Persistent and transient productive inefficiency in a regulated industry: electricity distribution

M. Filippini [*] W. Greene [†] G. Masiero [‡]

Published in:
Energy Economics (2018), 69: 325-334.
< https://doi.org/10.1016/j.eneco.2017.11.016 >

Abstract

The productive efficiency of a firm can be decomposed into two parts, one persistent and one transient. This distinction seems to be appealing for regulators. During the last decades, public utilities such as water and electricity have witnessed a wave of regulatory reforms aimed at improving efficiency through incentive regulation. Most of these regulation schemes use *benchmarking*, namely measuring companies' efficiency and rewarding them accordingly. Focusing on electricity distribution, we sketch a theoretical model to show that an imperfectly informed regulator may not disentangle the two parts of the cost efficiency. Therefore, the regulator may fail to set optimal efficiency targets, which also undermines quality. We then provide evidence on the presence of persistent and transient efficiency using data on 28 New Zealand electricity distribution companies between 2000 and 2011. First, we estimate a total cost function by means of traditional stochastic frontier models for panel data. These come up with an estimation of the persistent part or the transient part of the cost efficiency. Finally, we use the more recent generalized true random effects model that allows for the simultaneous estimation of both transient and persistent efficiency. We also find some evidence that persistent efficiency is associated to higher quality, and wrong efficiency targets are associated to lower quality compliance.

Keywords: cost efficiency, regulation, persistent and transient productive efficiency, electricity distribution.

JEL classification: C1, C23, D24.

---

[*]Institute of Economics (IdEP), Università della Svizzera italiana (USI); Department of Management, Technology and Economics, ETH Zurich, Switzerland. Corresponding author. E-mail address: massimo.filippini@usi.ch. We are indebted to Heinke Wetzel for providing the dataset used in the empirical part of the paper, and to Nilkanth Kumar for his helpful advice. We thank two anonymous reviewers for their many insightful comments and suggestions.

[†]Department of Economics, Stern School of Business, New York University, USA.

[‡]Department of Management, Information and Production Engineering, University of Bergamo, Italy; Institute of Economics (IdEP), Università della Svizzera italiana (USI), Switzerland.

1 Introduction

During the last twenty years, several countries have introduced reforms in public utility sectors, such as water, electricity and telecommunications. Regarding electricity, two key elements of these reforms are the introduction of competition in the supply and generation of electricity, and the introduction of new regulation methods in the transmission and distribution of electricity considered as natural monopolies. The new methods apply the incentive regulation theory (Laffont and Tirole, 1993). They provide incentives for productive efficiency by compensating the company with its savings.[1] Several incentive-based schemes make use of information on the level of overall productive efficiency (or cost efficiency) of an electricity distribution company, i.e. technical and allocative efficiency.

The level of productive efficiency can be decomposed into two parts, one persistent and one transient (Colombi et. al., 2014). The presence of structural problems in the organization of the production process or systematic shortfalls in managerial capabilities can generate the persistent part. Conversely, the presence of non-systematic management problems in the short term, the hiring of new workers that require some initial learning time for their tasks, or the adjustment in the production process due to new regulation or new techniques may determine the transient part. Failing to distinguish between these two types of efficiency in the regulation process may lead to serious consequences. Since the application of price cap regulation is based on erroneous estimation of persistent and transient efficiency gains, service quality provided by electricity distribution companies could suffer. Also, investment decisions could be postponed and incentives for innovation adoption could be weakened.

In this paper we show that imperfect information exposes the regulator to wrong efficiency target setting, which may worsen the already existing problem of quality provision. We argue that one reason for this might be the underestimation of the persistent component of inefficiency. To overcome this problem, there are recent econometric methods to estimate stochastic frontier models that

[1] See Joskow and Schmalensee (1986) for a review of incentive regulation models. Incentive regulation in electricity distribution has been investigated by several authors. See, for instance, Cullmann and Nieswand (2016) for Germany, Blázquez-Gómez and Grifell-Tatjé (2011) for Spain, and Weyman-Jones (1990) and Jamasb and Pollitt (2007) for the UK.

could be applied to enable a distinction between the two inefficiency components. These methods could represent a useful tool for regulation authorities.

We build our argument around a theoretical part, a methodological part and some suggestive evidence that follows from the empirical analysis. In the first part of the paper, we sketch a theoretical model that shows the importance of the distinction between persistent and transient inefficiency in the application of a price-cap regulation method. The problem of information acquisition on firms efficiency under price cap regulation is a critical one and has been widely investigated in the literature (e.g. Bernstein and Sappington, 1999; Iossa and Stroffolini, 2002). However, the regulatory implications of the lack of information on different types of efficiency have not been debated yet. The theoretical model illustrates how imperfect information on persistent and transient inefficiency may undermine the effectiveness of price-cap regulation and worsen service quality.

In the second part of the paper, we apply the most recent econometric method (GTRE model) proposed by Colombi et al. (2014) and Filippini and Greene (2016) to provide evidence on the presence of persistent and transient inefficiency in the electricity distribution sector using a sample of 28 New Zealand electricity distribution companies. This approach shows that a methodological improvement that disentangle the two efficiency components may provide a response to the regulatory problem, as discussed in the theoretical framework. Indeed, we find significant differences in the efficiency estimates as well as in the efficiency rankings obtained with the most recent econometric method as compared to traditional stochastic frontier methods. However, none of the regulation authorities around the world has made a distinction so far between persistent and transient inefficiency in the electricity distribution sector. Therefore, the new method could be used by regulatory authorities to refine the price cap mechanism.

The literature on the estimation of cost efficiency of regulated industries through classical econometric approaches is abundant.[2] Among the most frequent applications of frontier modelling we find several articles on energy services (e.g. Chen, Pestana and Borges, 2015; Ghosh and Kathuria, 2016), transportation services (e.g. Filippini et al., 2015; Walter, 2011), and water services (e.g.

[2]See Ramos-Real (2005) for a review of part of these studies.

Phillips, 2013). Finally, very few studies focus on electricity distribution in New Zealand (Filippini and Wetzel, 2014; Nillesen and Pollit, 2011; Scully, 1999). These studies either rely on classical models (Pitt and lee, 1981) or true random effects (TRE) models (Greene, 2005a, 2005b). As a consequence, they provide information either on the persistent part or on the transient part of the inefficiency.

In the third part of the paper we provide some suggestive evidence that persistent efficiency and service quality are related but the regulatory system does not use this information in setting efficiency targets. In particular, we show that the level of persistent efficiency estimated by our econometric approach is correlated to service quality, i.e firms that are relatively more efficient tend to show higher service quality. Moreover, the number of New Zealand electricity distribution companies that do not comply with the regulated level of quality is remarkable (around 37%). Clearly, this is only a suggestive evidence since we rely on limited information to assess whether the lack of association between quality compliance and persistent efficiency is due to reasons other than the wrong setting of efficiency targets. Ideally, the best support to the results of our theoretical model should be provided by comparing a regulated setting with no distinction between transient and persistent efficiency with a setting that takes into account this distinction, which is not possible. However, we can show that differences arising from different estimation methods in terms of efficiency ranking are statistically significant.

We organize this paper as follows. Section 2 sketches the theoretical model that investigates how regulation may fail when persistent and transient inefficiency are ignored. Section 3 introduces the cost model specification and the estimation approaches, while Section 4 describes the data. In Section 5, we present the estimation results and discuss some suggestive evidence. Section 6 summarizes and concludes.

2 A model of persistent and transient inefficiency in a regulated industry

To understand the implications of transient and persistent inefficiency of electricity distribution companies in New Zealand, we sketch a model where firms maximize the current value of future profits under price cap regulation. In the market there are N identical firms acting as local monopolies.[3] Each firm chooses price (p_t) and service quality (q_t) in each regulatory period t ($t \geq 1$) as well as the level of managerial effort (e_t). Managerial effort allows obtaining transitory efficiency gains, which are immediately exploitable in terms of cost reductions, and lagged (or persistent) efficiency gains that cut costs now as well as in the future. We show that the regulator cannot achieve optimal efficiency targets if imperfectly informed on persistent efficiency. Regulation failure may lead to postponed expenditures, which worsens service quality, or increases in monopoly rents. In addition, higher pressure to meet current minimum quality standards may undermine a firm's compliance and also result in delayed expenditures and poorer quality in the future.[4]

Consider first the following demand for electricity distribution services faced by each firm:

$$s(p_t, q_t) = q_t (\theta - p_t), \tag{1}$$

where $q_t \in [0, +\infty)$ is service quality, p_t is unit price, and θ is a parameter indicating the reservation price for a unit of quality.[5]

The following equation describes total costs at the end of the regulatory period t:

$$c(p_t, q_t, e_t, e_{t-1}, .., e_1) = \gamma s(p_t, q_t) + \left(\eta - \sum\nolimits_{j=1}^{t} \alpha e_{t-j}\right) + \beta (1 - e_t) + f(q_t) + g(e_t),$$

[3] In practice, firms are single-product monopolists. For a model of price cap regulation with multi-market monopolists when the costs of serving different markets vary, see for instance Cowan (1997a).

[4] Recently, Di Giorgio et al. (2015) proposed a theoretical approach to separate structural (or institutional) inefficiency from managerial inefficiency in public and private nursing homes. The model applies to a different regulatory setting - global budget instead of price cap regulation - and does not elaborate on the implications for the regulatory mechanism.

[5] Following the seminal paper by Mussa and Rosen (1978), we also considered the alternative functional form $s(p_t, q_t) = q_t\theta - p_t$, where quality affects willingness to pay but not the slope of the demand. The main results are unchanged. Calculation details are available upon request.

(2)

where γ is the unit cost of electricity distribution services, η is a long-lasting (or persistent) cost component and β is a temporary (or transient) cost component. Transient costs (β) can be reduced immediately by current effort e_t, with $e_t \in [0, 1]$. Conversely, persistent costs (η) can only be reduced in the future by delayed and long-lasting efficiency effort. Delayed long-lasting savings from effort are $\sum_{j=1}^{t} \alpha e_{t-j}$, where α is the marginal impact of effort on persistent costs, with $\eta \geq \alpha t$. In $t = 1$ there are no lagged effects of effort since no effort is assumed in period 0 before the regulation starts, i.e. $e_0 = 0$.

Cost persistency may arise for several reasons. For instance, management habits lead to inertia, which prevents improving tasks or solving problems immediately. In addition, environmental and social constraints related to shareholders' preferences, access to inputs, or the fulfillment of legal rules may affect the timing of efficiency gains. Finally, unions' bargaining power may succeed in postponing the achievement of efficiency targets.

Both quality and managerial effort are costly to the firm. We assume increasing marginal cost functions where $f(q_t) = q_t^2/2$ is the cost of quality and $g(e_t) = e_t^2/2$ is the cost of managerial effort.[6,7]

2.1 Price cap regulation

We assume that the firm maximizes the current value of future economic profits in each period subject to a price cap constraint. Firm's intertemporal profits at time t can be written as

$$V_t = \sum_{t=k}^{\infty} \delta^{t-k} \pi_t \left(p_t, q_t, e_t, e_{t-1}, .., e_1\right), \tag{3}$$

where

$$\pi_t \left(p_t, q_t, e_t, e_{t-1}, .., e_1\right) = p_t s(p_t, q_t) - c\left(p_t, q_t, e_t, e_{t-1}, .., e_1\right), \tag{4}$$

and $\delta \leq 1$ is the discount factor for future profits.

[6]For instance, one can think to the cost of remunerating the performance of the manager through an increase in the wage.

[7]Other functional forms for the cost of quality and managerial effort could be used without affecting the final results, provided that the realistic assumption of increasing marginal costs is preserved.

The energy authority sets a price cap for the regulatory period t according to the following rule:

$$\frac{p_t s(p_t^*, q_t^*)}{p_{t-1} s(p_t^*, q_t^*)} \leq \frac{CPI_t}{CPI_{t-1}} - X_t, \tag{5}$$

where CPI_t and CPI_{t-1} are Consumer Price Indexes and p_{t-1} is the reference price from the previous regulatory period, with initial price $p_0 = P_0$.[8] X_t is the expected efficiency gain (persistent and transient) based on past performance or average performance of other firms in the market.[9] Let us assume that prices do not inflate, i.e. $CPI_t = CPI_{t-1}$.[10] Therefore, the price cap rule in Eq. (5) can be written as:

$$p_t \leq p_{t-1} (1 - X_t) . \tag{6}$$

In addition, a minimum quality standard (MQS) is set by the regulator to limit the risk of poor quality service:

$$q_t \geq q_{min} = q_{t-1}. \tag{7}$$

This standard can be implemented by periodical controls on service quality. The combination of price cap and MQS mechanism applies to the case of electricity distribution companies in New Zealand (Shen and Yang, 2012; New Zealand Commerce Commission, 2015). Firms are subject to regulation under the Commerce Act 1986, which defines a price and quality threshold regime since 2001. This regulatory mechanism identifies companies whose performance may warrant further examination. Quality thresholds are based on two criteria: reliability and engagement with consumers to determine their demand for service quality. The reliability criterion requires that unplanned interruptions should not exceed the previous five-year average. The interruption indicators used are SAIDI (System Average Interruption Duration Index - minutes per connected customer) and SAIFI (System Average Interruption Frequency Index - interruptions per

[8] We consider a standard rule, though alternative price-cap schemes are possible. See Cowan (1997b) for a comparison of different price-cap schemes in terms of allocative efficiency.

[9] See Bernstein and Sappington (1999) for a review of the relevant basic principles to determine the X factor.

[10] Dropping the inflation component CPI_t/CPI_{t-1} does not affect the insights of our analysis but simplifies the subsequent equations.

connected customer). Since 2010, a more rigorous system is in place based on Default Price-quality Path (DPP) and Customized Price-quality path (CPP). In the current analysis we simply assume that quality standards are set according to interruptions in the previous period.

When choosing price, quality and efficiency effort each firm takes into account the effects not only on its current period profits but also on its demand and costs in the following periods. This dependence needs to be taken into account when solving the model for the equilibrium levels of price, quality and effort. Profits in period t depend upon efficiency effort in periods $t-j$. In addition, the value function represented by the flow of all future profits depends on all future levels of price, quality and efficiency effort. In equilibrium the firm selects price, quality and efficiency effort that maximize its intertemporal profit given its subsequent choices of price, quality and efficiency effort. To simplify the analysis we now set unit costs at zero, therefore $\gamma = 0$.[11] Because efficiency effort affects profits in the subsequent period and expected profits are the sum of concave functions in price, quality and efficiency effort, we can write the following first-order conditions for the firm using Eqs. (1)-(2), (3)-(4) and the price cap constraint defined by Eq. (6):[12]

$$\begin{aligned}
&\frac{\partial V_t}{\partial p_t} = \frac{\partial \pi_t\left(\boldsymbol{p_t}, q_t, e_t, e^*_{t-1}, .., e^*_1\right)}{\partial p_t} = q_t\theta - 2q_t p_t - \lambda_1 + \delta\lambda_1\left(1 - X_{t+1}\right) = 0 \\
&\frac{\partial V_t}{\partial e_t} = \frac{\partial \pi_t\left(p_t, q_t, \boldsymbol{e_t}, e^*_{t-1}, .., e^*_1\right)}{\partial e_t} + \sum_{j=1}^{\infty} \delta^j \frac{\partial \pi_{t+j}\left(p^*_{t+j}, q^*_{t+j}, e^*_{t+j}, .., \boldsymbol{e_t}, e^*_{t-1}, .., e^*_1\right)}{\partial e_t} = \\
&\quad = -e_t + \beta + \alpha\tfrac{\delta}{1-\delta} = 0 \\
&\frac{\partial V_t}{\partial q_t} = \frac{\partial \pi_t\left(p_t, \boldsymbol{q_t}, e_t, e^*_{t-1}, .., e^*_1\right)}{\partial q_t} = p_t\left(\theta - p_t\right) - q_t + \lambda_2 - \delta\lambda_2 = 0 \\
&\frac{\partial V_t}{\partial \lambda_1} = p_{t-1}\left(1 - X_t\right) - p_t \geq 0 \\
&\frac{\partial V_t}{\partial \lambda_2} = q_t - q_{t-1} \geq 0,
\end{aligned} \tag{8}$$

where λ_1 and λ_2 are slack variables. When the quality constraint is not binding,

[11] From Eq. (9) we see that the marginal cost of electricity distribution affects price and quality in equilibrium. However, the equilibrium level of efficiency effort is not affected. Since we focus on the effects of persistent and transient components of efficiency, we avoid further mathematical complications by assuming that marginal costs of electricity distribution are negligible. Clearly, marginal costs of quality and efficiency effort are still present in the following analysis.

[12] Note that we rule out any firm's strategic behaviour that raises prices to anticipate future constraints and influence efficiency targets. The model assumes that firms do not have information on the timing of introduction of the new price cap regulation or the discount factor on future profits is relatively large.

the solution to the constrained maximization is:

$$\begin{array}{l} p_t^* = p_{t-1}^* (1 - X_t) = \Phi_t \\ e_t^* = \beta + \alpha \frac{\delta}{1-\delta} \\ q_t^* = p_{t-1}^* (1 - X_t) \left[\theta - p_{t-1}^* (1 - X_t)\right] = \Phi_t (\theta - \Phi_t) > q_{t-1}^* \\ \lambda_1^* = \frac{q_t^* \theta - 2 q_t^* p_t^*}{[1-\delta(1-X_{t+1})]} = \frac{\Phi_t (\theta - \Phi_t)(\theta - 2\Phi_t)}{[1-\delta(1-X_{t+1})]} \\ \lambda_2^* = 0 \end{array} \tag{9}$$

where $\Phi_t = P_0 \prod_{j=1}^{t} (1 - X_j)$. Conversely, when minimum quality standards are binding, we have:

$$\begin{array}{l} p_t^* = p_{t-1}^* (1 - X_t) = \Phi_t \\ e_t^* = \beta + \alpha \frac{\delta}{1-\delta} \\ q_t^* = q_{t-1}^* = \Phi_{t-1} (\theta - \Phi_{t-1}) \\ \lambda_1^* = \frac{q_{t-1}^* \theta - 2 q_{t-1}^* p_t^*}{[1-\delta(1-X_{t+1})]} = \frac{\Phi_{t-1} (\theta - \Phi_{t-1})(\theta - 2\Phi_t)}{[1-\delta(1-X_{t+1})]} \\ \lambda_2^* = \frac{q_t^* - p_t^* (\theta - p_t^*)}{1-\delta} = \frac{\Phi_{t-1} (\theta - \Phi_{t-1}) - \Phi_t (\theta - \Phi_t)}{1-\delta}. \end{array} \tag{10}$$

where $\Phi_{t-1} = P_0 \prod_{j=1}^{t-1} (1 - X_j)$.

Note that the equilibrium level of efficiency effort increases with persistent (α) and transient (β) marginal efficiency gains and the discount factor (δ) on future earnings. Contrary to the price cap constraint, the quality constraint may or may not be binding. From solution (9) we can see that an increase in the efficiency target may lead to lower quality if $\partial q_t^* / \partial X_t = p_{t-1}^* \left[2 p_{t-1}^* (1 - X_t) - \theta\right] < 0$ because the price cap affects quality through the efficiency target X_t. This happens if the efficiency target is set at too high levels, or $X_t > 1 - \theta / 2 p_{t-1}^*$, which is increasingly more likely with subsequent reductions in prices. Therefore, the situation may deteriorate over time even if quality is initially above the minimum quality standard. For high efficiency targets the level of quality is expected to fall towards solution (10) with binding minimum quality standards.

From Eqs. (2) and (9)-(10) we see that a perfectly informed regulator estimates transient efficiency gains of $E^T = \beta e_t^*$. Note, however, that transient efficiency gains in each period are already considered in the first-period efficiency target (X_1) to set the price cap $p_1 \leq P_0 (1 - X_1)$. Therefore, transient efficiency gains should not reduce the price further in the following periods, and should not be added to efficiency targets set for the following periods (X_t with $t > 1$). Clearly, this does not imply that the level of transient efficiency in the regulatory period is zero.

As for persistent efficiency, note that in the first period ($t = 1$) persistent efficiency gains are zero ($E_1^P = 0$) since no effort is assumed in period 0 before the first regulatory period. For $t > 1$ persistent efficiency gains are $E_t^P = \sum_{j=1}^{t-1} \alpha e_{t-j}^*$. Since persistent efficiency gains generated in periods $t-2, .., 1$ are taken into account in previous periods efficiency targets, the only persistent effect not yet considered in the current period is the one-period lead time effect of effort made in the previous period: $\alpha e_{t-1}^* = E_t^P / (t-1)$.

Using the above result for the equilibrium price (9) we can write the price cap constraint (6) as $p_t \leq P_0 \prod_{j=1}^{t} (1 - X_j) = P_0 (1 - X_1) \prod_{j=2}^{t} (1 - X_j)$, where X_1 represents the first-period efficiency target which includes only transient efficiency gains. The remaining efficiency terms X_j represent new persistent efficiency gains that can be obtained in each of the following periods. Optimal efficiency targets per unit of output can then be written as:

$$X_1^* = \frac{E^T}{s(p_1^*, q_1^*)} = \frac{\beta \left(\beta + \alpha \frac{\delta}{1-\delta} \right)}{s(p_1^*, q_1^*)} \tag{11}$$

$$X_{t>1}^* = \frac{E_t^P / (t-1)}{s(p_t^*, q_t^*)} = \frac{\alpha \left(\beta + \alpha \frac{\delta}{1-\delta} \right)}{s(p_t^*, q_t^*)}. \tag{12}$$

We can see that optimal efficiency targets increase with persistent (α) and transient (β) marginal efficiency gains and the discount factor (δ) on future earnings. In this case, transient and persistent efficiency gains are fully internalized by the informed regulator.

2.2 Imperfect information on efficiency structure

When the regulator is imperfectly informed about achievable efficiency gains, efficiency targets cannot be set at the optimal level defined by Eqs. (11)-(12). This point has been debated in the theoretical literature (see for instance Bernstein and Sappington, 1999), and recent empirical research developed more sophisticated methods to address imperfect information on firm's productivity measurement. In fact, some regulators around the world are using empirical methods to estimate the level of efficiency. However, an additional concern arises about the regulator's ability to disentangle different types of efficiency, even when total efficiency gains are correctly estimated.

Let us assume that the regulator underestimates persistent inefficiency (or overestimates transient efficiency). This happens for instance when the regulator does not expect a delay in the effect of the efficiency effort, hence $E_1^P(e_1) > 0$. In other words, the α component is expected to contribute to cost reductions (also) in the current period. Consequently, we can modify the cost function above (2) using $\sum_{j=0}^{t} \alpha e_{t-j}$ instead of $\sum_{j=1}^{t} \alpha e_{t-j}$. Solving the maximization problem with the modified (dystopian) cost function that 'anticipates' some efficiency gains, we obtain $e_t^* = \beta + \alpha \frac{1}{1-\delta}$, which is higher than the equilibrium level of effort in Eq. (9).[13] The new efficiency targets will be:

$$X_1^{IF} = \frac{E^T + E_1^P}{s(p_1^*, q_1^*)} = \frac{(\beta + \alpha)\left(\beta + \alpha \frac{1}{1-\delta}\right)}{s(p_1^*, q_1^*)} > X_1^* \tag{13}$$

$$X_{t>1}^{IF} = \frac{E_t^P / t}{s(p_t^*, q_t^*)} = \frac{\alpha\left(\beta + \alpha \frac{1}{1-\delta}\right)}{s(p_t^*, q_t^*)} > X_{t>1}^*. \tag{14}$$

Because the regulator estimates a higher efficiency effort, he will set tighter efficiency targets. This will reduce firm's marginal profitability. A possible consequence is that firms may want to postpone or lessen important expenditures leading to poor quality service in the current and future periods. From Eq. (9) we saw that $\partial q_t^* / \partial X_t < 0$. Consequently, tighter quality controls are probably required to avoid too low quality levels.

2.3 Unobserved quality

Poor information regarding quality may exacerbate the possible consequences of too high efficiency targets leading to disappointing quality compliance. Let us assume that service quality is affected by random shocks (interruptions) during the regulatory period. Therefore, the regulator cannot observe the true level of quality:

$$q_t^* = \hat{q}_t + \tilde{\epsilon}, \tag{15}$$

[13] A similar conclusion can be drawn if the regulator does not expect at all long-lasting savings, and assumes that all cost inefficiency is transitory. We can substitute $\sum_{j=1}^{t} \alpha e_{t-j}$ with ϕe_t (with $\phi \geqslant \alpha$) into the cost function and solve again the constrained maximization problem. Then, we get $e_t^* = \beta + \phi$, which is higher than $e_t^* = \beta + \alpha \frac{\delta}{1-\delta}$ obtained in Eq. (9) if $\frac{\phi}{\alpha} > \frac{\delta}{1-\delta}$, i.e. for relatively low levels of the discount factor (δ) or relatively high expectations of marginal efficiency (ϕ).

where $\hat{q}_t$ is the observed level of service quality and $\tilde{\epsilon}$ is the exogenous shock distributed as:

$$\tilde{\epsilon} \in \{\epsilon, -\epsilon\}, \quad \Pr[\tilde{\epsilon} = \epsilon] = \rho, \quad \Pr[\tilde{\epsilon} = -\epsilon] = 1 - \rho. \tag{16}$$

The probability ρ is unknown to the regulator.

Because of exogenous interruptions, the regulator can only verify if the observed level of quality is in the acceptable range $q^*_{t-1} - \epsilon \leq \hat{q}_t \leq q^*_{t-1} + \epsilon$. Since an observed level of quality below q^*_{t-1} is tolerated if it is not too low, i.e. $\hat{q}_t \geq q^*_{t-1} - \epsilon = q_{min}$, this may provide room for speculation by electricity distribution companies. Under pressure, companies may be more prone to rely on positive shocks, thus speculating on the expected level of service quality. Indeed, the expected level of observable quality is:

$$\hat{q}_t = \rho(q^*_t + \epsilon) + (1 - \rho)(q^*_t - \epsilon) = q^*_t + \epsilon(2\rho - 1). \tag{17}$$

This may fall below q_{min} if:

$$q^*_t + \epsilon(2\rho - 1) < q^*_{t-1} - \epsilon. \tag{18}$$

This inequality depends on the efficiency target set for the current period (X_t). Since q^*_t is decreasing for relatively high levels of the efficiency target, the above inequality is more likely to be satisfied when the regulator underestimates cost persistency. Therefore, too high efficiency targets may increase violations of MQS, *ceteris paribus*, leading to lower quality compliance.

3 Cost model specification and estimation method

Our theoretical model hypothesizes that there are two components of inefficiency. If the authority neglects these components, the effectiveness of the regulation may be undermined. Consequently, the effort to separate transient and persistent inefficiency could improve the performance of the electricity distribution market. In the following empirical analysis, we show that data availability and the application of econometric models allow to disentangle transient and persistent inefficiency. Moreover, we provide some evidence that the presence of persistent efficiency may result in poorer quality levels.

The total cost of an electricity distribution company can be specified as a function of input prices and outputs. Moreover, as discussed in Filippini and Wetzel (2014) and in previous studies on the cost structure of electricity distribution companies, in the cost model specification it is important to include a number of output characteristic variables.[14] These variables should take into account the heterogeneity of the electricity distribution companies' production environment.

Generally, the explanatory variables considered in the specification of a cost function for electricity distribution companies are: the quantity of electricity distributed, the number of customers and the factor prices, some output characteristics such as customer density, network size, service area, service quality and load factor.

In this analysis, we specify a total cost function with two outputs and three output characteristics. Unfortunately, the cost model specification does not include input prices since these data are lacking. Consequently, we hypothesize that all electricity distribution companies are exposed to the same input prices.[15]

The total cost can be written as:

$$TC = c\left(Y, CU, NL, LF, Q, T\right), \tag{19}$$

where Y and CU represent the output measured by the electricity supplied in kilowatt-hours and the number of final consumers, respectively. NL, LF and Q are output characteristics: NL is the network length, LF denotes the load factor, and Q is service quality measured by $SAIDI$, an index of the average interruption duration of the system. Finally, T is a time trend that captures changes in the cost over time. In order to be able to compute three type of economies, i.e. economies of output density, economies of customers' density and economies

[14]For a discussion on the estimation of cost functions in the energy sector see Farsi and Filippini (2009).

[15]This assumption is used also in previous studies using data from electricity distribution companies in New Zealand (Nillesen and Pollit, 2011; Filippini and Wetzel, 2014) as well as by the regulator. It is worth noting that New Zealand is characterized by an open economy that works on free market principles. This means that the level of interest rates, the price of inputs, the average salaries are similar across different regions. The market for inputs is quite competitive and prices are expected to be similar across distribution companies. Finally, in our GTRE model, part of the possible differences in prices are captured by the individual company effect.

of scale, we use the network length instead of customer density previously used by Filippini and Wetzel (2014). As indicated by the microeconomic theory of production, the cost function should be concave in input prices, non-decreasing in input prices and output, and linearly homogeneous in input prices.[16]

For the estimation of the cost function (19), we use a translog functional form. The translog has the advantage that it does not impose *a priori* restrictions on the nature of the technology. However, in case the model specification includes some variables relatively highly correlated, then the estimation of the translog cost function can suffer from multicollinearity. In our case, some of the explanatory variables, such as the number of customers, the network length and the load factor, are highly correlated and cause problems in the econometric estimation. Therefore, we estimate a reduced version of the translog, where all interaction variables between the two outputs and output characteristics have been dropped. Based on Eq. (19) the reduced translog cost function can be expressed as:

$$\begin{aligned}\ln TC_{it} = {} & \beta_0 + \beta_Y \ln Y_{it} + \beta_{CU} \ln CU_{it} + \beta_{LF} \ln LF_{it} + \beta_Q \ln Q_{it} + \\ & + \beta_{NL} \ln NL_{it} + \frac{1}{2}\beta_{YY}(\ln Y_{it})^2 + \frac{1}{2}\beta_{CUCU}(\ln CU_{it})^2 + \\ & + \frac{1}{2}\beta_{LFLF}(\ln LF_{it})^2 + \frac{1}{2}\beta_{QQ}(\ln Q_{it})^2 + \frac{1}{2}\beta_{NLNL}(\ln NL_{it})^2 + \\ & + \beta_{YCU} \ln Y_{it} \ln CU_{it} + \beta_t T_t + \varepsilon_{it}, \end{aligned} \tag{20}$$

where the subscripts i and t denote the firm and year, respectively; and the βs are unknown parameters to be estimated. The error term in Eq. (20) is still general and will be specified later from an econometric point of view (see Table 1).

As discussed in more details in Filippini and Greene (2016), several different panel data stochastic frontier models (SFA) can be used to estimate the level of productive inefficiency. Some of these models estimate the persistent part of productive inefficiency, although with different levels of precision. Others estimate the transient component. Moreover, some recent developed models provide information on both types of productive inefficiency.

[16]Due to the fact that in the model specification (19) we are not considering input prices, some of the properties of the cost function e.g. the concavity in input prices, cannot be verified.

In this paper, we decided to use three alternative stochastic frontier models: two classical models and one relatively new model. The first model is the basic version of the random effects model proposed by Pitt and Lee (1981) (RE hereafter); the second model is the so-called true random effects model (TRE hereafter) proposed by Greene (2005a, 2005b); and the third model is the generalized true random effects model (GTRE) recently introduced by Colombi et al. (2014) and Filippini and Greene (2016).

The RE model considers the individual random effects as inefficiency rather than unobserved heterogeneity as in the usual panel data models. This model provides information on the persistent part of the inefficiency. One problem with the RE is that any time-invariant, firm-specific unobserved heterogeneity is considered as inefficiency. As a result, the values obtained by this model are not precise and tend to underestimate the level of persistent efficiency in electricity distribution.

The TRE proposed by Greene (2005a and 2005b) extends the SFA model in its original form (Aigner, et al., 1977) by adding an individual random effect in the model. In general terms, for the TRE the constant term, β_0, in Eq. (20), is replaced with a series of firm-specific random effects. The improvement provided by this model is to control for unobserved variables that are constant over time. However, any time-invariant component of inefficiency is captured by the firm-specific constant terms. Therefore, the TRE tends to overestimate the level of efficiency. Generally, the TRE provide information on the time-varying part of the inefficiency.

The third model (GTRE) offers the possibility to estimate at the same time the persistent and transient part of inefficiency. Colombi et al. (2014) proposed a theoretical platform to distinguish persistent from transient inefficiency. Filippini and Greene (2016) suggest a practical completion to the development by proposing a straightforward, transparent empirical estimation method of the GTRE. It is worth noting that TRE model provides similar estimates of the transient part of inefficiency obtained using the GTRE model. This is because, the transient part of inefficiency in the TRE and GTRE models is specified in the same way. Table 1 summarizes the three econometric specifications including the structure of the error term in Eq. (20). In comparison to the RE and the

TRE models, the error term in the GTRE model is composed of four parts. Two of them (h_i and u_{it}) are half-normal distributed and measure the two inefficiency components. The other two parts (w_i and v_{it}) represent the usual noise and are two-sided normally distributed.

4 Data

The dataset used in this study is a panel of 28 New Zealand's electricity distribution businesses (EDBs) between 2000 and 2011.[17] The panel is constructed mainly by exploiting information in the "NZ EDB Database" from Economic Insights (Economic Insights, 2009). This database consists of financial and production data on electricity distribution companies. As required by the New Zealand electricity regulation, financial and production data are yearly published in the Electricity Information Disclosures.

In terms of the number of connected customers, the size of companies in our sample varies between $4,100$ and $680,000$. Total cost is defined as the sum of variable cost and capital cost. Variable cost includes the operating expenses for labor, materials and services.[18] The capital cost is the sum of capital depreciation (or amortization) and the opportunity cost of alternative use of assets. Regrettably, consistent information on capital cost for the whole period under observation are not present in the Electricity Information Disclosures. However, we can use the assets value and some assumptions regarding depreciation and opportunity cost to approximate the capital cost. For the assets value, we use the so-called optimized deprival value (ODV), which is the annual monetary value of the system fixed assets of each distribution company. The ODV captures the loss of value that a company would bear if deprived of assets.[19] Following Lawrence

[17]Few companies have been excluded because of lack of information. Further, a new company recently established has been excluded because of too few years of operation. Conversely from the dataset used by Filippini and Wetzel (2014), our dataset includes only companies that have already introduced unbundling of their activities. For more details on the data and definition of variables see Filippini and Wetzel, 2014.

[18]In New Zealand, the generation, the transmission and the distribution of electricity are separated. Hence, there is no vertical integration. Ownership unbundling characterizes the whole period of our data (2000-2011). The Electricity Industry Reform Act (EIRA) in 1998 created three competing publicly-owned companies in the generation sector and the ownership separation of distribution from retailers (Shen and Yang, 2012).

[19]The ODV methodology used for asset valuations in the New Zealand's electricity distribu-

et al. (2009) and Filippini and Wetzel (2014), we set a common depreciation rate of 4.5% of ODV and an opportunity cost rate of 8% of ODV. Consequently, capital cost is nearly 12.5 $(4.5 + 8)$ percent of the annual ODV.[20] Total cost is adjusted for inflation using the consumer price index for New Zealand provided by the OECD (base 2005).

In addition to outputs, we consider three output characteristic variables: the load factor, network quality and network length. The load factor captures the intensity in utilization of the distribution network. This is measured by the ratio between the electricity supplied and the maximum distribution transformer demand multiplied by the total number of hours in one year. Lower costs are expected for distribution companies with higher rates of network utilization. Therefore, the estimated coefficient of the load factor is expected to show a negative sign.

The network quality characteristic is measured by SAIDI. This is the average number of interruption minutes for a consumer within a given period. The impact of SAIDI on total costs is rather unclear. On the one hand, higher quality, that is a lower SAIDI, may require more investments and hence may induce higher capital costs. The higher quality may also lead to lower operational costs. For the estimation of our models, we decided to use a weighted level of quality instead of the actual level of quality. The weighted level of quality is the moving average of the actual level of quality over the previous five years. During the period under study, the regulator in New Zealand set a minimum level of quality using the previous five-year average of SAIDI. By using the regulated level of SAIDI, we can limit the endogeneity problem related to quality and take into account the relatively high volatility of SAIDI.

Finally, the network length is measured in kilometers to approximate the service area size. We expect a positive coefficient, indicating that companies with a larger area size operate at higher costs than companies with smaller area size do. Some descriptive statistics of the variables used in this study are

tion sector is described in detail by the New Zealand Commerce Commission (2004).

[20]Clearly, the limitations arising from the use of ODV as a proxy for capital cost are not ignored. Nevertheless, in the absence of a consistent alternative measure this is plausibly the best proxy for capital cost.

provided in Table 2.

5 Results

The estimation results for the three models are given in Table 3.[21] These results show that the coefficients of output, number of customers and network length are positive and significant across all different estimators. In general, the estimated coefficients are relatively similar across the estimators, except for the coefficients of the two outputs and the coefficient of quality. λ is the ratio of the standard deviation of the inefficiency term u_{it} to the standard deviation of the stochastic term v_{it}. This ratio is significant and reflects the relatively low contribution to the decomposed error term ε_{it}. The standard deviations of the time fixed symmetric effects (σ_w) and the time fixed one-sided effects (σ_h) are also significant.

Since total costs and regressors are in logarithms and normalized to the median value, the first order coefficients are interpretable as cost elasticities evaluated at the sample median. All these coefficients have the expected sign and are highly significant. For instance, the output coefficients suggest that the increase in costs due to a one percent increase in the number of KWh of electricity distributed, keeping all other explanatory variables constant, varies between 0.16 and 0.33%. As we see from Table 3 the interaction term between the two outputs is negative and statistically significant. This suggests the presence of cost complementarities between electricity and number of customers, i.e. companies with a higher number of customers have a relatively low marginal cost for distributing electricity.

The coefficient of the network length suggests that the increase in costs due to a one percent extension in the network, keeping all other explanatory variables constant, is approximately 0.2%. Further, the coefficient of the number of customers suggests that the increase in costs due to a one percent increase in the number of customers, keeping all other explanatory variables constant, varies between 0.36 and 0.52%.

The coefficient of the time trend is positive and indicates that total costs of electricity companies increased over time. This result is apparently counterin-

[21] The estimates are obtained using the software NLogit, version 5.

tuitive from a theoretical point of view since the time trend should capture the presence of technical change. However, from an econometric point of view we should keep in mind that the time trend captures the impact of several factors that change over time and affect in the same way all companies. For instance, one possible explanation for the positive sign could be a general increase of input prices for all companies (not controlled in the model) or a new regulation that imposes extra costs to the company.

The cost elasticity with respect to the load factor is negative in all specifications of the cost model, indicating that a 1% improvement in the load factor reduces costs by approximately 0.1%. Finally, the quality index measured by the regulated level of interruptions (SAIDI) has a negative and significant impact on costs, though this impact is quite small. A 1% decrease in quality (i.e. higher number of interruptions) decreases costs between 0.01 and 0.03%, *ceteris paribus*. This suggests that reducing service quality allows firms to save on costs.

5.1 Persistent and transient efficiency

The firm's inefficiency for the RE and the TRE models are estimated using the conditional mean of the inefficiency term proposed by Jondrow et al. (1982). Following Filippini and Greene (2016) and using a result from Colombi (2010) based on the moment generating function for the closed skew normal distribution, we compute the inefficiency scores for the GTRE specification.

Table 4 provides descriptive statistics for the cost efficiency estimates for the 28 electricity distribution companies obtained from the econometric estimation of the three models. The estimation results for the new cost frontier model (GTRE) provide estimates of the persistent (PGTRE) as well as the transient component of cost efficiency (TGTRE). The RE model produces values of the cost efficiency that are time-invariant and, therefore, should reflect the persistent part of the cost efficiency. On the other hand, the TRE model produces values that are time varying and, therefore, should reflect the transient part of the cost efficiency.

The results reported in Table 4 show that the estimated average values of persistent efficiency vary from 78% in the RE model to 88% in the GTRE model. Moreover, the estimated average values of the transient efficiency vary from 94%

in the TRE model to 88% in the GTRE model. Note that the values of the persistent and transient efficiency obtained with the GTRE model are different from the values obtained with the TRE and the RE models. This suggestive evidence implies that efficiency scores obtained with the RE and TRE models do not provide precise information on the level of persistent and transient efficiency. Finally, Table 5 provides the Pearson's correlations between the estimated levels of cost efficiency obtained from the three model specifications. The correlation between the levels of transient cost efficiency obtained with TRE and GTRE models is relatively high (0.78). However, the correlation between the values of the persistent cost efficiency obtained with RE and GTRE models is lower (0.43). This suggests that the result obtained with the RE model is not measuring the persistent efficiency of the firms correctly. Moreover, the values of the Spearman's rank correlation confirm the values obtained with the pairwise Pearson's correlation.[22] As suggested by Greene (2005b), the reason of such differences could be that all unobserved time invariant heterogeneity in the RE model is captured by the individual effect, which is also used to compute the level of efficiency.

The evidence on the presence of persistent efficiency casts doubts on the effectiveness of a price cap regulation that does not distinguish the two parts of efficiency. As suggested by the theoretical model in Section 2, when persistent and transient efficiency are not estimated correctly the regulator may assume wrong efficiency targets.

5.2 Efficiency and quality

Theoretically, the regulatory implications of persistent and transient inefficiency could be assessed by comparing the regulated setting with an ideal setting without price cap and quality regulation. Unfortunately, this experimental design cannot be performed with our dataset. Still, some figures are worth discussing and maybe can stimulate opportunities for future research.

The theoretical model suggests that, if the regulator underestimates cost persistency, service quality may decrease because of the efficiency target is too

[22]The difference in the efficiency ranking of firms obtained from the three models is also confirmed by the Kruskal-Wallis rank test.

ambitious. In New Zealand, one of the objectives of the regulator is to provide services at a quality that reflects consumer preferences, and the use of total factor productivity should allow setting efficiency targets according to this objective (Brown and Moselle, 2008). Within this framework, the regulator does not disentangle the two parts of efficiency, i.e. the persistent and the transient components.

Note that, if persistent efficiency is ignored the levels of quality and quality compliance may suffer. Some preliminary evidence is provided by the estimated level of persistent efficiency that appears to be positively correlated with quality ($\rho = 0.2$). Further, the number of firms that do not comply with the regulated quality level is remarkable. Around 37% of firms on average across the whole period are below the regulated quality standard, ranging from a minimum of 15% in 2001 and a maximum of 68% in 2007.

To further investigate this issue, we can build a measure of the information bias in the estimation of persistent efficiency. We compare the estimates of inefficiency from a conventional stochastic frontier pooled model (Aigner et al., 1977) with the estimates of the GTRE model. In most countries (e.g. Austria and Germany), national authorities adopt simple frontier models based on cross-sectional data to estimate the level of efficiency of electricity distribution companies. This may lead to biased results since persistent efficiency is not correctly taken into account. Therefore, we can use the results of the GTRE model as a benchmark of correct information on persistent efficiency and compare them with the efficiency results of a pooled model that ignores the magnitude of persistent efficiency and considers all the inefficiency as transient. In this way, we obtain a proxy of the information bias in the estimation of persistent efficiency. As expected, we observe that the larger the information bias the higher is the number of firms that do not comply with the regulated quality ($\rho = 0.18$). Although this correlation is small, the result may cast doubts on the ability of the regulator to ensure quality compliance when an information bias is present on the estimation of persistent efficiency.

5.3 Economies of scale

The estimation results reported in Table 3 can also be used to compute the value of economies of scale under different econometric models. More precisely, the inclusion of the number of customers and the network length in the cost function allows us to derive economies of output density (EOD), economies of customer density (ECD) and economies of scale (ES). We follow Roberts (1986) and Filippini (1998) to define these three types of economies.

First, economies of output density are defined as the inverse of the percentage change in total costs following a percentage change in the output, assuming that input prices, the load factor, the number of customers and the network length are unchanged. This corresponds to the inverse of the elasticity of total costs with respect to output: $EOD = 1 \div \left(\frac{\partial lnTC}{\partial lnY}\right)$. Second, economies of customer density are expressed as the inverse of the percentage change in total costs generated by a percentage change in the output and the number of customers, holding constant input prices, the load factor and the network length. Therefore, economies of customer density can written as: $ECD = 1 \div \left(\frac{\partial lnTC}{\partial lnY} + \frac{\partial lnTC}{\partial lnCU}\right)$. Finally, economies of scale are defined as the inverse of the percentage change in total costs caused by a percentage change in the output, the number of customers and the size of the service area, holding all input prices and the load factor fixed. Economies of scale (ES) can then be expressed as: $ES = 1 \div \left(\frac{\partial lnTC}{\partial lnY} + \frac{\partial lnTC}{\partial lnCU} + \frac{\partial lnTC}{\partial lnNL}\right)$.

Based on definitions above, we observe economies of output density if $EOD > 1$, i.e. if the average cost of an electricity distribution utility decreases as the volume of electricity sold to a fixed number of customers in a service area of a given size increases. Similarly, there are economies of customer density if $ECD > 1$. This measure is relevant for analyzing the cost of distributing more electricity to a fixed service area as it becomes more densely populated. Finally, economies of scale are present when $ES > 1$.

Table 6 reports the estimates of the three types of economies for a medium sized firm.[23] All indicators are greater than 1, ranging from 1.119 in the GTRE

[23] All economies have been evaluated at the values for the load factor, SAIDI and consumer density of the median company. For the interpretation of the results, it is important to note that a proportional increase in electricity supplied and the number of consumers imply, keeping the value of the consumer density constant, an increase in the network length.

model to 1.127 in the TRE model.[24] These results tend to support the hypothesis that the electricity distribution sector is characterized by economies of scale, output and consumer density as obtained in other empirical studies.[25] For instance, a recent study by Tovar et al. (2011) on the Brazilian electricity distribution sector found a median value of economies of scale of approximately 1.48. It is worth noting that the values reported in Table 6 are slightly larger than the values provided in a previous study by Filippini and Wetzel (2014) using New Zealand data. This difference may be due to the relatively simple stochastic frontier model used in Filippini and Wetzel (2014).

6 Conclusions

The level of productive efficiency of a firm can be split in two parts: a persistent and a transient component. This distinction can be important in the application of incentive-based regulation schemes, such as the price cap that uses inefficiency scores in the definition of prices in water, electricity and telecommunication sectors. If the regulator ignores or underestimates persistent efficiency, efficiency targets can be wrongly set. As a consequence, this may lead to quality distortion and lower quality compliance.

Generally, the empirical literature on efficiency analysis of firms has not paid a lot of attention to the distinction between these two components. Some scholars (Colombi et al., 2014; Kumbhakar and Tsionas, 2012; Kumbhakar et al., 2012; Filippini and Greene, 2016) have recently proposed econometric approaches to provide separate estimates of the two components of efficiency. Some of these approaches are relatively cumbersome. In this paper, we apply the estimator proposed by Filippini and Greene (2016) to assess the level of persistent and transient efficiency in the New Zealand electricity distribution sector. The estimator is based on maximum simulated likelihood using all the sample distribu-

[24] Using the Delta method, we computed the 95% confidence intervals for the three types of economies for the three models. The confidence intervals for the RE model are different from the confidence intervals obtained using the GTRE and the TRE models. This is not surprising because in the RE model the output coefficient is not statistically significant. The confidence intervals in the GTRE and the TRE models are relatively similar.

[25] For a review of previous studies on economies of scale and density in transmission and distribution of electricity see Ramos-Real (2005).

tional information to obtain the estimates, and is very effective and strikingly simple to apply.

Few studies analyze the level of cost efficiency of New Zealand electricity distribution companies (Filippini and Wetzel, 2014; Nillesen and Pollitt, 2011; Scully, 1999) and none of these studies distinguishes the two components of cost efficiency. Our empirical results show that the transient and the persistent parts of productive efficiency are relatively different in absolute value and differ from productive efficiency measured by previous approaches. From a regulatory point of view, following the theoretical model, the results imply that differentiated measures of efficiency should be used in regulation. For instance, the regulator could set (transient) efficiency targets for each year within the regulatory period. In addition, (persistent) efficiency targets that require more years to be achieved could be verified only at the end of the regulatory period. Due to the presence of persistent inefficiency, the regulatory period should be longer than the usual five-year period defined by the regulation authorities. Therefore, a more flexible price-cap rule combining short-run efficiency targets (to reduce transient inefficiency) and long-run efficiency targets (to reduce persistent inefficiency) could provide some regulatory improvements.

We found some suggestive evidence that higher levels of persistent efficiency are positively correlated with quality levels. However, electricity distribution companies seem to suffer systematically from poor quality compliance. Moreover, quality compliance is decreasing with the information bias of the regulator in the estimation of persistent efficiency. Further research is needed to confirm these findings.

References

Aigner D, Lovell CAK, Schmidt P (1977) Formulation and estimation of stochastic frontier production function models. Journal of Econometrics 6(1):21-37

Blázquez-Gómez L, Grifell-Tatjé E (2011) Evaluating the regulator: Winners and losers in the regulation of Spanish electricity distribution. Energy Economics, 33(5):807-815

Bernstein JI, Sappington DEM (1999) Setting the X factor in price-cap regulation plans. Journal of Regulatory Economics 16:5-25

Brown T, Moselle B (2008) Use of total factor productivity analyses in network regulation case studies of regulatory practice. The Brattle Group Ltd.

Chen Z, Pestana C, Borges M, (2015) A Bayesian stochastic frontier analysis of Chinese fossil-fuel electricity generation companies. Energy Economics 48:136-144

Cowan SGB (1997a) Tight average revenue regulation can be worse than no regulation. The Journal of Industrial Economics 45 (1):75-88

Cowan SGB (1997b) Price-cap regulation and inefficiency in relative pricing. Journal Regulatory Economics 12:53-70

Colombi R (2010) A skew normal stochastic frontier model for panel data. Proceedings of the 45th Scientific Meeting of the Italian Statistical Society, University of Padua, June 29-July 1

Colombi R, Kumbhakar SC, Martini G, Vittadini G (2014) Closed-skew normality in stocastic frontiers with individual effects and long/short-run efficiency. Journal of Productivity Analysis 42:123-136

Cullmann A, Nieswand M (2016) Regulation and investment incentives in electricity distribution: An empirical assessment. Energy Economics, 57:192-203

Di Giorgio L, Filippini M, Masiero G (2015) Structural and managerial cost differences in nonprofit nursing homes. Economic Modelling 51:289-298

Economic Insights (2009) Economic Insights NZ EDB Database. http://www.comcom.govt.nz/assets/Imported-from-old-site/industryregulation/Electricity/PriceQualityPaths/ContentFiles/Documents/comcom-economicinsightedbdatabaseandanalysisdatafiles-aug2009.zip, accessed February 2012

Farsi M, Filippini M (2009) An analysis of cost efficiency in Swiss multi-utilities. Energy Economics 31(2):306-315

Filippini M, Greene W (2016) Persistent and transient productive inefficiency: a maxi-

mum simulated likelihood approach. Journal of Productivity Analysis 45(2): 187-196.

Filippini M, Koller M, Masiero G (2015) Competitive tendering versus performance-based negotiation in Swiss public transport. Transportation Research Part A: Policy and Practice 82: 158-168

Filippini M, Wetzel H (2014) The impact of ownership unbundling on cost efficiency: Empirical evidence from the New Zealand electricity distribution sector. Energy Economics 45:412-418

Ghosh R, Kathuria V (2016) The effect of regulatory governance on efficiency of thermal power generation in India: A stochastic frontier analysis. Energy Policy 89:11-24

Greene W (2005a) Reconsidering heterogeneity in panel data estimators of the stochastic frontier model. Journal of Econometrics 126(2):269-303

Greene W (2005b) Fixed and random effects in stochastic frontier models. Journal of Productivity Analysis 23(1):7-32

Jamasb T, Pollitt M (2007) Incentive regulation of electricity distribution networks: Lessons of experience from Britain. Energy Policy, 35(12): 6163-6187

Jondrow J, Knox Lovell CA, Materov IS, Schmidt P (1982) On the estimation of technical inefficiency in the stochastic frontier production function model. Journal of Econometrics 19(2-3):233-238

Joskow PL, Schmalensee R (1986) Incentive regulation for electric utilities. Yale Journal on Regulation 4:1

Kumbhakar SC, Lien G, Hardaker JB (2012) Technical efficiency in competing panel data models: a study of Norwegian grain farming. Journal of Productivity Analysis 41(2):321-337

Kumbhakar SC, Tsionas EG (2012) Firm heterogeneity, persistent and transient technical inefficiency: A generalized true random effects model. Journal of Applied Econometrics 29(1):110-132

Iossa E, Stroffolini F (2002) Price cap regulation and information acquisition. International Journal of Industrial Organization 20:1013-1036

Laffont J-J, Tirole J (1993) A theory of incentives in procurement and regulation. MIT press

Lawrence D, Diewert E, Fallon J, Kain J (2009) Electricity distribution industry productivity analysis: 1996-2008. Report prepared by Economic Insights for the New Zealand Commerce Commission

Mussa M, Rosen S (1978) Monopoly and product quality. Journal of Economic Theory 18:301-317

New Zealand Commerce Commission (2004) Handbook for optimised deprival valuation of system fixed assets of electricity lines businesses

New Zealand Commerce Commission (2015) Regulated Industries: Electricity archive. http://www.comcom.govt.nz/regulated-industries/electricity/, accessed September 2015

Nillesen PHL, Pollitt MG (2011) Ownership unbundling in electricity distribution: empirical evidence from New Zealand. Review of Industrial Organization 38:61-93

Phillips MA (2013) Inefficiency in Japanese water utility firms: a sotchastic frontier approach. Journal of regulatory Economics 44(2):197-214

Pitt MM, Lee L-F (1981) The measurement and sources of technical inefficiency in the Indonesian weaving industry. Journal of Development Economics 9(1):43-64

Ramos-Real FJ (2005) Cost functions and the electric utility industry. A contribution to the debate on deregulation. Energy Policy 33(1):69-87

Roberts MJ (1986) Economies of density and size in the production and delivery of electric power. Land Economics 378-387.

Shen D, Yang Q (2012) Electricity market regulation reform and competition - Case study of the New Zealand electricity market, In: Wu Y, Shi X, Kimura F (eds) Energy market integration in East Asia: Theories, electricity sector and substitutes, ERIA Research Project Report 2011-17, Jakarta, pp. 103-139

Scully GW (1999) Reform and efficiency gains in the New Zealand electrical supply industry. Journal of Productivity Analysis 11(2):133-147

Tovar B, Ramos-Real FJ, Fagundes de Almeida EF (2011) Firm size and productivity. Evidence from the electricity distribution industry in Brazil. Energy Policy 39(2): 826-833.

Walter M (2011) Some determinants of cost efficiency in German public transport. Journal of Transport Economics and Policy 45(1):1-20

Weyman-Jones TG (1990) RPI–X price cap regulation. Utilities Policy, 1(1):65-77

	Model I	*Model II*	*Model III*
	RE (Pitt and Lee)	**TRE**	**GTRE**
Model	$lnTC_{it} = \beta_0 + \beta' \mathbf{x}_{it} + v_{it} + u_i$	$lnTC_{it} = \beta_0 + w_i + \beta' \mathbf{x}_{it} + v_{it} + u_{it}$	$lnTC_{it} = \beta_0 + (w_i - h_i) + \beta' \mathbf{x}_{it} + v_{it} + u_{it}$
Full random error ε_{it}	$\varepsilon_{it} = u_i + v_{it}$ $u_i \sim N^+(0, \sigma_u^2)$ $v_{it} \sim N(0, \sigma_v^2)$	$\varepsilon_{it} = w_i + u_{it} + v_{it}$ $u_{it} \sim N^+(0, \sigma_u^2)$ $v_{it} \sim N(0, \sigma_v^2)$ $w_i \sim N(0, \sigma_w^2)$	$\varepsilon_{it} = w_i + h_i + u_{it} + v_{it}$ $u_{it} \sim N^+(0, \sigma_u^2)$ $h_i \sim N^+(0, \sigma_h^2)$ $v_{it} \sim N(0, \sigma_v^2)$ $w_i \sim N(0, \sigma_w^2)$
Persistent inefficiency estimator	$E(u_i \mid \varepsilon_{i1}, ..., \varepsilon_{iT})$	None	$E(h_i \mid \varepsilon_{it})$
Transient inefficiency estimator	None	$E(u_{it} \mid \varepsilon_{it})$	$E(u_{it} \mid \varepsilon_{it})$

Table 1: Econometric specifications of the stochastic cost frontier.

Variable	Unit of measurement	Mean	Std. Dev.	Min	Max
Total cost (TC)	Thousand 2005$	$38.5 * 10^3$	$64.0 * 10^3$	$3.4 * 10^3$	$378.0 * 10^3$
Electicity supplied (Y)	MWh	$970.5 * 10^6$	$1753.6 * 10^6$	$37.9 * 10^6$	$10700.0 * 10^6$
Consumers (CU)	Number	62775	114387	4108	679612
Load factor (LF)	Percentage	62.9	7.8	30.4	84.7
SAIDI (Q)	Minutes	221.4	192.5	15.0	1918.0
Network length (NL)	Km	5147.5	6007.3	239.0	30035.5

Number of observations: n=305

Table 2: Descriptive statistics.

Variable	**RE**	**TRE**	**GTRE**
β_0	16.359	16.560	16.449
	(499.981)	(1474.362)	(1380.217)
lnY	0.164	0.241	0.330
	(1.602)	(11.657)	(16.345)
$lnCU$	0.516	0.425	0.361
	(4.210)	(18.300)	(16.140)
$lnLF$	-0.139	-0.122	-0.144
	(-1.190)	(-2.785)	(-3.478)
lnQ	-0.011	-0.015	-0.025
	(-0.580)	(2.183)	(3.816)
$lnNL$	0.210	0.207	0.213
	(2.494)	(15.017)	(16.526)
$lnY{*}lnY$	0.277	0.202	0.043
	(1.147)	(2.155)	(0.460)
$lnCU{*}lnCU$	0.424	0.326	0.211
	(2.023)	(3.435)	(2.227)
$lnLF{*}lnLF$	-0.257	-0.164	-0.150
	(-0.499)	(-0.619)	(-0.719)
$lnQ{*}lnQ$	0.011	0.014	0.033
	(0.335)	(0.934)	(2.421)
$lnNL{*}lnNL$	-0.143	-0.157	-0.181
	(-1.686)	(-15.810)	(-18.351)
$lnY{*}lnCU$	-0.333	-0.232	-0.088
	(-1.441)	(-2.548)	(-0.976)
T	0.020	0.020	0.021
	(11.868)	(22.558)	(21.754)
σ_w	-	0.192	0.189
	-	(36.563)	(37.417)
λ	4.981	1.734	4.933
	(2.030)	(6.130)	(6.320)
σ^2	0.328	0.914	0.119
	(4.489)	(20.260)	(21.530)
σ_h	-	-	0.888
	-	-	(13.471)
Log likelihood	336.692	335.893	328.043

Note: σ_w= standard deviation of time fixed symmetric effects; σ_h= standard deviation of time fixed one-sided effects; $\lambda = \sigma_{u_{it}} \ / \ \sigma_{v_{it}}$; $\sigma^2 = \sigma^2_{u_{it}} + \sigma^2_{v_{it}}$; number of observations: n=305.

Table 3: Estimated first and second order coefficients from cost frontier models (asymptotic t-ratios in parentheses).

Variable	Mean	Std. Dev.	Minimum	Maximum
RE	0.782	0.143	0.515	0.984
TRE	0.940	0.032	0.803	0.987
TGTRE	0.878	0.062	0.644	0.990
PGTRE	0.884	0.021	0.866	0.946

Table 4: Cost efficiency scores.

	RE	TRE	TGTRE	PGTRE
RE	1	0.031	-0.235	0.425
TRE	0.315	1	0.779	-0.069
TGTRE	-0.235	0.779	1	-0.653
PGTRE	0.425	-0.069	-0.653	1

Table 5: Correlation coefficients.

	Economy of output density	Economy of consumer density	Economy of scale
RE	6.089	1.470**	1.123***
TRE	4.142***	1.698***	1.127***
GTRE	3.031***	1.902***	1.119***

Significance levels: ***p<0.01, **p<0.05, *p<0.1.

Table 6: Economies of scale, output and consumer density