ECONOMIC DISCUSSION PAPERS

Efficiency Series Paper 3/2013

Using the latent class approach as a supervised method to cluster firms in DEA: An application to the US electricity transmission industry

Manuel Llorca, Luis Orea, Michael G. Pollitt

Departamento de Economía

Universidad de Oviedo

Available online at: http://economia.uniovi.es/investigacion/papers

USING THE LATENT CLASS APPROACH AS A SUPERVISED METHOD TO CLUSTER FIRMS IN DEA: AN APPLICATION TO THE US ELECTRICITY TRANSMISSION INDUSTRY

Manuel Llorca
Oviedo Efficiency Group, Department of Economics, University of Oviedo

Luis Orea
Oviedo Efficiency Group, Department of Economics, University of Oviedo

Michael G. Pollitt
Electricity Policy Research Group and Judge Business School, University of Cambridge

26 February 2013

Abstract

In this paper we advocate using the so-called latent class model (LCM) approach to control for technological differences in traditional efficiency analysis of regulated electricity networks. Our proposal relies on the fact that latent class models are designed to cluster firms by uncovering differences in technology parameters. Moreover, our approach can be viewed as a supervised method for clustering data as it takes into account the same (production or cost) relationship that is analyzed later, often using non-parametric frontier techniques. The simulation exercises confirm our expectations and show that the proposed approach outperforms other alternative sample selection procedures. The proposed methodology is illustrated with an application to a sample of US electricity transmission firms for the period 2001-2009.

Keywords: electricity transmission, utilities regulation, latent class approach, non-parametric analysis.

JEL classification: D22, L51, L94

1. Introduction

Electricity networks are often regulated by implementing incentive-based regulation schemes that use some sort of benchmarking, i.e. a comparison of utilities' performance with best-practice references. As shown by Coelli *et al*. (2005), the most commonly method used by energy regulators to measure relative firms' inefficiency is data envelopment analysis (DEA). Unlike the parametric approach that requires the specification of a particular functional form for the cost or production functions to be estimated, non-parametric methods impose fewer assumptions on the shape of firms' technology.

However, a key issue that is sometimes not taken into account by regulators (and researchers) is the existence of heterogeneity or unobserved differences among firms. Moreover, it is often assumed in this setting that the whole set of benchmarked firms share the same technology, and hence differences in behaviour are attributed to an inefficient use of factors that are under control of the companies. Possible differences among utilities associated with different technologies are either overlooked or are addressed using simple sample selection procedures, most of them based on factors that may affect performance such as geographic location or utilities' size. Therefore, the efficiency scores obtained from these analyses might be biased and some firms might be penalized (or rewarded) in excess if their underlying technology is less (more) productive than the technology used by other firms operating with more (less) advantageous conditions. This is particularly important in the case of incentive regulation and benchmarking of electricity networks where the results of efficiency analysis have important financial implications for the firms.

In this paper we advocate using a more comprehensive approach to control for technological differences in a preliminary stage, i.e. before carrying out a traditional efficiency analysis of regulated electricity networks. In particular, we propose using the so-called latent class model (LCM) approach to split the sample of utilities into a number of different classes, where each class is associated with a different technology. We argue that this approach is an appropriate statistical procedure to cluster firms in these settings for two reasons. First, they are specifically designed to cluster firms by searching for differences in production or cost parameters, which is exactly what regulators are looking for. Second, our approach can be viewed as a supervised method for clustering data as it takes into account the same (production or cost) relationship that is analyzed later, often using non-parametric frontier techniques.

The same idea is currently being developed by Agrell *et al.* (2013) in a very recent working paper where they use the LCM approach to control for technological differences in an application to Norwegian power distribution firms. Our paper reinforces the results obtained by these authors from both a theoretical and an empirical point of view. In particular, we carry out a simulation analysis to examine whether the latent class approach outperforms other alternative procedures of splitting a sample of observations - such as cluster analysis or simply using the median of some relevant variables - before the non-parametric stage. The simulation exercises confirm our expectations and show that the proposed approach outperforms other alternative sample selection procedures. On the other hand, we illustrate this procedure with an application to the US electricity transmission firms examined in Llorca *et al.* (2013). We find two statistically different groups of firms that should be compared or treated separately. In order to confirm the results from the simulation exercise, we compare the partition of the sample obtained through this method with those from alternative clustering procedures.

This paper is organized as follows. Section 2 introduces the two-stage procedure that is proposed to control for unobservable differences in firms' technology (environment) in energy regulation. Section 3 introduces the simulation analysis performed and its main outcomes. Section 4 uses data from the US electricity transmission industry to compare the relative performance of our approach and alternative procedures. Section 5 concludes.

2. A two-stage procedure to deal with unobserved heterogeneity in energy regulation

As Brophy Haney and Pollitt (2009) pointed out, regulators have been using several statistical methods to determine the performance of energy utilities. Obtaining reliable measures of firms' performance requires dealing with controllable factors and monitoring for the different environmental conditions under each firm operates. However, both regulators' reports and academic studies do not usually deal with these technological differences. Statistical methods have recently been developed to address this issue. In most of these methods, heterogeneity is understood as an unobserved determinant of the production/cost frontier, while inefficiency is interpreted as the 'distance' to the frontier once heterogeneity has been taken into account.

Following Greene (2005a, 2005b) we can distinguish two sorts of models that allow us to achieve this aim, namely the so-called true fixed/random effects models and the latent class stochastic models, also known as finite mixture models. Both approaches have their own strengths and weaknesses. In the true fixed/random effects models, unobserved heterogeneity is captured through a set of firm-specific intercepts that are to be estimated simultaneously with other parameters. Hence, using this approach implies assuming that there are as many technologies as firms.[1] However, as it imposes common slopes for all firms, all of them share the same marginal costs, economies of scale and other technological characteristics.

In contrast, the latent class model approach allows estimating different parameters for firms belonging to different groups as can be easily seen where the general specification of a cost function in this framework is expressed as follows:

$$\ln C_{it} = \alpha_j + \beta_j \ln X_{it} + v_{it|j} \qquad (1)$$

where i stands for firms, t for time and $j = 1,\ldots, J$ for class. C_{it} is a measure of firms' cost, X_{it} is a vector of explanatory variables, and the random term v_{it} follows a normal distribution with zero mean and variance σ_v^2. The number of classes J should be chosen in advance by the researcher or regulator. As both α_j and β_j, are j-specific parameters, the technological characteristics vary across classes.

Letting θ_j denote all parameters associated with class j, the conditional likelihood function of a firm i belonging to class j is $LF_{ij}(\theta_j) = \prod_{t=1}^{T} LF_{it}(\theta_j)$. The unconditional likelihood for firm i is then obtained as the weighted sum of their j-class likelihood functions, where the weights are the probabilities of class membership, P_{ij}. That is:

$$LF_i(\theta, \delta) = \sum_{j=1}^{J} LF_{ij}\,(\theta_j) P_{ij}(\delta_j), \qquad 0 \le P_{ij}(\delta_j) \le 1, \qquad \sum_{j=1}^{J} P_{ij}(\delta_j) = 1 \qquad (2)$$

[1] This idea can be considered to underlie the negotiations between regulators and utilities, where utilities wield uniqueness as a reason to avoid being compared with their peers.

where, $\theta=(\theta_1,\ldots,\theta_j)$, $\delta=(\delta_1,\ldots,\delta_j)$ and the class probabilities are parameterized as a multinomial logit model:

$$P_{ij}(\delta_j) = \frac{\exp(\delta_j' q_i)}{\sum_{j=1}^{J}\exp(\delta_j' q_i)}, \qquad j = 1,\ldots,J, \qquad \delta_J = 0 \tag{3}$$

where q_i is either an intercept or a vector of individual-specific variables. Therefore, the overall likelihood function resulting from (2) and (3) is a continuous function of the vectors of parameters θ and δ, and can be written as:

$$ln\,LF\,(\theta,\delta) = \sum_{i=1}^{N} ln\,LF_i\,(\theta,\delta) = \sum_{i=1}^{N} ln\{\sum_{j=1}^{J} LF_{ij}\,(\theta_j)P_{ij}\,(\delta_j)\} \tag{4}$$

Maximizing the above maximum likelihood gives asymptotically efficient estimates of all parameters. A necessary condition to identify the whole set of parameters is that the sample must be generated from at least two different technologies or two noise terms.

Three comments are in order. First, in this framework each firm belongs to one and only one class. Therefore, the probabilities of class membership just reflect the uncertainty that researchers or regulators have about the true partition of the sample. The estimated parameters can be used to compute posterior class membership probabilities using the following expression:

$$P(j|i) = \frac{LF_{ij}(\hat{\theta}_j)P_{ij}\,(\hat{\delta}_j)}{\sum_{j=1}^{J} LF_{ij}(\hat{\theta}_j)P_{ij}\,(\hat{\delta}_j)} \tag{5}$$

These posterior probabilities of membership can then be used to allocate each firm to a particular class, e.g., each firm is allocated to the class with the higher posterior probability.

On the other hand, only between-groups and not individual heterogeneity is controlled using a latent class model because all firms belonging to a particular group share the same technology. This situation is possible in energy economics if, as happens in our application, firms operating in areas with different environmental conditions must choose between a limited number of technical standards[2] to expand and maintain their networks. As each class has a different set of parameters, the latent class approach is able to control for the aforementioned differences in environmental conditions and technologies.

Finally, it should be noted that the random term in (1) follows a symmetric distribution and hence it does not include a traditional one-sided inefficiency term. In other words, unlike previous studies estimating latent class stochastic frontier models,[3] we advocate using a non-frontier model in a first stage as a "statistical" tool to cluster firms before carrying out a traditional efficiency analysis of regulated electricity networks (second stage). Compared to other sample-separating methods, our proposal relies on the fact that latent class models are designed to cluster firms by searching for differences in production or cost parameters, which is exactly what regulators are looking for.[4] Moreover, our approach can be viewed as a supervised method for

[2] These standards are either proposed by the International Electrotechnical Commission or the Institute of Electrical and Electronics Engineers.

[3] See, for instance, Orea and Kumbhakar (2004).

[4] Another tool that could be used in the first stage to split the sample and to reduce heterogeneity among firms is the k-means cluster analysis method. Although this procedure was proposed by Lloyd in 1957 (it was not published until 1982), this name was first used by MacQueen (1967). This method is a popular unsupervised algorithm for clustering data which is widely used in scientific research. The aim of cluster

clustering data as it examines the same (production or cost) relationship that is analyzed later in the traditional efficiency analysis. Indeed, the simulation exercises carried out in next section show that the proposed approach outperforms other alternative sample selection procedures, such as cluster analysis or the simple use of median of some relevant variables.

As mentioned above, although it is possible to estimate a stochastic cost frontier in the first stage of our procedure, we propose obtaining the efficiency scores later. There are three reasons for this. First, ignoring the asymmetric error term traditionally associated with inefficiency prevents the appearance of convergence problems in practice when estimating a latent class model, which by nature is highly non-linear. This facilitates replication of the procedure when researchers or regulators compare different specifications of the underlying technology. Second, this empirical strategy allows us to compute efficiency scores using more flexible representation of firms' technologies if non-parametric techniques such as DEA are employed. Finally, DEA is the method mainly used by regulators.

In a second stage we apply DEA. As noted by Zhou *et al.* (2008), DEA has become a very popular tool in energy and environmental studies, especially for benchmarking electric utilities. It is a type of efficiency analysis which involves mathematical programming to construct a frontier of best performing companies. Farrell (1957) was the first to propose this type of frontier analysis and since then there have been many authors who have developed and applied different models which have enlarged the literature in DEA methodology (see Coelli *et al*., 2005).

In this paper, as we will assume that the output level cannot be modified by firms we will use an input-oriented DEA model. This assumes that technical inefficiency can be viewed as a proportional reduction in input usage or cost while maintaining the output levels constant. In our simulation exercise we impose constant returns to scale (CRS), so any efficient firm should be operating at an optimal scale level. The optimization problem in this case can be represented as:

$$\begin{aligned} & min_{\theta,\lambda}\theta, \\ st \quad & -q_i + Q\lambda \geq 0, \\ & \theta x_i - X\lambda \geq 0, \\ & \lambda \geq 0 \end{aligned} \tag{6}$$

where λ is a vector of constants and θ is a scalar calculated for each observation which represents the efficiency score for the *i*-th firm. q_i and x_i are the vectors of inputs and outputs for the *i*-th firm respectively, while Q and X are the input and output matrices for all I firms. This linear programming problem must be solved I times and gives an efficiency score θ equal or lower than one for each firm. It should be noted in addition that in our empirical application we do not assume that all the companies exhibit

---

analysis is to divide the observations into homogeneous and distinct groups by taking advantage of the information contained in variables or attributes of interest. It involves minimizing the following objective function:

$$J = \sum_{j=1}^{k}\sum_{i=1}^{n}\left\|x_i^{(j)} - c_j\right\|^2$$

where k is the chosen number of clusters, n the number of data points, and $\left\|x_i^{(j)} - c_j\right\|^2$ is the distance between a certain data point $x_i^{(j)}$ and the cluster centre, c_j.

constant returns to scale as we use a variable returns to scale (VRS) specification,[5] which only requires adding the convexity constraint $I1'\lambda = 1$ to the minimization problem in (6). $I1$ is a vector of ones, and multiplying by the vector of weights λ basically ensures that firms are only compared with firms of a similar size.

3. Simulation analysis

In this section we carry out a simulation exercise to examine whether a latent class approach is a good procedure to find groups of comparable companies within a sample when we want to apply a benchmarking with DEA, commonly used in regulatory processes.

The simulation exercise can be summarized as follows. We initially generate 1,000 observations of two hypothetical outputs (Y_1, Y_2) using an uniform distribution between 0 and 1. We have chosen this distribution instead of the normal distribution because these variables cannot take negative values, and outputs in DEA must be positive. Inefficiency levels are obtained assuming that the inefficiency term, represented as u^+, is a positive half-normal distribution with zero mean and σ_u^2 variance. Random noise is simulated assuming that the noise term v follows a normal distribution with zero mean and σ_v^2 variance. We impose $\sigma = \sqrt{\sigma_u^2 + \sigma_v^2}$ equal to 1 which, given the specification that we have chosen (see below), implies that the size of the random term in our function is relatively low, i.e. our levels of generated efficiency are quite high. We also fixed $\gamma = \sigma_u^2/(\sigma_u^2 + \sigma_v^2)$ equal to 0.5, which implies that the weights of inefficiency and noise in the function are the same. Given the previous values, this implies that $\sigma_u = \sigma_v = 0.71$, and therefore is equivalent to generating a value of $\lambda = \sigma_u/\sigma_v$ equal to 1.[6]

Firms' costs are simulated following the normalized linear specification proposed by Bogetoft and Otto (2011) for the regulation of electrical Distribution System Operators in Germany. This functional form allows us to easily introduce heteroscedasticity in our data generation process. Following this type of specification, our cost function can be expressed as follows:

$$\frac{C_i}{Y_{1i}} = \beta_1 + \beta_2 \frac{Y_{2i}}{Y_{1i}} + u_i^+ + v_i \qquad (7)$$

where β_1 and β_2 stands for the marginal costs of the outputs Y_1 and Y_2 and define our technologies. With this functional form, we are imposing constant returns to scale. This prevents size effects when comparing our sample separating methods. As the random noise term takes both positive and negative values, we impose on all technologies that (β_1 + β_2) =10 to get positive costs. Technologies thus differ in the relative weight of each β, i.e. in relative marginal costs. In particular, have simulated three possible technologies:

- Technology A: $\beta_2 = \beta_1 \quad , \quad (\beta_1 = 5,\ \beta_2 = 5)$
- Technology B: $\beta_2 = 2\beta_1 \quad , \quad \left(\beta_1 = \frac{10}{3},\ \beta_2 = \frac{20}{3}\right)$

[5] Although is quite common to presume that electricity transmission firms are natural monopolies, this is confirmed by the increasing returns to scale obtained from different authors: Huettner and Landon (1978), Pollitt (1995), Dismukes *et al.* (1998) and Llorca *et al.* (2013).

[6] Although the values of these parameters have been arbitrary chosen, the results obtained from the simulation are consistent respect to changes in them as long as we keep the underlying efficiency at 'normal' levels.

- Technology C: $\beta_2 = 4\beta_1 \quad , \quad (\beta_1 = 2,\ \beta_2 = 8)$

Both coefficients are the same in technology A, while marginal costs are increasingly different in the other two technologies, B and C.

Next, we will examine the robustness of our results by adding differences between outputs. In particular, we modify the original statistical distribution of one of the outputs by doubling and quadrupling its range of values, that is:

- Distribution 1: $Y_1 \sim U(0,1) \quad , \quad Y_2 \sim U(0,1)$
- Distribution 2: $Y_1 \sim U(0,1) \quad , \quad Y_2 \sim 2 \cdot U(0,1)$
- Distribution 3: $Y_1 \sim U(0,1) \quad , \quad Y_2 \sim 4 \cdot U(0,1)$

Taking into account that we always apply the technology A to the first 500 observations and then B or C to the following 500 observations, and that we have three output distributions, 6 possible scenarios are obtained. In Table 1, we show the scenarios and the percentage of success predicting the underlying class membership. This percentage is computed by comparing the ratios of β that can be recovered after applying an OLS regression to different groups of observations with the real ones. The estimated ratios are also shown in Table 1. The estimated values give an idea about how well each procedure is able to identify the different underlying technologies.

[Insert Table 1]

The first empirical exercise has to do with the case in which DEA is applied using the real separation of our data (i.e. the first 500 observations belonging to technology A and the later 500 observations belonging to B or C). By construction the percentage of success in this case is 100%. For this reason, this exercise is used as a benchmark to study the performance of four sample separation methods: the median of the cost[7], cluster analysis considering the outputs, cluster analysis including both outputs and cost, and the latent class model (that involves both output and cost information).

Looking at the percentages of success and the β-ratios we can confirm that the LCM is the method that is most precise in assigning observations to technologies and is also the best at identifying the relationship between technologies. As we move to a different scenario where there are more unequal features among groups we observe that there is a clear divergence in the behaviour of the procedures: whereas the LCM improves its percentage of success in the prediction[8], the alternative procedures only slightly improve their performances.

In Table 2, we show the average efficiencies that are obtained after DEA is applied separately to each group of firms. The last column shows the sums of squared differences with respect to the real separation case. Leaving aside the 'true' partition of the sample, i.e. the real separation case, the LCM is the approach that gives the largest efficiency scores. Moreover, the sums of squared differences in the last column indicate that the LCM is not only the procedure that gives us a higher average efficiency level (and closest to the real separation case) but also is the best at predicting individual efficiencies.

[Insert Table 2]

[7] The sample separation using the median of a variable can be viewed as a cluster analysis in which a dummy variable that takes value 1 for values above the median, and 0 otherwise, is considered as a classification variable.

[8] The estimated probabilities for the most likely latent class also increase, so the LCM not only improves its prediction capacity but also the precision with which each observation is assigned.

Once unobserved heterogeneity is taken into account and 'removed' in a first stage, larger efficiency scores are obtained when carrying out a traditional DEA analysis. It should be noted that as inequalities between groups rise, the average efficiency score obtained using the LCM as a sample separation method even exceed the average efficiency score from the real separation case. This shows that an imperfect assignment of firms to groups can lead us to obtain higher levels of efficiency or, in other words, the closest frontier to a firm is not always its real reference frontier. This result is quite interesting from a firm's perspective as in some cases this procedure can be more favourable for them than a 'true' benchmarking. In spite of this, it seems that the efficiency level is a good indicator of how well each procedure performs and assigns firms to groups.

In Figure 1 we show the positive correlation that exists between efficiency and success in assigning observations to technologies using the LCM approach. This figure allows us to examine the discriminatory power of the model when there are either larger differences between technologies (illustrated as the shift from the blue to the red line) or between output data generation processes (illustrated as movements along the red and blue lines). As expected, the percentages of success are much larger when the two technologies differ notably in their characteristics. It is worth mentioning that this increase in percentages of success is especially important when there is no separating information on the output side, i.e. when both outputs are similarly generated. When additional information for splitting the sample is contained in the way both outputs are generated, both efficiency levels and percentages of success increase regardless of whether the technologies are similar or diverse.

[Insert Figure 1]

In summary, the above results clearly indicate that the LCM deals with unobserved heterogeneity much better than the others. This conclusion is one of the main contributions of the paper because it provides evidence in favour of using the LCM as a simple statistical sample separation method in energy regulation.[9] We attribute this better performance to the fact that the LCM, unlike other methods, splits the data taking into account the objective of the second stage, where a relationship between outputs and inputs (or costs) is estimated in order to compute inefficiency scores. Alternative sample separating methods only try to find statistical differences in the mean values of a set of variables. In this sense, and borrowing the terminology used for dimension reduction, this approach can be interpreted as a 'supervised' method to split the data.

4. Application to the US electricity transmission industry

We next illustrate the proposed procedure with an application to the US electricity transmission industry. The database used in this paper is the same as in Llorca *et al.* (2013) and contains 405 observations on 59 US electricity transmission firms for the period 2001-2009.

Following the literature,[10] we specify a standard cost function with four outputs where our cost variable is Totex (which includes operation and maintenance expenses,

[9] In this sense, our paper can be used to justify the approach first suggested by Agrell *et al.* (2013).

[10] As is highlighted by Brophy Haney and Pollitt (2012), benchmarking of electricity transmission utilities is a challenging task due to the small number of transmission utilities that usually operate in the jurisdiction of a particular regulator. This likely explains why there are few empirical papers published on

annual depreciation on capital assets, and annual return on the balance of capital). The four outputs are: *Peak Load* (PL), which is the maximum peak load of the year during 60 minutes; *Electricity Delivered* (DE), which is the total annual energy delivered by the system; *Total Energy*[11] (TE), which stands for the total energy of the system, including total net own generation, total purchases from others, net exchanges in the system (received-delivered), net transmission for others and transmission by others; and *Network length* (NL), which is a measure of the geographic spread of each company and is obtained as the sum of all transmission lines in miles regardless of the number of power cables on each power line. The four outputs considered (explanatory variables) and the cost variable (dependent variable) will be used later on in the DEA stage.

To analyse robustness, we extend the standard model by adding four environmental variables to split the sample of transmission utilities. Three of these are weather variables: *Temperature* (TMIN), which represents the annual minimum temperature in Fahrenheit degrees; *Wind* speed (WIND), which is the average of the daily mean wind speeds in knots; and *Precipitation* (PRCP), which is the average of daily precipitation in inches. The last environmental variable is the *Growth in Demand* (GDEM) for each firm over time. The descriptive statistics of the full set of variables are shown in Table 3.

[Insert Table 3]

The specification of the cost function used in the sample-separating stage of our procedure is quite simple in order to avoid convergence problems and facilitate the replication of the procedure. Unlike our simulation, we have preferred to use a Cobb-Douglas (or logarithm) specification of the cost function due to its widespread use and acceptance in previous empirical studies. Convergence problems prevented us from estimating the LCM for more than two classes. However, these problems did not appear using the Cobb-Douglas functional form. As we do not know the true number of underlying technologies, this is an interesting advantage of the logarithm specification of the model. The coefficients for the Cobb-Douglas specification are shown in Table 4. Except for total energy (TE) for one of the classes, all estimated coefficients are statistically significant and positive.[12]

[Insert Table 4]

In Figure 2 we illustrate the individual efficiency scores obtained after applying DEA as the number of classes is increased. As expected, the average efficiency score for the so-called non-separation model[13] is 65%, much lower than the average efficiency obtained from the model with two classes (77%). The most comprehensive model that is estimated is a LCM with 7 classes. Although the average efficiency score for this model

efficiency analysis of electricity transmission firms. Exceptions are Huettner and Landon (1978), Pollitt (1995), Dismukes *et al.* (1998) and von Geymueller (2009). However, none of these articles deal with unobserved heterogeneity or technological differences.

[11] Although *Electricity Delivered* and *Total Energy* are both variables that measure electricity flows, it can be observed in Table 3 that they are quite different since *Total Energy* includes transmission for others. We have decided to include both output variables as they help to increase the efficiency scores obtained with DEA in the second stage.

[12] As shown in the Appendix, for two groups the linear specification gives reasonable parameter estimates. It also gives similar group membership probabilities and efficiency scores. For instance, the percentage of coincidence in the assignment of observations is 87.7%, and the average efficiency score is higher using the Cobb-Douglas (77.03%) than using the linear form (72.45%).

[13] Note that in the non-separation model the sample of firms is not divided into several groups and hence can be viewed as a model with one class.

goes up to 87%, the largest change in efficiencies occurs when we move from one class to two classes. The values of both the AIC and BIC criteria for model selection are shown in Figure 3. While the AIC always decreases when we move to a larger number of classes, the BIC statistic has its minimum value for two classes. Taking the BIC statistic into account and given that the main improvement in efficiency levels is observed when we move from one class to two, we chose the model with two classes as our preferred model.[14]

[Insert Figure 2]

[Insert Figure 3]

We show in Figure 4 the efficiency scores obtained using different methods to split the sample into two groups of firms. As expected, the lowest efficiency levels are obtained when there is no separation of firms and when we use cluster procedures where we include the size of the network and the cost as separating variables.[15] A simple division using the median of cost seems to produce larger scores of efficiency in the second stage. As in the simulation exercise, the largest efficiency scores are obtained when the LCM is used as a statistical tool to account for unobserved differences among firms.

[Insert Figure 4]

We next introduce some environmental variables (i.e., three weather variables and demand growth) as sample-separating variables in the first stage of procedure to analyse the robustness of our results. Table 5 shows the estimated coefficients for the extended LCM. The coefficients of the variables included in the cost function do not undergo major changes with the exception of electricity delivered (DE), which is no longer statistically significant in one of the classes. Regarding the sample-separating variables, temperature, wind and growth of the demand are statistically significant, which implies that they have helped the procedure to split the sample in two groups. Despite this, there are not many differences between our previous LCM that ignored this information and the extended LCM that includes separating variables. For instance, the percentage of coincidence in allocating observations of both specifications is quite high (88%). In addition, Figure 5 shows that both LCMs give us larger efficiency scores than extended k-means procedures that include environmental variables (alone or with information about the cost function). As with the simulation exercise, these results suggest that the latent class approach is the best procedure for finding 'homogeneous' groups of firms when we do not have information about the environment in which these firms operate. When this information is available, the LCM still outperforms other sample separating methods.

[Insert Table 5]

[Insert Figure 5]

[14] Both criteria are based on the maximum value of the logarithm of the likelihood function and the number of parameters estimated. However, as the BIC penalizes more adding successive parameters, it is our favoured criterion.

[15] The separation when we take into account all the outputs and the cost or just the network and the cost is the same.

5. Conclusions

In energy regulation, differences in technologies or unobserved heterogeneity between firms are often not taken into account despite the theoretical importance of environmental features on utilities performance. As in Agrell *et al.* (2013), in this paper we propose using a latent class approach as a statistical method to split the sample into groups of more comparable firms before carrying out a traditional efficiency analysis using DEA, the most common frontier technique used by regulators in utility benchmarking.

We have demonstrated through a simulation exercise that the latent class approach allocates each observation to its reference group better than the alternative procedures and that the efficiency scores obtained in the second stage are larger. It has also been shown that when large differences between technologies or output distributions arise, the discriminatory capacity and the assignment success of the procedure increases and the second-stage efficiency levels converge to the true underlying levels. An additional outcome of our simulation exercise is the large correlation between average efficiency levels and the percentage of success allocating observations into classes. This outcome is very important because it suggests that the average efficiency level obtained in a second stage can be used in practice as a good proxy of the relative performance of any sample-separating method that has been carried out before the traditional efficiency analysis.

Finally, we illustrate the proposed method with an application to a sample of US electricity transmission firms for the period 2001-2009. We find that the largest change in efficiency scores occurs when we move from a model without any partition of the sample to a LCM that only splits the sample into two classes. Like the simulation exercise, our empirical application suggests using a latent class approach as a statistical method to deal with unobserved heterogeneity and differences in environmental characteristics.

References

Agrell, P.J., Farsi, M., Filippini, M. and Koller, M. (2013), *Unobserved heterogeneous effects in the cost efficiency analysis of electricity distribution systems*, CER-ETH Economics working papers series 13/171, CER-ETH - Center of Economic Research (CER-ETH) at ETH Zurich.

Bogetoft, P. and Otto, L. (2011), *Benchmarking with DEA, SFA, and R*, Springer, New York.

Brophy Haney, A. and Pollitt, M. (2009), "Efficiency analysis of energy networks: an international survey of regulators", *Energy Policy*, 37: 5814-5830.

Brophy Haney, A. and Pollitt, M. (2012), *International benchmarking of electricity transmission by regulators: Theory and practice*, Electricity Policy Research Group Working Paper 1226, University of Cambridge.

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

Dismukes, D.E., Cope III, R.F. and Mesyanzhinov, D. (1998), "Capacity and economies of scale in electric power transmission", *Utilities Policy*, 7, 3, 155-162.

Farrell, M.J. (1957), "The measurement of productive efficiency", *Journal of the Royal Statistical Society*. A, 120, 253-281.

Greene, W. (2005a), "Fixed and random effects in stochastic frontier models", *Journal of Productivity Analysis*, 23, 7-32.

Greene, W. (2005b), "Reconsidering heterogeneity in panel data estimator of the stochastic frontier model", *Journal of Econometrics*, 126, 269-303.

Huettner, D.A., and Landon, J.H. (1978), "Electric utilities: scale economies and diseconomies" *Southern Economic Journal*, 44, 4, 883-912.

Llorca, M., Orea, L. and Pollitt, M. (2013), *Efficiency and environmental factors in the US electricity transmission industry*, Efficiency Series Papers 02/2013, Oviedo Efficiency Group.

Lloyd, S.P. (1982), "Least squares quantization in PCM", *IEEE Transactions on Information Theory*, vol. it-28, 129-137.

MacQueen, J.B. (1967), *Some methods for classification and analysis of multivariate observations*, Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. University of California Press, 281–297.

Orea, L. and Kumbhakar, S. (2004), "Efficiency measurement using stochastic frontier latent class model", *Empirical Economics*, 29, 169-183.

Pollitt, M. (1995), *Ownership and performance in electric utilities*, Oxford University Press, Oxford.

Von Geymueller, P. (2009), "Static versus dynamic DEA in electricity regulation: the case of US transmission system operators", *Central European Journal of Operations Research*, 17: 397-413.

Zhou, P., Ang, B.W. and Poh, K.L., (2008), "A survey of Data Envelopment Analysis in energy and environmental studies", *European Journal of Operational Research*, 189, 1-18.

Table 1. Success of the procedures identifying technologies

Simulation	Procedure	% Success	Underlying technology	
			Group 1 (β1 / β2)	Group 2 (β1 / β2)
A&B	Simulation	-	1.000	0.500
(OD 1)	Real separation	100.00	1.080	0.597
	Median (C)	49.60	0.890	0.756
	Cluster (Y_1, Y_2)	46.50	0.849	0.815
	Cluster (Y_1, Y_2, C)	49.30	0.894	0.763
	LCM	65.70	1.063	0.555
A&C	Simulation	-	1.000	0.250
(OD 1)	Real separation	100.00	1.080	0.331
	Median (C)	50.20	0.678	0.575
	Cluster (Y_1, Y_2)	46.50	0.646	0.642
	Cluster (Y_1, Y_2, C)	49.90	0.684	0.593
	LCM	79.20	1.162	0.337
A&B	Simulation	-	1.000	0.500
(OD 2)	Real separation	100.00	1.077	0.597
	Median (C)	55.00	0.834	0.799
	Cluster (Y_1, Y_2)	54.00	0.822	0.812
	Cluster (Y_1, Y_2, C)	55.10	0.822	0.802
	LCM	79.30	1.110	0.596
A&C	Simulation	-	1.000	0.250
(OD 2)	Real separation	100.00	1.077	0.331
	Median (C)	57.20	0.723	0.562
	Cluster (Y_1, Y_2)	54.00	0.656	0.609
	Cluster (Y_1, Y_2, C)	58.30	0.714	0.529
	LCM	87.90	1.099	0.337
A&B	Simulation	-	1.000	0.500
(OD 3)	Real separation	100.00	1.076	0.598
	Median (C)	57.40	0.868	0.765
	Cluster (Y_1, Y_2)	53.90	0.833	0.785
	Cluster (Y_1, Y_2, C)	57.80	0.863	0.754
	LCM	90.60	1.097	0.583
A&C	Simulation	-	1.000	0.250
(OD 3)	Real separation	100.00	1.076	0.331
	Median (C)	60.60	0.779	0.493
	Cluster (Y_1, Y_2)	53.90	0.674	0.576
	Cluster (Y_1, Y_2, C)	61.80	0.772	0.486
	LCM	94.70	1.102	0.328

Table 2. Efficiencies with DEA

Simulation	Procedure	Av. Eff.	SSD
A&B	Real separation	76.73	-
(OD 1)	No separation	67.15	135,252
	Median (C)	73.29	118,885
	Cluster (Y_1, Y_2)	70.38	108,590
	Cluster (Y_1, Y_2, C)	72.91	118,993
	LCM	73.35	40,268
A&C	Real separation	75.16	-
(OD 1)	No separation	54.70	533,029
	Median (C)	65.26	322,549
	Cluster (Y_1, Y_2)	61.65	379,683
	Cluster (Y_1, Y_2, C)	64.65	338,483
	LCM	78.93	139,993
A&B	Real separation	83.61	-
(OD 2)	No separation	73.28	151,038
	Median (C)	76.75	121,264
	Cluster (Y_1, Y_2)	75.93	119,387
	Cluster (Y_1, Y_2, C)	76.22	124,400
	LCM	85.75	51,232
A&C	Real separation	83.14	-
(OD 2)	No separation	63.22	507,703
	Median (C)	69.29	372,090
	Cluster (Y_1, Y_2)	68.14	386,773
	Cluster (Y_1, Y_2, C)	68.32	389,718
	LCM	85.87	45,663
A&B	Real separation	89.26	-
(OD 3)	No separation	78.75	180,309
	Median (C)	80.36	158,268
	Cluster (Y_1, Y_2)	79.99	160,646
	Cluster (Y_1, Y_2, C)	80.20	159,799
	LCM	90.76	29,598
A&C	Real separation	89.24	-
(OD 3)	No separation	70.49	511,481
	Median (C)	73.45	429,212
	Cluster (Y_1, Y_2)	72.86	446,596
	Cluster (Y_1, Y_2, C)	72.95	438,210
	LCM	90.15	16,511

Table 3. Descriptive statistics

	Variable	Units	Mean	Max.	Min.	Std.Dev.
Totex	Cost	US$	144,602,000	667,127,000	20,713,600	120,324,000
Peak Load	Output	MW	6,173	23,111	380	5,533
Electricity Delivered	Output	MWh	6,280,310	74,584,700	56,730	8,839,980
Total Energy	Output	MWh	34,557,900	116,415,000	2,339,000	26,752,600
Network Length	Output	Miles	4,064	16,292	1,087	3,253
Minimum Temperature	Weather	°F	-10.35	19.90	-59.80	16.51
Wind Speed	Weather	Knots	6.84	9.60	4.63	1.01
Precipitation	Weather	Inches	0.07	0.16	0.01	0.03
Growth in Demand	Other	%	0.03	244.11	-74.96	17.72

Table 4. Parameter estimates for the Cobb-Douglas specification

	LCM – CD			
	CLASS 1		**CLASS 2**	
Variable	**Coefficient**	**t-ratio**	**Coefficient**	**t-ratio**
Constant	14.257	7.852	8.211	16.037
ln PL_{it}	0.808	4.853	0.144	3.109
ln DE_{it}	0.044	1.900	0.054	5.258
ln TE_{it}	-0.261	-1.357	0.415	7.817
ln NL_{it}	0.184	2.038	0.136	6.192
Sigma	0.380	22.982	0.119	11.332
Log LF	-39.666			

Prior class probabilities	0.444	0.556

Table 5.

Parameter estimates for the Cobb-Douglas specification with separating variables

	LCM – CD (Weather, Demand)			
	CLASS 1		CLASS 2	
Variable	**Coefficient**	**t-ratio**	**Coefficient**	**t-ratio**
Constant	13.957	8.070	8.286	17.580
ln PL_{it}	0.800	4.907	0.166	3.881
ln DE_{it}	0.042	1.457	0.060	5.823
ln TE_{it}	-0.237	-1.300	0.401	7.785
ln NL_{it}	0.182	2.085	0.123	5.133
Sigma	0.381	22.078	0.111	11.382
Log LF	-26.726			
Prior class probabilities	0.479		0.521	

Estimated prior prob. for class membership		
Variable	**Coefficient**	**t-ratio**
Constant	-0.088	-0.416
$TMIN_i$	-0.065	-3.001
$WIND_i$	-0.373	-2.153
$PRCP_i$	11.910	1.535
$GDEM_i$	0.092	1.744

Figure 1. Average efficiency and percentage of success for the LCM

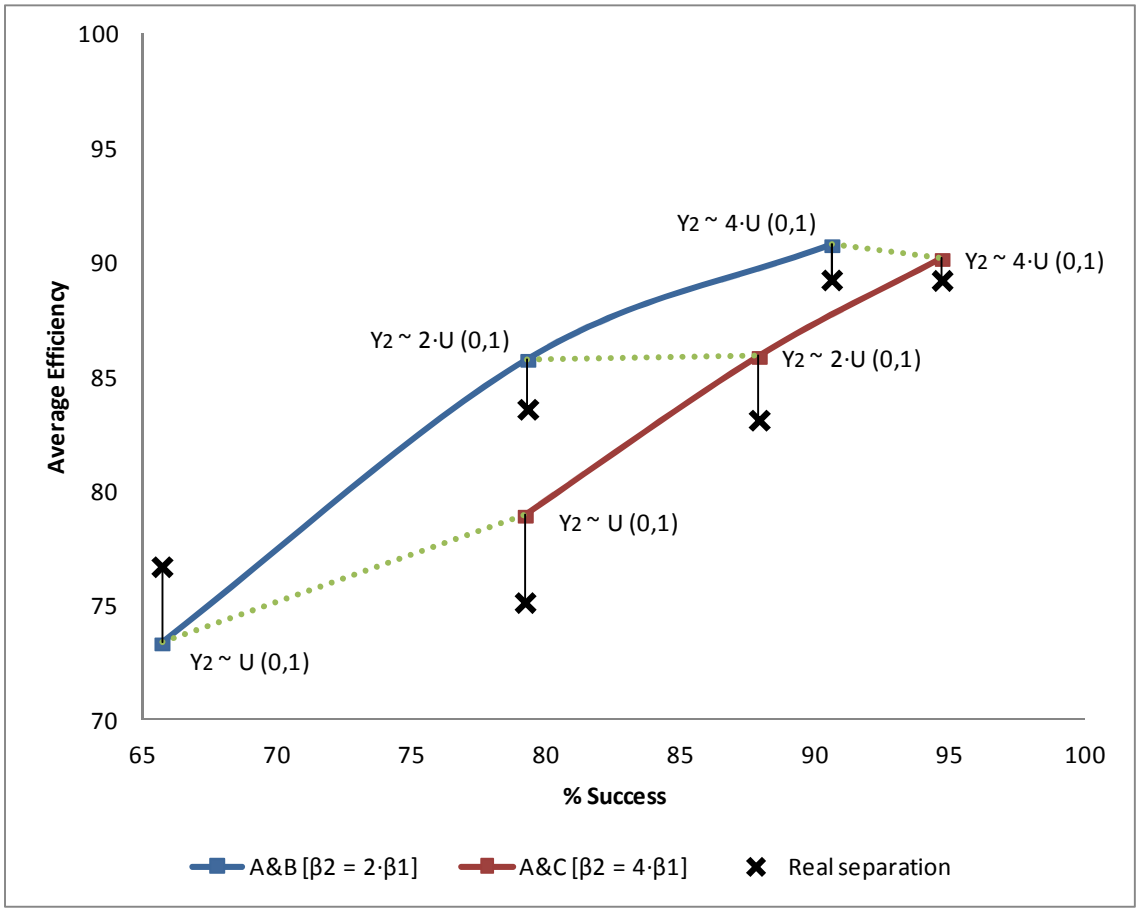


Figure 2. Efficiency scores obtained with LCM

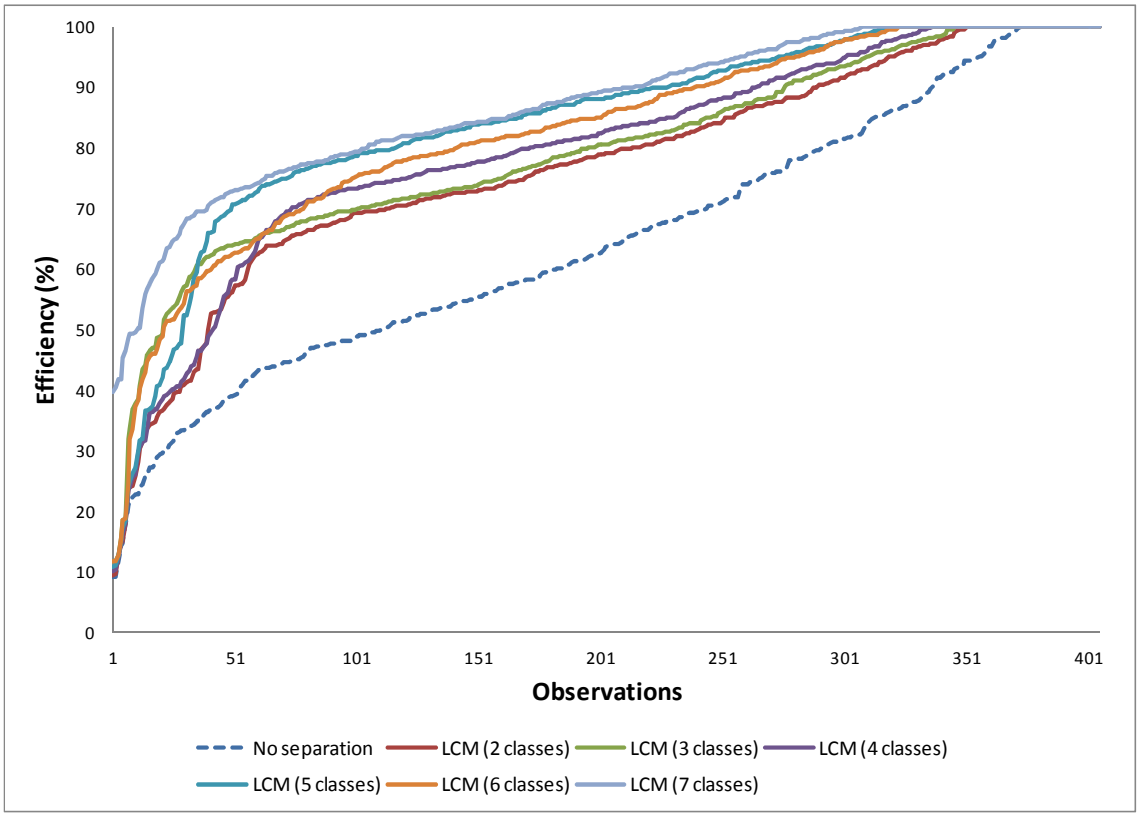


Figure 3. AIC and BIC for the different LCM

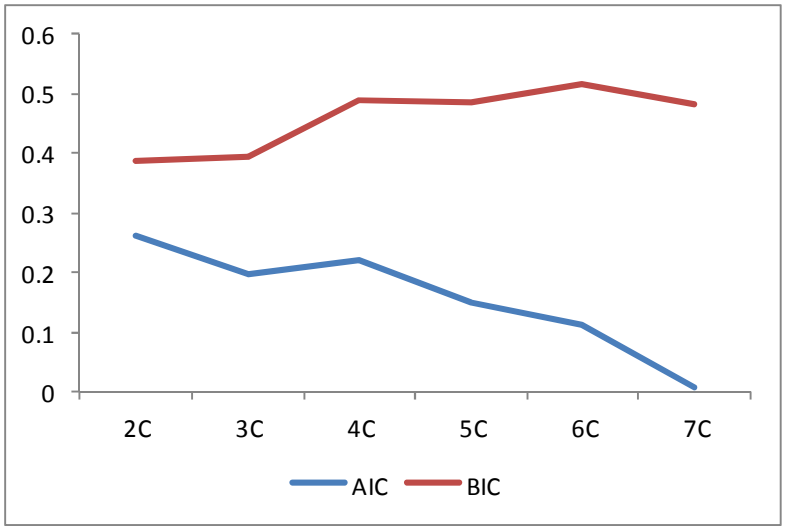


Figure 4. Efficiency scores obtained with different procedures

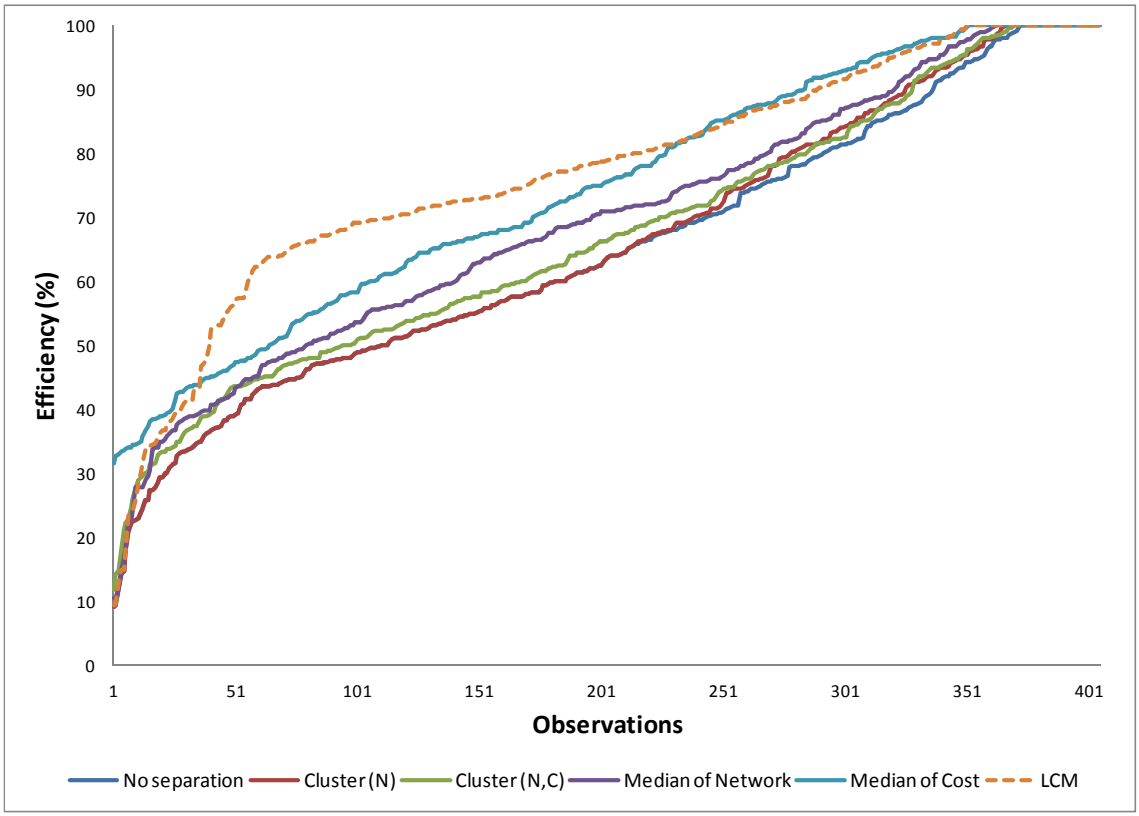
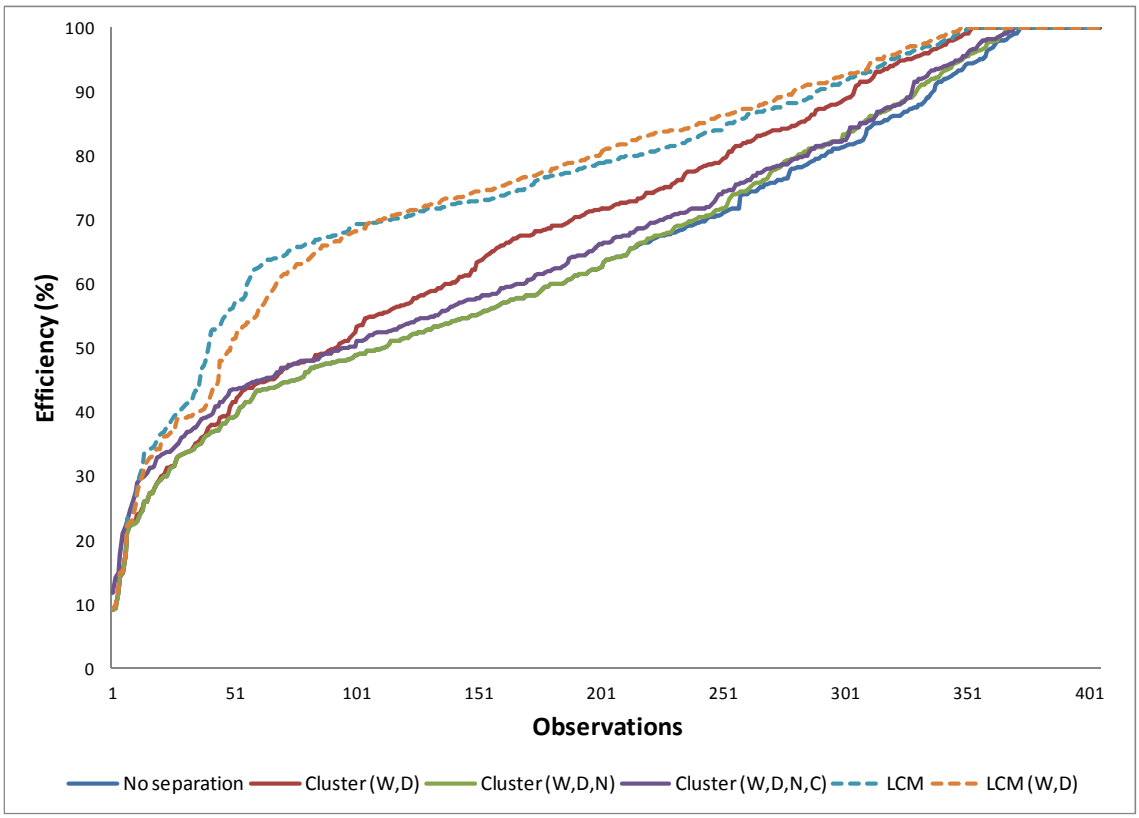


Figure 5. Efficiency scores including environmental variables



APPENDIX

Parameter estimates for the linear specification of the LCM

	LCM - Linear			
	CLASS 1		CLASS 2	
Variable	**Coefficient**	**t-ratio**	**Coefficient**	**t-ratio**
Constant	16,326,900	1.044	23,246,900	10.070
PL_{it}	22,756.088	12.640	3,442.459	4.990
DE_{it}	4.194	7.653	0.546	5.318
TE_{it}	-0.698	-1.287	1.521	10.312
NL_{it}	6,211.166	3.343	4,234.141	6.327
Sigma	57,039,600	15.028	16,185,400	22.864
Log LF	-7,580.103			
Prior class probabilities	0.444		0.556	