers are not always clear and some of the shortcomings are hidden. A serious flaw in ANEEL's model is its maintained assumption of NDRS for the cost function linking operating costs (OPEX) directly to just a small number of aggregate measures of outputs. More specifically, ANEEL proposed the following single input and three outputs specifications of cost function in its initial proposal for distribution utilities: 19

Input: OPEX, Outputs: Network length (km), Number of customers, Electric power (KwH) sold.

Some old companies became much larger after the expansion of rural electrification through the “Light for All” program introduced by the Government of Brazil. The Government's decision to provide universal electric service may be noble and commendable, but it imposes different costs for the distributors throughout the country. The proportion of customers served by this program is very different across distributors. The average cost of serving new customer units is higher, and the growth

---

18 The total expenditure of electricity distribution companies in Brazil was approximate R$ 26.9 billion and the operating expenditure of eight largest electricity transmission companies was about R$ 2.6 billion in 2008, when the exchange rate was approximately US$ 1 = R$ 1.6.

19 ANEEL also adjusted for some complexity factors as environmental variables in a second-stage regression. As shown in the simulation study reported in Section 4, these adjustments do not address the problem caused by the NDRS assumption. In addition, ANEEL also proposed a multi-output Cobb-Douglas cost model.
in size has been accompanied by an increase in average costs.

We analyze the actual data for the 28 largest Brazilian distribution firms whose electric power sales exceed 1 TwH used by ANEEL (Banker 2011). The sample consists of a 7 year panel from 2003 to 2009 (196 firm year observations in total). We split the panel into seven separate annual cross-sectional subsamples. We estimate the BCC and NDRS DEA models separately with the cross-sectional dataset for each year. After obtaining the annual inefficiency scores for the 28 firms, we construct the two half-F test statistics and the Kolmogrov-Smirnov test statistic as described in Section 5.1. Table 6 displays the results of returns to scale tests.

<table>
<thead>
<tr>
<th>Test Statistic</th>
<th>Test-stat.</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hypothesis (NDRS vs VRS): the technology exhibits no-decreasing returns to scale</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\frac{\sum_{j=1}^{N} \ln \theta_j^{NDRS}}{\sum_{j=1}^{N} \ln \theta_j^{RS}}$</td>
<td>1.420</td>
<td>0.000</td>
</tr>
<tr>
<td>$\frac{\sum_{j=1}^{N} (\ln \theta_j^{NDRS})^2}{\sum_{j=1}^{N} (\ln \theta_j^{RS})^2}$</td>
<td>1.485</td>
<td>0.001</td>
</tr>
<tr>
<td>Kolmogrov-Smirnov test</td>
<td>0.224</td>
<td>0.000</td>
</tr>
</tbody>
</table>

The test results reject NDRS assumption in Brazilian distribution industry when compared with the more general VRS assumption at the 1% level for all three models for all alternative test statistics. In untabulated results, we find that there are 7 to 10 firms out of 28 each year that apparently exhibit deceasing return to scale.\textsuperscript{20}

The evidence with both non-parametric and parametric methods reveals a region of

\textsuperscript{20} In a revised proposal, ANEEL augments the sample but the existence of apparent DRS for large firms persists.
apparent DRS for the large electricity distributors in Brazil.21 As postulated in our conceptual model, we attribute these surprising empirical findings to the fact that the larger firms in Brazil may have a large proportion of customers or locations that are difficult to serve. If the simple output specification proposed by the regulator captures most differences in marginal costs of customers (which requires that they are all almost identical), then under the NDRS assumption these large firms are classified as inefficient. On the other hand, if there are differences in marginal costs of serving different customers, as is likely, the apparent region of DRS signals model misspecification. The difference due to model specification in benchmarked costs can be large, amounting to a difference of as much as R$808 million (47.1% of operating cost) for the largest firm in the Brazilian distribution industry in 2009.

6. Concluding Remarks

To justify the existence of local natural monopolies, economic theory posits that increasing returns to scale prevail in a variety of industries such as electricity distribution, telecommunication and water and sewage. Without such scale economies it is difficult to justify the existence of regional monopolies in these industries. Based on this theoretical consideration, several regulators around the world, backed by prescriptions from some academic studies and testimonies, have implemented incentive regulation that relies on production (or cost) frontiers estimated maintaining the assumption of constant returns to scale (CRS) or non-decreasing returns to scale (NDRS), instead of allowing variable returns to scale (VRS).

21 As documented by Banker, Conrad and Strauss (1986), returns to scale need not be the same throughout the sample. We may find regions where IRS prevail and regions where DRS prevail.
Prior empirical research, however, provides at best mixed empirical results for
returns to scale in these industries. Several studies using parsimonious specifications
of outputs and inputs in the production (or cost) function find a region of decreasing
returns to scale for large firms, appearing to contradict economic theory. In this paper,
we address the question of empirically imposing the NDRS assumption in
benchmarking models for incentive regulation. Conceptually, we recognize that the
empirical specification of a production function is only an abstraction and not
necessarily the reality. Even though the full and complete specification of a
production relationship between all inputs, outputs and contextual variables may
exhibit NDRS, we show that it is not appropriate to impose such an assumption on the
simpler empirical abstraction.

Most incentive regulation based on estimated production frontiers specifies a
relatively simple functional relationship between costs and a few outputs and
contextual variables. We consider the situation when firms get larger by increasing
customers that are costlier to serve, but data on who the costlier customers are or their
proportion, are not available. If the output is specified only in terms of the number of
customers or total sales without explicitly capturing the impact of different types of
customers in the production function, the estimated production function will appear to
exhibit variable returns to scale (VRS) with a region of decreasing returns. It is
incorrect, therefore, to impose CRS or NDRS in the empirical estimation with simple
output specification even though the true production function capturing the complex
realities exhibits NDRS as required by economic theory.
To demonstrate this proposition, we conduct extensive Monte Carlo simulation experiments. Even for a simple setting with two outputs (regular and complex customers) and single input (costs), where the empirical model specifies the output only as total customers, our simulation results indicate that the VRS model performs much better than the models that impose NDRS or models that correct for the customer mix in a two-stage DEA model with NDRS. The evidence very clearly indicates that the NDRS model should not be used in incentive regulation when detailed data on customers or service areas with different marginal costs are not available to the regulator to construct a production (or cost) model that captures prevailing operating complexities.

It will be important for future research to develop empirical models that capture the production realities in this setting so that differences between managerial inefficiency, scale inefficiency, operating complexities and specification errors can be analyzed in a reliable manner. In the meanwhile, policy implications for regulators are clear. They should not use NDRS or CRS models, but instead use the VRS model, unless and until detailed models that capture differences in marginal costs of different customers are developed and validated.

While we have presented our analysis here in the context of use of DEA for incentive regulation, the concepts are general and extend to many other situations where DEA is applied. For instance, even if the decision making units have the flexibility to adjust their scale size, it does not imply that their efficiency should be

---

22 Similar insights obtain when we conduct the simulations with more complex production functions involving several different outputs and multiple cost categories.
measured relative to a production function estimated imposing the CRS assumption. This paper shows that if all relevant inputs, outputs and contextual variables are not included, or cannot be measured with precision, then the VRS model perform the best without imposing any additional structure. Furthermore, an ad hoc second stage analysis intended to “remove” the impact of differences within each output (or input) does not suffice.
Reference


Huang, Y., Chen, K., Yang, C., 2010. Cost Efficiency and Optimal Scale of Electricity Distribution Firms in Taiwan: An Application of Metafrontier Analysis. Energy


