Using Supervised Environmental Composites in Production and Efficiency Analyses: An Application to Norwegian Electricity Networks

Luis Orea

Department of Economics, University of Oviedo, Oviedo, Spain

Christian Growitsch

Hamburg Institute of International Economics, Hamburg, Germany

*Tooraj Jamasb **

Durham University Business School, Durham, UK

*Corresponding author: Durham University Business School, Mill Hill Lane, Durham DH1 3LB, United Kingdom, Email: tooraj.jamasb@durham.ac.uk

Using Supervised Environmental Composites in Production and Efficiency Analyses: An Application to Norwegian Electricity Networks

Abstract

Although supervised dimension reduction methods have been extensively applied in different scientific fields, they have hardly ever been used in production economics. Nonetheless, these methods can also be useful in regulation of natural monopolies, where firms' cost and performance are affected by a large number of environmental factors. As economic theory suggests the presence of other relevant production or cost drivers, the traditional *all-inclusive* assumption is not satisfied. This paper shows that purging the data allows us to address this issue when analyzing the effect of weather and geography on efficiency in the context of the Norwegian electricity distribution networks.

Keywords: supervised composites, environmental conditions, electricity networks.

JEL: D24, L51, L94

1. Introduction

New technologies allow researchers to collect and analyze large amounts of data at relatively low cost. Examples of large-size data sets are plentiful, among others, in computational biology, climatology, geology, neurology, health science, economics, and finance. The availability of more data along with new scientific problems has reshaped statistical thinking and data analysis. Reducing the dimensions of data is a natural and sometimes necessary in policy analysis. The act of replacing a set of regressors with a lower-dimensional function is called *dimension reduction*; and the reduction is labeled as *sufficient* or *supervised* when this reduction is achieved without loss of information.[1]

In a seminal paper, Li (1991) introduced the first method for sufficient dimension reduction, i.e. sliced inverse regression (SIR), and since then various types of inverse regressions have been proposed.[2] These inverse regression methods are computationally simple and practically useful, and have been widely used in diverse fields such as biology, genome sequence modeling, pattern recognition involving images (e.g. face recognition, character recognition, etc.), or speech (e.g. auditory models). However, the potential of sufficient dimension reduction methods for reducing economic data has barely been explored. A notable exception is Naik et al. (2000) that use the SIR techniques to aggregate marketing data.

[1] Fisher (1922) formulated the concept of sufficient statistic as a means of reduction without loss of information. See Cook (2007) and Adragni and Cook (2009) for a formal definition and overviews of *sufficient dimension reduction* in regression.

[2] The literature on sufficient dimension reduction is still evolving. See Cook (1998) for an early summary of this literature. For a brief survey, see Bura and Yang (2011) and references therein.

Electricity, Gas, and water distribution networks are natural monopolies and, as a result, are, in many countries, subject to economic regulation by sector regulators. Since the 1990s, many regulators have adopted incentive-based regulation regimes that use efficiency and productivity analysis techniques for benchmarking performance of network utilities and determining their allowed revenues and prices (see, e.g., Jamasb and Pollitt, 2001; 2007). The results of efficiency analysis have important financial implications for the firms.

In this paper we implement the supervised dimension reduction techniques in the context of economic regulation and efficiency analysis of distribution networks in a study of the electricity sector in Norway. We use cost function modeling focusing on the effect of weather and geography on cost efficiency of the networks. Weather and geographic conditions can affect the performance of networks. Severe weather conditions and difficult terrains can increase investment and operating costs as well as service interruptions and costs of replacing damaged equipment and restoring power. The sector regulator in Norway has access to more than 60 different weather and geographic condition variables that can potentially affect the performance of the networks. However, for practical reasons only a few of these variables can be included in parametric and non-parametric efficiency analysis models.

While controlling for the potential effect of environmental factors on the performance and regulation of utilities is important, the empirical evidence is limited, with Yu et al. (2009a) and Growitsch et al. (2012) as notable exceptions. Therefore, a data dimension reduction strategy that maximizes the use of available information and minimizes the loss of information on the heterogeneity of the operating environment of the networks

contributes to the accuracy and fairness of the regulation process and outcomes. The proposed method can be easily incorporated in regulatory efficiency analysis.

Taking into account the effect of weather and geography on the cost of production and quality of service is a challenging task. It is difficult to formulate hypotheses or impose restrictions derived from production theory on the technology and weather or other non-discretionary influences from a large number of factors with complex interactions. This leads to a 'dimensionality' issue which becomes acute in non-parametric settings or when flexible parametric functional forms are to be estimated (e.g., Translog).

The most common strategy to work with high dimensional data in production and efficiency analysis is dimension reduction (e.g., principal components or explanatory factor analyses), where the input dimension space is reduced into a small number of composites for further analysis using a linear combination of the original variables.[3] This approach is used in the abovementioned studies and in Nieswand et al. (2009), Adler and Ekaterina (2010) and Zhu (1998). The main drawback of principal component analysis (PCA) and similar *unsupervised* dimension reduction techniques is that they might mis-specify the fundamental relationship between a dependent variable (for instance cost) and weather and geography because they ignore information on the dependent variable when reducing the dimension of the data. Therefore, predictions

[3] An alternative approach to avoid the curse of dimensionality is variable selection and regression shrinkage (see Fan and Li, 2006 for overviews). While *all* candidate explanatory variables indirectly have (i.e. through the composite) an effect on the dependent variable in variable dimension reduction, some of them are dropped from the model in lasso (Tibshirani, 1996) and other variable selection models. Although the latter approach might help in obtaining more interpretable models, such a procedure, however, might result in leaving out some relevant variables that are highly correlated with one or several selected variables.

might be biased because relevant predictive variables can be underweighted, while factors irrelevant for cost can be overweighed.

In contrast, the *supervised* methods for reducing the dimensions of data assume that the dependent variable is known, and take into account the relationship between the variable to be predicted (e.g., cost), and the vector of explanatory variables that are to be aggregated (e.g., weather and geographic variables). A common characteristic of these methods is that they aim to replace a high-dimensional vector of explanatory variables with a lower-dimensional function, provided that it captures all the information about the dependent variable. Another feature of these techniques is that they require no pre-specified model for the production or cost function and, except for the traditional exogeneity assumption, no other assumptions are made about the error term. This characteristic is especially important in efficiency analysis where the error term is composed into a (symmetric) noise term and a (asymmetric) random efficiency term.

In this study we use non-nested model selection tests to examine the relative performance of PCA and two of the most common supervised methods, viz. the sliced inverse regression of Li (1991) and its parametric counterpart PIR (for parametric inverse regression). To our knowledge, the present paper is the first attempt to apply supervised methods to production economics. As controlling for the effect of environmental factors is important in efficiency analyses and regulation of electricity networks, we also examine whether efficiency analyses are robust to using unsupervised vs. supervised aggregation methods. This is achieved by using a model of firms' inefficiency. While previous papers have examined the robustness of efficiency results

with respect to controlling or not for environmental factors (see e.g., Growitsch et al., 2012), they have not carried out the above mentioned robustness analysis.

We need to address an important empirical issue caused by the fact that our dependent variable (i.e. electricity distribution costs) not only depends on weather and geographic factors to be aggregated, but also on a set of economic and technical variables (e.g., number of customers, energy delivered, network length, etc.) that might be correlated with the environmental variables. If this is the case, the exogeneity assumption of Li (1991) is violated and this procedure will likely give poor predictive composites. Indeed, the sliced inverse regression and other sufficient dimension reduction methods do not consider the presence of other explanatory variables than those to be aggregated, a feature that we label as the *all-inclusive* assumption. In order to relax this assumption and address the potential problems of endogeneity, we propose purging the data before using any supervised method. Moreover, to purge the data, we propose a partial regression approach that requires estimating auxiliary regressions where original variables are replaced with residuals obtained from previous regressions.

The proposed empirical strategy to tackle this issue is one of the contributions of the present paper. In addition, as we are primarily interested in weather and geographic factors, we also adapt conventional non-nested model selection tests to analyse a subset of the variables in a multiple regression. The adaptation is carried out by reversing the procedure used to purge the data mentioned above.

Our test results show that SIR and its parametric counterpart are superior to traditional variable reduction techniques. Moreover, using reasonable upper and lower bounds to identify utilities operating in areas with "normal" environmental conditions, we find evidence of superior ability of SIR and PIR methods to identify utilities disadvantaged

(enjoying) by adverse (favorable) environmental conditions. If SIR and PIR are the preferred specifications, this implies that some utilities would be penalized (rewarded) too strongly if a less rigorous approach (e.g., principal component) is instead used by the regulator. A traditional efficiency analysis indicates that, although there is scope for inefficient performance in the Norwegian electricity distribution sector, the model that utilises PCA composites does not capture any performance variation among the utilities.

Section 2 introduces SIR and other popular sufficient dimension reduction methods. Section 3 describes the data, the composite variables derivation, and the specification of the cost function used in the empirical exercise. Section 4 presents the main results using PCA, SIR and other supervised methods. Section 5 presents the conclusions.

2. Sufficient dimension reduction methods

Since Li (1991) introduced the sliced inverse regression, sufficient dimension reduction methods have become popular to use with large-scale data sets in fields such as biology, genome sequence modeling, or face recognition because massive datasets have been available in these fields for a relatively long time. The huge dimensionality problem in these fields boosted up the researchers interest in this issue and since Li's seminar paper various types of inverse regressions have been proposed, including the sliced average variance estimation (SAVE) (Cook and Weisberg, 1991), principal hessian directions (PhD) (Li, 1992), parametric inverse regression (PIR) (Bura and Cook, 2001a), and partial SIR (Chiaromonte et al., 2002) with categorical variables. The popularity of these inverse regression methods is due to the fact that most of them are computationally simple and have been demonstrated to be practically useful in the abovementioned fields.

We next describe briefly the procedure of the original sliced inverse regression method and later on we relax the all-inclusive assumption underlying the sufficient dimension reduction method.

Let Y_i denote the variable to be predicted (e.g., firms' cost) and x_i be a *p*-dimensional column vector of regressors that are to be aggregated (e.g., weather and geographic variables). Let X denote a matrix of *n* observations of x_i', Y be a vector of *n* observations of Y_i, and Σ_x be the covariance of X. A common characteristic of most popular sufficient dimension reduction methods is that they aim to replace the *p*-dimensional vector X with a smaller set of linear combinations of the original variables, provided that it captures all of the information that X contains about the variables to be predicted. A statistical formulation for addressing this issue is given in Li (1991), using the following fairly general model:

$$Y_i = f(\beta_1' x_i, \dots, \beta_K' x_i) + \varepsilon_i \qquad (1)$$

where $(\beta_1,\dots,\beta_K)$ are px1 vectors of unknown coefficients, and ε_i is a random term which is assumed to be independent of x_i. This model makes no assumption about either $f(\cdot)$ or the distribution of the error term. The response variable Y is related to the *p*-dimensional regressor x_i only through the reduced K-dimensional variable $(\beta_1' x_i, \dots, \beta_K' x_i)$. When this model holds, the projection of the *p*-dimensional explanatory variable x_i onto the K-dimensional subspace $(\beta_1' x_i, \dots, \beta_K' x_i)$ captures all the information about Y_i.

SIR and other sufficient dimension reduction methods are developed to find the space generated by the K unknown β vectors, called the *effective dimension reduction* (e.d.r.) space. This space should be estimated from the data and is based on the spectral

decomposition of a kernel matrix M that belongs to the central subspace, that is $S(M) \subset S_{Y|X}$.[4] Most sufficient dimension reduction methods are based on the same logic but use different kernel matrices. Once a sample version of M is obtained, the eigenvectors corresponding to the largest eigenvalues of $\widehat{M}$ are used to estimate a basis of the central subspace.

For effective dimension reduction, Li (1991) proposed to reverse the conventional viewpoint in which Y is regressed on X (forward regression) and showed that if X is standardized to have the zero mean and the identity covariance, the inverse regression curve $\eta(Y)=E(X|Y)$ fall into the e.d.r. space. Hence a principal component analysis on an estimation of E(X|Y) and its covariance matrix $\Sigma_\eta=\text{cov}[E(X|Y)]$ can be used to estimate e.d.r. directions.

The computation of SIR is quite simple and can be carried out in five steps:

1) Standardize X to $Z = \hat{\Sigma}_x^{-1/2}(X - \overline{X})$ where $\hat{\Sigma}_x^{-1/2}$ is any square root of the inverse of the sample covariance matrix for X (e.g., the inverse of the lower triangular Cholesky factor of $\hat{\Sigma}_x$) and $\overline{X}$ is the sample mean of X. In this scaling, Z has a zero mean and identity sample covariance matrix.
2) Divide the range of Y into H slices and compute the mean of Z in each slice, $\bar{z}_h$, h=1,…, H.
3) Form the weighted covariance matrix of these slice mean vectors, $\widehat{M}_{SIR} = \Sigma_{h=1}^{H} \hat{p}_h \bar{z}_h \bar{z}_h'$, where $\hat{p}_h$ is the proportion of cases falling into slice h.
4) Find the eigenvalues and eigenvectors for $\widehat{M}_{SIR}$.

[4] The central subspace is the smallest dimension-reduction subspace of $\mathbb{R}^p$ that provides the greatest dimension reduction in the predictor vector.

5) Let η_k (k=1,…, *d*) be the *d* largest eigenvector of $\widehat{M}_{SIR}$. The e.d.r. directions can be estimated by rescaling the previous eigenvectors, i.e. $\hat{\beta}_k = \hat{\Sigma}_x^{-1/2}\hat{\eta}_k$.

After standardizing X, step 2 produces a nonparametric estimate of the inverse regression curve E(Z|Y), which is the slice mean of Z after slicing the range of Y into several intervals and partitioning the whole data into several slices according to the Y value.[5] These slice means of Z are used in step 3 to obtain an estimate of the conditional covariance matrix Σ_η=cov[E(X|Y)], i.e. matrix $\widehat{M}_{SIR}$. In step 4, a PCA is applied to these slice means of Z in order to find the e.d.r. directions. Finally, step 5 simply retransforms the scale back to the original one.

Several comments are in order. First, the above eigenvalue decomposition differs from that of the traditional PCA. Here, the dependent variable is used to form slices while the standard PCA does not use any information from Y. Second, Chen and Li (1998) showed that SIR can be interpreted as a (multiple linear) regression analysis and established that SIR searches for the best e.d.r. directions without knowledge of the functional form f(·). Third, Li (1991) and Cook and Weisberg (1991) pointed out that the performance of SIR is not sensitive to the number of slices used in partitioning the data matrix. Finally, under general conditions the linearity assumption by Li (1991) holds asymptotically for high dimensional data (see Hall and Li, 1993). Therefore, when the dimensions are high, SIR can work well. These features lead us to focus our analysis on SIR or its parametric counterpart PIR, as outlined below.

[5] Note that SIR uses the response variable only to create slices and only the order of Y is needed in slicing. Hence it is indifferent whether we use a monotonic transformation of the response variable such us the logarithm of Y.

In SIR, Z are regressed on a discrete version of Y resulting from slicing its range in order to obtain a rather crude estimate of the inverse curve E(Z|Y). If Y is a continuous variable, its transformation to a categorical variable might imply ignoring much information about the inverse relationship between Z and Y. PIR is another first-moment based method for effective dimension reduction that aims to estimate the central subspace without imposing any restrictions on the nature of the response variable Y. This is achieved using least squares estimates of each standardized explanatory variable on a set of arbitrary functions of Y. The fitted values of E(Z|Y) are then used as an estimate of M, that is

$$\widehat{\mathrm{M}}_{\mathrm{PIR}} = (\mathrm{F_i}'\mathrm{F_i})^{-1}\mathrm{F_i}'\mathrm{Z_i} \qquad (2)$$

where $F_i=F_i(Y_i)$ is a matrix of centered functions of Y. For instance, Bura and Cook (2001a) suggest using a quadratic function of Y. Because polynomial functions of Y are used, this method is known as a polynomial inverse regression.

Although SIR and other supervised methods are statistical in nature, it is useful to see the challenges faced by any supervised method using a simple theoretical model. This not only helps to understand better the nature of these challenges but also to shed light on the nature of measurement errors in (un)supervised composites.

Assume that the theoretically model to be estimated is $Y = \alpha_1 x_1 + \alpha_2 x_2$. Note that the coefficients of both observed variables, α_1 and α_2, capture the theoretical effect of each variable on Y. From a theoretical point of view, this effect does not rely on how x_1 and x_2 are aggregated using statistical techniques. An equivalent way to write the model above using a *theoretical* composite $f=f(x_1, x_2)$ is $Y = \alpha \cdot [(\alpha_1/\alpha)x_1 + (\alpha_2/\alpha)x_2] = \alpha \cdot f(x_1, x_2)$, where $\alpha = \alpha_1 + \alpha_2$ measures the overall effect of both variables. The

latter specification simply indicates that both explanatory variables should be weighted using their relative effects on Y in order to estimate the same α. In other words, the *empirical* composite should aggregate the explanatory variables taking into account (somehow) the dependent variable. This is precisely the aim of the *supervised* methods for reducing the dimension of the data. In contrast, unsupervised methods only use information on how x_1 and x_2 are statistically distributed, how large their variances are, or whether they are highly correlated.

So far we have assumed that the p-dimensional vector X_i includes the whole set of available variables and, hence, equation (1) does not consider the presence of omitted explanatory variables that might be correlated with X_i. In particular equation (1) ignores both the basic economic theory on production economics and the available empirical literature on the electricity distribution networks that demonstrate that utilities' costs are increasing functions of *economic-in-nature* variables, such as number of customers, amount of energy delivered, network length, and labor and capital prices, etc. Under the presence of additional regressors, not only we have a traditional endogeneity problem but also the performance of supervised composites might degenerate drastically. If we add the economic variables to equation (1), the theoretical-consistent model to be estimated can be written as

$$Y_i = f(c_i, \theta) + g(Q_i, P_i, \gamma) + \varepsilon_i \qquad (3)$$

where θ and γ are parameters to be estimated, $c_i = \left(\hat{\beta}_1' x_i, \dots, \hat{\beta}_K' x_i\right)$ is the matrix containing the estimated composites that were obtained using PCA, SIR and other supervised methods, and $g(\cdot)$ can be viewed as the traditional cost function in production economics which, by theory, should be increasing in outputs, Q_i, increasing in input prices, P_i, and linearly homogeneous in input prices. As weather is a complex

phenomenon and its overall effect on cost is unknown, we take an agnostic position and do not make specific assumptions about the probable (partial) effect of specific weather composites on distribution costs.

In order to address both the likely endogeneity problem and poor performance of supervised composites, we rewrite (3) as follows:

$$Y_i^* = Y_i - g(Q_i, P_i) = f(c_i, \theta) + \varepsilon_i \qquad (4)$$

Hence, the above equation suggests that previous methods on sufficient dimension reduction can still be used after we have properly *corrected* the dependent variable for the effect of the economic variables. In order to achieve this objective we propose the so-called *partial* regression approach used to measure the contribution of a subset of the explanatory variables (in our case, economic drivers of costs) in multiple regression models taking into account that other explanatory variables (in our case, the weather/geographic composites) are already included in the model and that they might be correlated with the economic variables.

An adjusted dependent variable can be obtained following Frisch-Waugh (1933)'s theorem by estimating an auxiliary regression where original variables are replaced with residuals obtained from previous regressions (see Greene, 2002, p. 26). The aim of this regression is to net out the effect of the weather and geographic composites on both the dependent variable (i.e. costs) and the economic variables. This procedure allows us to isolate the effect of the economic variables on cost, i.e. $g(\cdot)$ to purge the dependent variable.

In particular, a properly *corrected* dependent variable can be obtained as follows. First, we regress Y_i on the weather and geographic variables using the original data and obtain the associate residuals:[6]

$$e_i = e_i(Y|X) = Y_i - \widehat{Y}_i(X_i) \quad (5)$$

Second, we regress each economic variable on the weather and geographic variables and obtain the residuals:

$$d_i = d_i(z|X) = z_i - \hat{z}_i(X_i) \quad \text{where } z = P \text{ or } Q \quad (6)$$

Finally, we regress the first set of residuals on the second set of residuals and obtain $g(\cdot)$ from the parameters of the *auxiliary* regression

$$e_i = g(d_i) + v_i \quad (7)$$

The coefficients of (7) allow us to compute the costs attributed *only* to economic variables taking into account that the weather and geographic composites are also costs drivers and that they might be correlated with the economic variables. This estimate is then used to obtain Y_i^*. As this the new dependent variable is purged of the effect of economic variables, it should provide a better starting point to implement our supervised methods to aggregate our set of weather and geographic variables.

[6] Note that here we are not interested in the interpretation of individual coefficients or in statistical properties (i.e. t-ratios) of the estimated coefficients. We carry out this regression in order to implement the two-step Frisch-Waugh (1933) procedure. One potential problem could be the number of geographic and weather variables. If these are too large, we could carry out this regression by replacing the individual variables with PCA composites. Recall that PCA does not require using information on the response variable. In this case, we use PCA as an intermediate tool in our procedure in order to control for endogeneity issues when using supervised variable reduction techniques.

3. Data, composite variables derivation, and cost function specification

The data used to illustrate the relative performance of the different variable reduction methods is a balanced panel for 128 Norwegian electricity distribution networks for the years 2001 to 2004. The data was provided by the sector regulator, the Norwegian Water Resources and Energy Directorate (NVE). Norway presents a particularly interesting case to study the effect of environmental factors. First, Norway was among the first countries to introduce efficiency benchmarking (based on the non-parametric data envelopment analysis technique) and incentive-based regulation in this sector in 1997. As a result, the managerial inefficiency of the networks has been reduced (Førsund and Kittelsen, 1998; Edvardsen et al., 2006). Second, unlike most countries, the Norwegian electricity sector consists of a large, though slowly declining (due to mergers and acquisitions), number of network utilities which allows the use of more sophisticated analytical methods, such as supervised methods and semi-parametric estimators. Finally, the Norwegian energy regulator has made efforts to take the effects of environmental factors on the cost and service quality performance into account and include these in the benchmarking models. In particular, the regulator intends to analyze a large number of geographic and weather variables with a view to use unsupervised dimension reduction techniques to construct composite indices. The composites are then to be used in a second-stage regression model to adjust for their effect on the estimated relative inefficiencies. This approach implies that the (geographic and weather) composites affect the firms' efficiency rather than their cost function.

We specify a simple cost model with two outputs (CU=customer numbers, and ED=energy delivered), two input prices (KP=capital price and LP=labor price), a proxy for size of the system (NL=network length), and two indices capturing technological

characteristics of the network (HV and TR). The two outputs are the number of final customers and the energy supplied measured in megawatt-hours (MWh). These two variables reflect the different marketable goods of the joint service of electricity distribution.[7] The first technological index is HV, i.e. the percentage of high-voltage network in total network, which is included as an explanatory variable to capture differences in maintenance or acquisition of equipment (likely more expensive for high-voltage network). This is an empirical question as high-voltage grid might be cheaper than low-voltage grid if the latter is mostly underground cable in urban areas. TR is the number of transformers and is used as proxy for assets and network capacity. Alternatively, we can interpret the coefficient of TR in terms of changes in the *ratio* of transformers to network length (i.e. changes in a new technological index) because network length is already included as a cost driver. The labor price is the average salary in electricity sector (NOK/month) and the price of capital is the price index for power sector goods.

Regarding the dependent variable, and following the Norwegian benchmarking approach, we incorporate quality of service into our model by using, as dependent variable, *social costs* instead of total production costs. In addition to operating expenses and capital expenditures, social costs, C, include external quality costs. External quality costs are calculated by the multiplication of the energy not supplied with the estimated

[7] Customer numbers and units of energy delivered are the most commonly used outputs in benchmarking of distribution network utilities (Giannakis et al., 2005; Yu et al. 2009a, 2009b). These output variables are important cost drivers and influence the pricing of distribution services.

customer willingness-to-pay for an uninterrupted energy supply. Summary statistics of the economic data are shown in Table 1.[8]

[Insert Table 1]

Given the abovementioned cost drivers, the *basic* cost function to be estimated can be written as:

$$\ln C_{it} = g_1^{TL}(CU_{it}, ED_{it}, NL_{it}, TR_{it}, \alpha) + g_2^{CD}(LP_{it}, KP_{it}, \gamma) + \varepsilon_{it} \quad (8)$$

where the subscripts i and t stand respectively for utility and time, α and γ are the parameters to be estimated, and our former dependent variable has been replaced with the log of social costs, $\ln C_{it}$. Superscript TL in $g_1^{TL}(\cdot)$ stands for Translog specification, and superscript CD in $g_2^{CD}(\cdot)$ indicates that only a Cobb-Douglas specification is estimated. We add the log of capital and labor prices to our cost function because they do not vary across utilities, but only vary over time. This precludes using quadratic terms and interactions with these variables.

In addition to the above economic variables, we incorporate a varied set of weather and geographic variables in our analysis. This includes variables for temperature, precipitation, wind speed, distance to coast, slope, population concentration, forests and so forth. Once we drop variables that are linear functions of others, our data consist of 35 weather variables and 32 geographic variables. In order to further reduce the number of environmental variables to a manageable number and to avoid the problem of multi-collinearity, we estimate composite factors for the geographic and weather conditions using two sufficient dimension reduction methods: SIR and its parametric counterpart, i.e. PIR. We use PCA as benchmark as it is the dominant method of dimension

[8] All monetary variables are in 2004 real terms.

reduction in production economics and efficiency analysis.[9] Then, we incorporate the estimated geographic and weather composites as additional cost drivers in our model. The final cost function that includes geographic and weather composites can then be written in a compact form as:

$$\ln C_{it} = g_1^{TL}(\alpha) + g_2^{CD}(\gamma) + f^{TL}(c_{it}, \theta) + \varepsilon_{it} \qquad (9)$$

where c_{it} is a previously computed vector of environmental composites, and function $f(\cdot)$ is estimated using a Translog specification. Growitsch et al. (2012) applied dimension reduction to separable types of variables (though not all weather and geographic variables together as we do here) because weather and geographic conditions are different in nature. This somewhat facilitates interpretation, but ignores the feedback effects among weather and geographic conditions.[10]

So far we have described the deterministic part of the model, i.e. the effect on the firms' cost function of both economic and environmental variables. However, the estimated residual in (12) can be used to explore whether relative efficiency depends on the method selected to control for geographic and weather conditions. In order to achieve

[9] We also used the most popular second-moment based methods, such as SAVE, SIR-II or pHd. Results from these alternative methods are not shown in the next sections due to space limitations, and because their performance was (and often considerably) worse than that of SIR and PIR. They are, however, available from the authors upon request.

[10] In a previous version of this paper we examined whether a separable or joint analysis is more appropriate applying the Voung tests to two competing models, one with two composites derived from weather and geographic information, and other model with one separable composite for weather and geography respectively. Our results indicated that we should not treat weather and geographic information differently when computing our composites.

this we follow the well-known stochastic frontier analysis (SFA) literature[11] and assume that the error term (ε_{it}) can be decomposed into a traditional noise term (v_{it}) and a random efficiency term (u_{it}), i.e. $\varepsilon_{it} = v_{it} + u_{it}$. In particular, we apply standard SFA techniques to a normal-half-normal model, and assume that $v_{it} \sim N(0,\sigma_v)$ and u_{it} is a nonnegative half normal, i.e. $u_{it} \sim N^{+}(0,\sigma_u)$.

In order to obtain the efficiency scores for each firm we use a three-stage approach suggested in various models in Kumbhakar and Lovell (2000). In the first stage, we estimate the cost function (12) using ordinary least squares, OLS.[12] The above distributional assumptions are then invoked to obtain estimates of the parameter(s) describing the variance of v_{it} and u_{it}, conditional on the first-stage residuals.[13] Finally, efficiency scores for each frirm are estimated for using the conditional distribution of u_{it} given ε_{it} introduced by Jondrow et al. (1982). The main advantage of this procedure is that no distributional assumptions are used in the first stage (i.e., in OLS) and that in the first stage the error components are allowed to be correlated. Standard distributional assumptions on u_{it} and v_{it} are only used in the second and third stages of the procedure.

[11] For an overview of this literature see Kumbhakar and Lovell (2000) and Coelli et al. (2005).

[12] Note that here the expectation of the composed error term is not zero as the expectation of the inefficiency term is positive. This implies that the OLS estimate of the intercept in (12) is biased. However, a consistent estimate of this coefficient can be obtained once the parameters describing the distribution of u_{it} are estimated.

[13] The second-stage estimators are obtained by maximizing the likelihood function associated with the residuals that can be obtained from an estimate of the first-stage cost function (12). The residuals from the first stage are $\hat{\varepsilon}_{it} = v_{it} + (u_{it} - E(u_{it}))$. If u_{it} follows a half-normal distribution, then $E(u_{it}) = \sqrt{2/\pi} \cdot \sigma_u$. Thus, the stochastic frontier model in the second stage is $\hat{\varepsilon}_{it} = v_{it} + u_{it} - \sqrt{2/\pi} \cdot \sigma_u$ where the parameters σ_u and σ_v are estimated by ML.

4. Results

Before presenting the parameter estimates, we first examine the relative performance of alternative composites using non nested model selection tests. In particular, for model selection we consider the widely-used the Bayesian Information Criterion (BIC), proposed by Schwarz (1978). BIC involves minimizing an index that balances the lack of fit (too few composites) and overfitting (too many composites) as it includes a penalty that increases with the number of regressors. Hence, models with lower BIC are generally preferred.[14]

If we apply directly any model selection criteria to equation (3), none of them will be quite informative about the relative performance of any methods for aggregating weather and geographic factors because the economic-in-nature variables explain most variation in electricity distribution costs.[15] As we are mainly interested in the correlation between the costs and weather and geographic factors, we adapt the traditional model selection tests to the analysis of a subset of the variables in a multiple regression setting. In order to address this issue, we make use of the so-called *partial* R^2 in multiple regression models. A *partial* R^2 allows us to measure the contribution of a subset of the explanatory variables (in our case, the weather and geographic composites) taking into

[14] An alternative model selection statistic is the widely-used Akaike's Information Criterion (AIC) proposed by Akaike (1973). We use the BIC formulation because it uses a larger penalty than AIC, and hence it tends to favor more parsimonious models than AIC, which in turn help estimate models coefficients with more precision (see Verbeek, 2008, p. 61).

[15] In our application, only one explanatory variable (e.g. customer number or energy delivered) explains about 97% of the cost variations.

account that: i) other explanatory variables (in our case, drivers of economic costs) are already included in the model, and ii) they might be correlated with the environmental variables. In particular, a *partial* R^2 for weather and geographic composites can be obtained as follows. First, regress Y_i on the economic variables Q_i and P_i and obtain the residuals. Second, regress each environmental variable on the economic variables Q_i and P_i, and obtain again the residuals. Finally, regress e_i on d_i, and obtain the R^2 statistic between these two sets of residuals. The computed R^2 from the last regression can be interpreted as a *partial* R^2 because it measures the (pure) relationship between costs and environmental variables once we control for the effect of economic variables.

The above model selection procedures have been criticized because the deterministic nature of these criteria means that no information is provided as to ‘‘how much’’ better is the chosen model (i.e., they do not allow probabilistic statements to be made regarding model selection). Therefore, we also apply the Vuong's (1989) model selection framework to select the most adequate model designed to test the null hypothesis that two competing models fit the data equally well versus the alternative that one model fits better. If the null is not rejected we can conclude that both competing models are equivalent. Otherwise, we conclude that the competing models can be statistically differentiated, and the sign of the test indicates which model dominates the other in the sense of being closer to the true model.

Based on the cost function (9), Figure 1 shows the overall R^2, and the partial R^2 and BIC values obtained by including different numbers of composites. Both weather and geography variables are used to compute up to 8 composites. The figure indicates that the overall goodness-of-fit for all models is very high. This high value is the consequence of size effect as most of cost variation is explained by the number of

customers, energy delivered or network length. In addition, these economic variables are strongly correlated with both weather and geographic variables and hence they are also capturing the effect on cost of environmental variables. For these two reasons, the overall R^2 statistic does not help us to analyze the relative performance of PCA, SIR and PIR when aggregating weather and geographic factors. Figures 1b and 1c show, as expected, that the partial R^2 and the partial BIC are significantly more informative as they isolate the contribution of weather and geographic composites. These figures clearly show the superior performance of SIR and its parametric counterpart, PIR. Only for large numbers of composites the performance of PCA is similar to that of SIR and PIR. Note, on the other hand, that the BIC figure reaches the minimum value when five composites are used. For this reason, a cost function with five SIR composites is, hereafter, our preferred model.

[Insert Figure 1]

The supervised composites have been computed after costs were purged using the procedure outlined in Section 3. The partial BIC values for SIR and PIR using actual cost are shown in Figure 2. Two comments are in order. First, the performance of both types of composites is similar for large numbers of composites. However, for small or moderate number of composites, the figure makes evident the superior performance of supervised methods that control for potential endogeneity based on corrected costs. Second, this better performance is especially evident for PIR composites. This result is likely due to the fact that SIR uses a crude measure of the inverse curve E(X|Y), while PIR uses a parametric but more precise estimate of E(X|Y) and, hence, more sensitive to the presence of additional regressors correlated with weather and geographic variables.

[Insert Figure 2]

As a robustness analysis, we examine again the relative performance of different dimension reduction methods in Table 2, but this time we use the Vuong test that compares the goodness-of-fit of two competing (unsupervised vs. supervised) models when using different number of composites. A positive sign in this table indicates that the model that appears on the right is closer to the true model than the model that is on the left, and vice versa. In Tables 2a and 2b we test the relative performance of the most commonly used dimension reduction method, PCA, against SIR and PIR. The estimated values of the Voung test are positive except when eight composites are used, indicating again that PCA tends to perform worse than SIR and PIR. It should be noted that the estimated differences in goodness-of-fit are statistically significant for moderate numbers of composites even though the Voung test might not perform well because our models are highly overlapped (Shi, 2011). This result is additional evidence in favour of SIR and PIR. On the other hand, Table 2c shows that both supervised methods perform similarly, although SIR clearly outperforms PIR when two or eight composites are used.

[Insert Table 2]

Table 3 shows the estimated coefficients of the cost function using our preferred model, i.e. the cost function with five SIR composites. For comparison, we present the parameter estimates of a model without environmental composites and the corresponding model using the commonly used dimension reduction method, PCA. The variables are measured in deviations from the geometric mean and hence first-order coefficients are elasticities (or derivatives) evaluated at the sample mean. Table 3 shows that all elasticities of social cost with respect to customer numbers, energy delivered,

and network length in our preferred model have the expected (positive) signs at the sample mean. The sum of the first order coefficients of the outputs is the scale elasticity and if this value is smaller than one, scale advantages indicate natural monopoly characteristics (see Salvanes and Tjøtta, 1998). Although the scale elasticity is similar using either SIR or PCA composites, there are notable differences in individual variables. For instance, the first-order coefficient of customer numbers (energy delivered) using PCA is much higher (lower) than that found using a SIR specification. This suggests that our results vary with the method used to construct environmental composites. The first-order coefficient of HV is positive and statistically significant, indicating that an increase in high-voltage network percentage yields higher maintenance costs. It might also indicate that high-voltage equipment are more costly. The second technological index, TR, has a positive first-order coefficient, indicating that a higher ratio of transformers to network length tends to increase the network costs. It is noteworthy that the coefficient of labor price is positive and significant. The coefficient of capital price is positive as well, though not statistically significant.

[Insert Table 3]

Regarding the environmental factors, the estimated coefficients are shown in the Appendix. Although we take an agnostic position and do not make specific assumptions about the sign of these coefficients, it is noteworthy that most of them are statistically significant. Hence, we can conclude that weather and geographic conditions matter and that they should be included as cost determinants.[16] Moreover, the estimated coefficients suggest the existence of remarkable differences among utilities in cost

[16] Growitsch et al. (2012) have found similar results using a similar specification of the technology of the Norwegian electricity distribution networks.

attributed to different environmental conditions. This is likely what regulators wish to control for. The estimated differences are summarized in Figure 3 where we plot kernel density functions of the percentage of cost attributed to (unfavourable) environmental conditions, measured in relation to the "average" firm, obtained using our preferred model based on SIR composites.[17]

The distribution of the estimated SIR environmental cost differences is slightly symmetric. However, we find a concentration mass in both tails of the distribution for the SIR model. The observations belonging to any of these two concentration masses are utilities that are likely operating in areas with either extreme bad or good environmental conditions. This issue is explored in more detail later, but Figure 3 shows that the SIR model predicts up to 44% lower costs for utilities operating in areas with favourable environmental conditions. A correlation analysis suggests that the lower costs are mostly associated with "urban" factors and relative importance of district heating systems, i.e. higher urbanisation and more output delivered from district heating in the concession area lead to lower costs. Higher shares of agricultural land around roads and lines reduce cost as well. Distribution costs also tend to decrease in warmer parts of Norway. On the other hand, for utilities operating in unfavourable environmental conditions the SIR model predicts up to 35% higher costs. These higher costs are mostly caused by weather conditions: an increase in snow, rain precipitations, and wind tend to increase costs. Geographic factors such as the hilliness and slope of ground and the presence of water bodies (e.g. wetlands or lakes) or shallow soil that may be moved by wind seem to be correlated with higher distribution costs.

[17] The so-called average firm is obtained by adding the first-order coefficients of the five composite variables as all explanatory variables are in deviations with respect to their respective means.

[Insert Figure 3]

As the above correlations make sense, we can conclude that our empirical strategy to control for environmental factors based on SIR composites performs well. Recall that a SIR specification of the model is preferred because the PCA specification is rejected using the Vuong test (see Table 2) and partial R^2 (BIC) statistic is higher (lower). We next examine the results in comparison to PCA composites. Figure 4 shows a scatter plot of the estimated environmental cost differences using SIR and PCA composites. As shown, the estimated cost differences from these two composites are not highly correlated. In addition, the coefficient of correlation (0.43) between the competing models is quite small. These results indicate that controlling for the effect of differences in environmental factors is a difficult task and, hence, the results of efficiency analysis in an incentive regulation framework might be biased if regulators do not select the proper method to construct environmental composites. For instance, regulators might reward some utilities in excess because they have been wrongly labelled as suffering from harsh environmental conditions.

[Insert Figure 4]

Figure 5 shows the box plots of the percentage of cost attributed to unfavourable environmental conditions, measured in relation to the "average" firm. We might assume that values smaller or larger (smaller) than the upper (lower) adjacent values are identifying utilities operating in areas with extreme bad (good) environmental conditions. It is conceivable that a firm lying within these bounds has 'normal' weather and geographic conditions. The range of values of the estimated environmental cost

differences using SIR is wider than those using a PCA model.[18] However, the number of observations with 'normal' weather or geographic conditions in the SIR model is somewhat lower than those detected using a PCA model. In other words, while PCA only detects 8 observations (in fact 2 utilities) with over-cost percentages larger than the upper adjacent value, SIR detects 55 observations (i.e. about 13 utilities) with extremely good (8 utilities) or bad (5 utilities) environmental conditions. We can intrepret this as an additional evidence of the higher ability of SIR methods to identify utilities suffering (or enjoying) extreme bad (good) environmental conditions. Moreover, PCA only detects utilities with extreme unfavourable conditions, while SIR is able to detect utilities with both good and bad extreme conditions. It is noteworthy that the SIR set of utilities does not nest the PCA set as the two utilities identified using the PCA model do not appear in the SIR set.

[Insert Figure 5]

So far we have analysed the results regarding the estimated cost function, i.e. the deterministic part of the model. We have made no distributional assumptions when estimating the coefficients of the cost functions in Table 3. We next assume, following the SFA literature, that the estimated error term can be decomposed into a noise term ($v_{it} \sim N(0,\sigma_v)$), and a one-sided random term capturing firms' inefficiency ($u_{it} \sim N^{+}(0,\sigma_u)$) in order to form an idea about the robustness of a traditional efficiency analysis if either unsupervised or supervised aggregation methods are used, an issue not examined yet.

[18] This result makes sense because PCA composites are computed ignoring their relationship with the dependent variable. As the potential explanatory power of PCA composites is weaker than that of SIR composites, OLS tends to soften the cost effect of weather and geographic conditions.

Table 4 shows the ML parameters estimates of (the log of) both standard deviations σ_u and σ_v describing the structure of the error terms, conditional on the first-stage estimated parameters. The variance of the inefficiency term is higher than the variance of the traditional noise term when no composites are used. In particular the statistic γ in Table 4 indicates that random shocks, which are captured by the traditional symmetric error term, explain 56% of the overall variance of the composed error term.[19] However, when either SIR or PCA composites are included in the model, the variance of the inefficiency term tends to be much lower. This result is consistent with our expectation as ignoring relevant factors in efficiency analysis tends to overestimate the efficiency scores because the relative differences among observations tend to be much higher.

[Insert Table 4]

Table 5 provides summary statistics of the estimated efficiency scores from the compared models. On average, the efficiency scores obtained from the more restrictive model (about 90.5%) are relatively lower than the efficiency scores using SIR or PCA composites (about 96 and 99% respectively). This increase is particularly high in specifications that use PCA composites. Moreover, the variance estimate of the inefficiency term in the PCA model is not statistically significant. This result indicates that, although there is scope for different degrees of inefficient performance in the Norwegian electricity distribution sector,[20] the model that uses PCA composites is not able to capture any variation in (relative) performance among utilities. Clearly this result (i.e. the lack of inefficiency) might not happen in other applications. We simply

[19] The expression $\gamma = \sigma_u^2/(\sigma_v^2 + \sigma_u^2)$ is often used in the SFA literature to measure the relative importance of inefficiency in total variation.

[20] This is expected as this sector is incentive regulated using cost benchmarking and efficiency analysis.

give some evidence about the consequences of using less accurate methods to control for environmental conditions in traditional efficiency analyses.

[Insert Table 5]

The methodology and results presented show that weather and geographic factors affect performance of electricity networks. This has implications for the regulator and the firms affected. Heterogenuity of the benchmarking samples are a constant cause of discussion among the regulators and firms as much revenue implications rests on the results obtained from the benchmarking models. At the same time, the number of potential factors affecting the performance of network utilities is large. Using supervised methods the regulators can effectively maximise the use of information among many factors with complex interactions. This in turn increases the accuracy and fairness of the regulatory process and results.

5. Summary and conclusions

Although supervised dimension reduction methods are commonly used with large-scale data sets in biology, genome sequence modeling or face recognition, they have hardly been used in applied economics, and in particular in production or cost function modeling. However, the methods can be particularly useful in regulated sectors and efficiency analysis where firms' performance is affected by weather and geographic conditions which involve a large and varied number of non-discretionary factors with complex interactions, leading to a 'dimensionality' problem. We have shown how these techniques can be used in production or cost function modeling with application to the effect of weather and geography on electricity distribution costs in Norway. To our

knowledge, this is the first attempt to apply these techniques to production and efficiency analyses.

We also addressed issues caused by the fact that our response variable (i.e. electricity distribution costs) not only depends on the weather and geographic factors to be aggregated, but also on a set of economic variables (e.g., number of customers, energy delivered, network length, etc.) that might be correlated with the environmental variables. In order to address these issues we have proposed using the Frisch-Waugh (1933)'s theorem that allows us to work with subsets of explanatory variables in a multiple regression setting.

Regarding our application, we first examined the relative performance of alternative composites using a model selection approach. Our test results are quite robust and clearly show the superior performance of SIR and its parametric counterpart, PIR. Moreover, PCA performs much worse than other supervised methods. However, its performance tends to be similar to that of SIR for large numbers of composites. Overall, our preferred model is a cost function with five SIR composites.

Regardless of the methodology used to avoid the dimensionality problem, the estimated coefficients of our environmental composites are statistically significant indicating that weather and geographic conditions matter and that they should be included as cost determinants. We also found large differences among utilities in costs attributed to environmental conditions. Our preferred model predicts up to 44% lower costs for utilities operating in areas with favourable environmental conditions. For utilities operating in unfavourable environmental conditions higher costs of up to about 35% are predicted. While lower costs are mostly associated with "urban" factors, district heating systems, and agrarian farms around roads and lines, higher costs are mostly caused by

weather conditions and two geographic factors reflecting the hilliness and slope of the ground and the presence of water bodies (e.g., wetlands or lakes), or shallow soil. Moreover, using reasonable upper and lower bounds to identify utilities operating in areas with "normal" environmental conditions, we found evidence of the superior ability of SIR methods to identify utilities disadvantaged by (enjoying) or adverse (favorable) environmental conditions.

Finally, we examined the robustness of a traditional efficiency analysis in order to use different aggregation methods to control for environmental factors. The variance of the inefficiency term tends to be much lower when either SIR or PCA composites are included in the model. This result is consistent with our expectation as ignoring relevant factors in efficiency analysis often tend to amplify the relative differences among observations. Our analysis also indicates that, although there is scope for different degrees of inefficient performance in the Norwegian electricity distribution sector, the model that utilizes PCA composites is not able to capture any performance variation among the utilities.

References

Adler, N. and Yazhemsky, E. (2010) "Improving discrimination in data envelopment analysis: PCA-DEA or variable reduction", *European Journal of Operational Research*, 202, 273-284.

Adragni, K.P. and Cook, D. (2009) "Sufficient dimension reduction and prediction in regression", *Philosophical Transactions of The Royal Society*, 367, 4385-4405.

Akaike, H. (1973) "Information theory and an extension of the maximum likelihood principle", *In 2nd International Symposium on Information Theory*, Petrov, B. and F. Csaki, T (Eds), 267-81, Budapest.

Bura, E. and Cook, R.D. (2001a) "Estimating the structural dimension of regressions via parametric inverse regression", *Journal of the Royal Statistical Society*, Series B, 63, 393-410.

Bura, E. and Yang, J. (2011) "Dimension estimation in sufficient dimension reduction: A unifying approach", *Journal of Multivariate Analysis*, 102, 130–142.

Bura, E., and Cook, R.D. (2001b), "Extending sliced inverse regression: The weighted chi-squared test", *Journal of the American Statistical Association*, 96, 996–1003.

Chen, C.H., and Li, C.H. (1998) "Can SIR be as popular as multiple linear regression", *Statistica Sinica*, 8, 289-316.

Chen, P. and Smith, A. (2007) "Dimension reduction using inverse regression and nonparametric factors", unpublished manuscript, Department of Agricultural and Resource Economics, University of California, Davis.

Chiaromonte, F., Cook, R. D. and Li, B. (2002) “Sufficient dimension reduction in regression with categorical predictors”, *Annals of Statistics*, 30, 475-497.

Coelli, T., D.S. Prasada, D.S. and Battese, G. (2005), *An introduction to Efficiency and Productivity Analysis*, Kluwer Academic Publishers, 2nd Edition.

Cook, R. D. and Weisberg, S. (1991) “Discussion of "Sliced inverse regression for dimension reduction", *Journal of the American Statistical Association*, 86, 28-33.

Cook, R.D. (1998), *Regression Graphics: Ideas for Studying Regressions through Graphics*, Wiley, New York.

Cook, R.D. (2007) “Fisher Lecture: Dimension reduction in regression”, *Statistical Science*, 22(1), 1-26.

Edvardsen, D.F., Førsund, F., Hansen, W., Kittelsen, S.A.C. and Neurauter, T. (2006) “Productivity and regulatory reform of Norwegian electricity distribution utilities”, in Coelli, T., Lawrence, D. (Eds.), *Performance Measurement and Regulation of Network Utilities*, Edward Elgar, Cheltenham.

Fan, J. and Li, R. (2006) "Statistical challenges with high dimensionality: Feature selection in knowledge discovery", Proceedings of the International Congress of Mathematicians, Madrid, Spain, European Mathematical Society.

Fisher, R. A. (1922) “On the mathematical foundations of theoretical statistics”, *Philosophical Transactions of The Royal Society*, 222, 309–368.

Førsund, F.A. and Kittelsen, S.A.C. (1998) “Productivity development of Norwegian electricity distribution utilities”, *Resource and Energy Economics*, 20, 207–224.

Frisch, R., and Waugh, F. (1933) "Partial time regressions as compared with individual trends", *Econometrica*, 1, 387-401.

Giannakis, D., Jamasb, T. and Pollitt, M. (2005) "Benchmarking and incentive regulation of quality of service: An application to the UK electricity distribution networks", *Energy Policy*, 33(1), 2256-2271.

Greene, W. (2002), *Econometric Analysis*, Prentice Hall, New York.

Growitsch, C, Jamasb, T. and Wetzel, H. (2012) "Efficiency effects of observed and unobserved heterogeneity: Evidence from Norwegian electricity distribution networks", *Energy Economics*, 34(2), 542–548.

Hall, P. and Li, K.C. (1993) "On almost linearity of low dimensional projections from high dimensional data", *The Annals of Statistics*, 21, 867-889.

Jamasb, T. and Pollitt, M. (2001) "Benchmarking and regulation: International electricity experience", *Utilities Policy*, 9(3), 107-130.

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

Jondrow, J., C.A.K. Lovell, S. Materov and P. Schmidt (1982) "On the estimation of technical efficiency in the stochastic frontier production function model", *Journal of Econometrics*, 19(2/3), 233-238.

Kumbhakar, S. and Lovell, C.A.K. (2000), *Stochastic Frontier Analysis*, University Press, Cambridge.

Li, K. (1991). Sliced inverse regression for dimension reduction (with discussion). *Journal of the American Statistical Association*, 86, 316–342.

Li, K. (1992) “On principal hessian directions for data visualization and dimension reduction: Another application of stein’s lemma”, *Journal of the American Statistical Association,* 87, 1025-1039.

Naik, P.A., Hagerty, M.R. and Tsai, C-L, (2000) "A new dimension reduction approach for data-rich marketing environments: Sliced inverse regression", *Journal of Marketing Research*, 37(1), 88-101.

Nieswand, M., Cullmann, A. and Neumann, A. (2009), “Overcoming data limitations in nonparametric efficiency measurement. A PCA-DEA application to natural gas transmission”, paper presented at INFRADAY 2009, October, Berlin.

Salvanes, K. and Tjøtta, S. (1998) “A test for natural monopoly with application to Norwegian electricity distribution”, *Review of Industrial Organization*, 13(6), 669-685.

Schwarz, G. (1978) “Estimating the dimension of a model”, *Annals of Statistics*, 6, 461-644.

Shi, X. (2011) “Size distortion and modification of classical Vuong tests”, unpublished paper, University of Wisconsin – Madison.

Tibshirani, R. (1996) "Regression Shrinkage and Selection via the Lasso", *Journal of the Royal Statistical Society*. Series B (Methodological), 58(1), 267-288.

Verbeek, M. (2008), *A Guide to Modern Econometrics*, third edition, John Wiley and Sons, Ltd.

Vuong, Q.H. (1989) “Likelihood ratio tests for model selection and non-nested hypotheses”, *Econometrica*, 57(2), 307–333.

Yu, W., Jamasb, T. and Pollitt, M. (2009a) "Does weather explain cost and quality performance? An analysis of UK electricity distribution companies", *Energy Policy*, 37(11), 4177-4188.

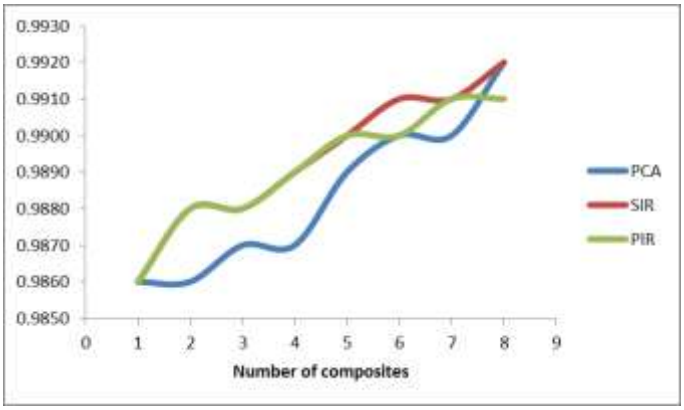
Yu, W., Jamasb, T. and Pollitt, M. (2009b) "Willingness-to-pay for quality of service: An application to efficiency analysis of the UK electricity distribution utilities", *Energy Journal*, 30(4), 1-48.

Zhu, J. (1998) "Data envelopment analysis vs. principle component analysis: An illustrative study of economic performance of Chinese cities", *European Journal of Operational* Research, 11, 50-61.

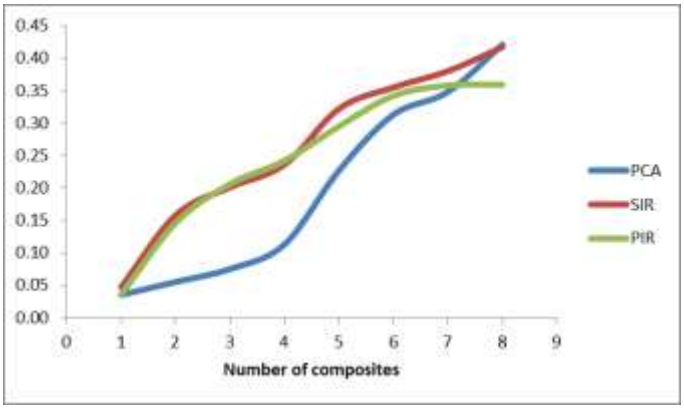
.

Figure 1. Overall and partial goodness-of-fit

(a) R-squared



(b) Partial R^2



(c) Partial BIC

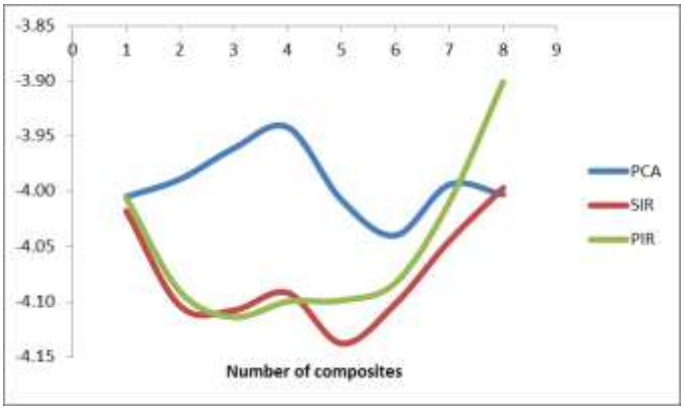
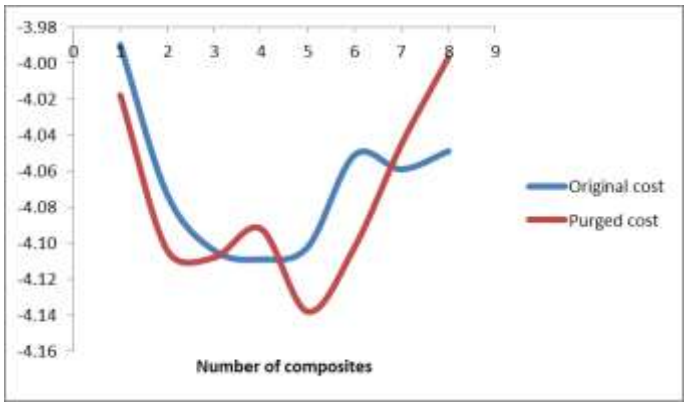


Figure 2. Partial BIC with actual and purged cost

(a) SIR composites



(b) PIR composites

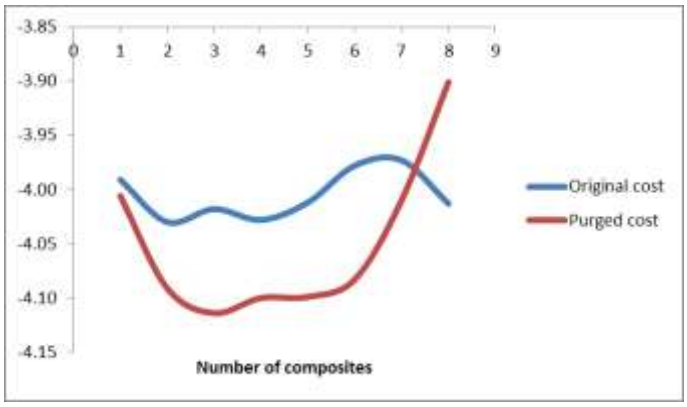


Figure 3. Histograms and Kernel density plots of estimated environmental cost differences using SIR composites.

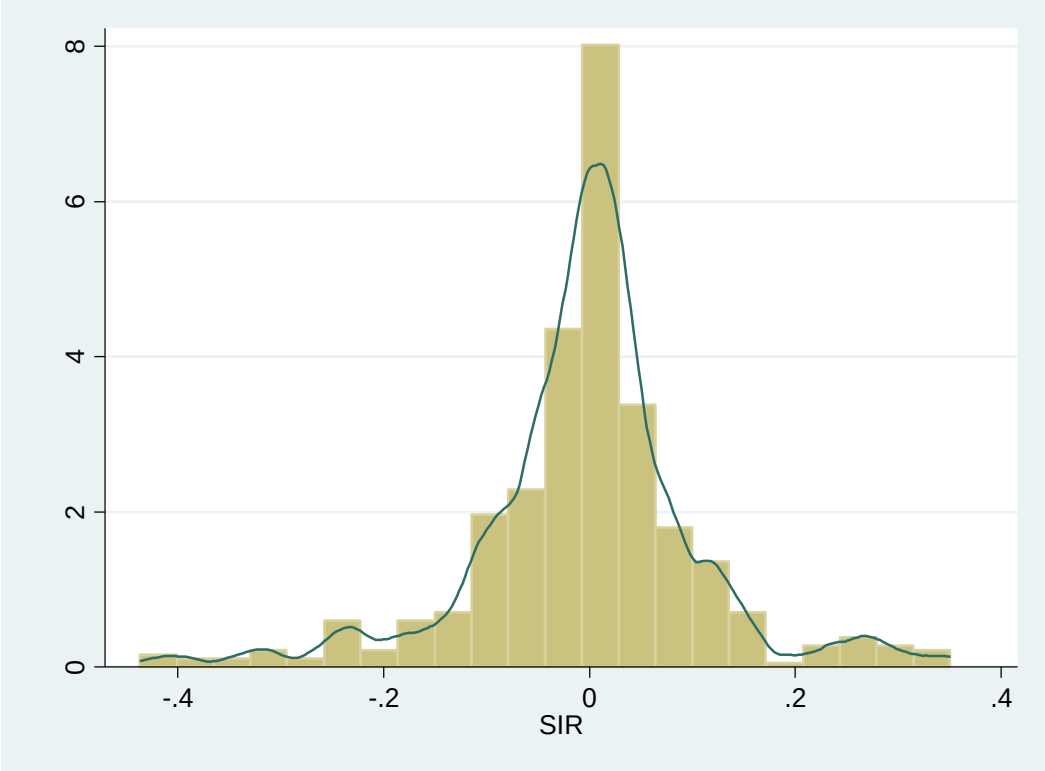


Figure 4. Estimated environmental cost differences using SIR and PCA models

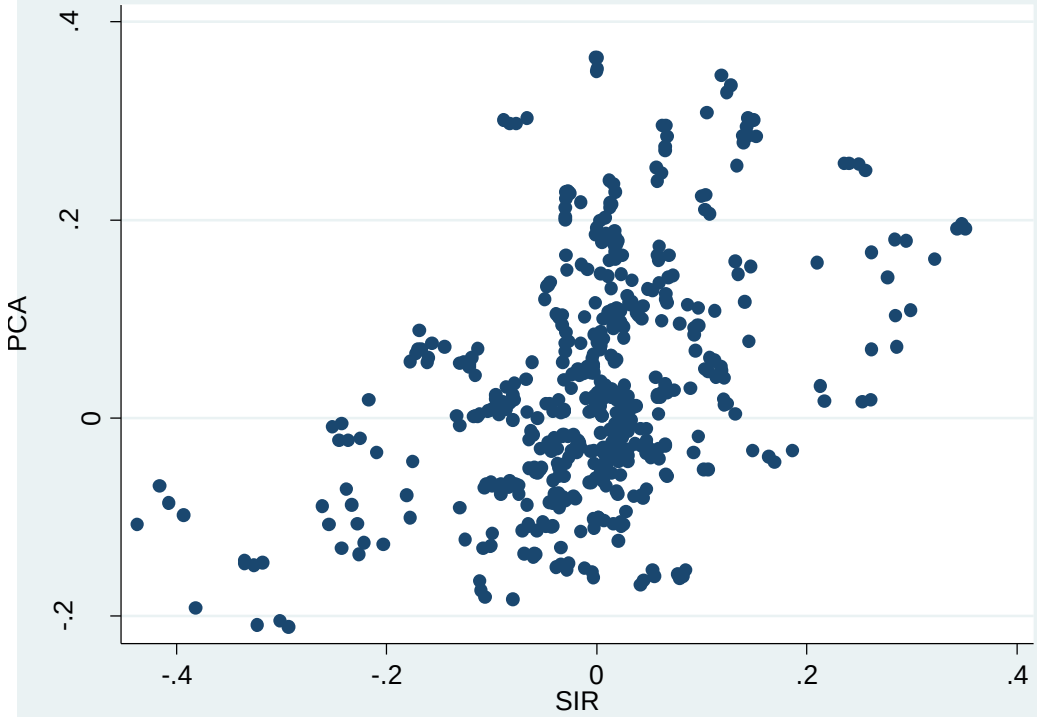


Figure 5. Box plots of % of estimated environmental cost differences.

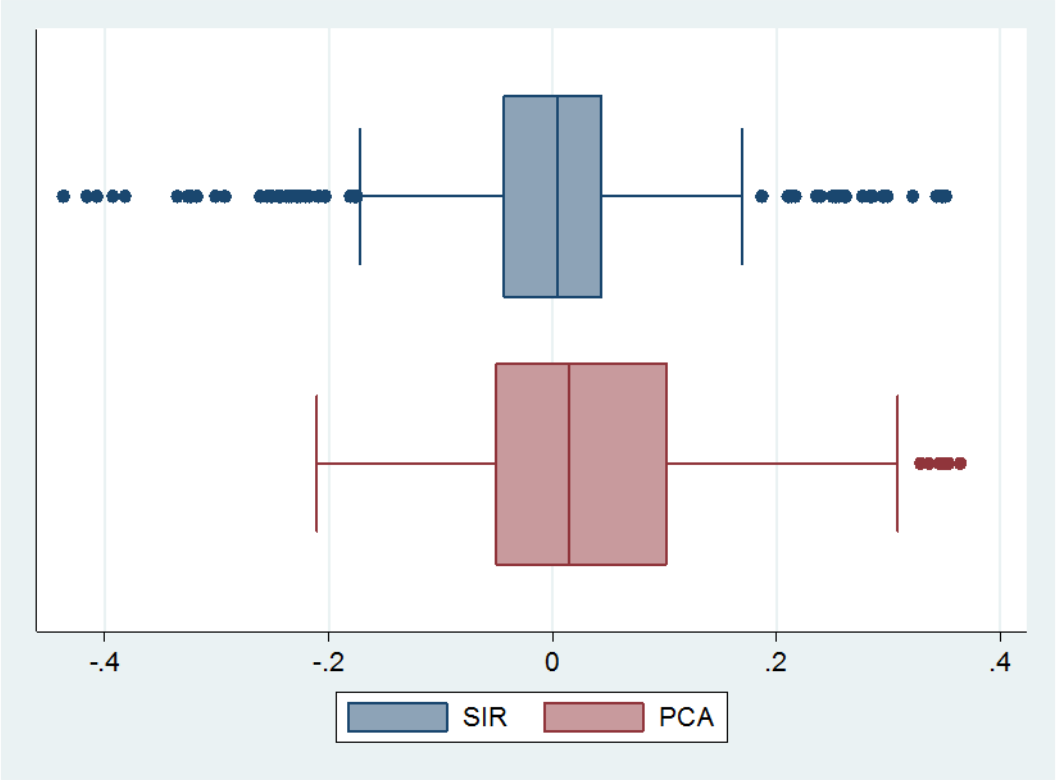


Table 1. Summary statistics of the economic data

Variable	Obs	Mean	Std. Dev.	Min	Max
ln(SOTEX)	512	10.4190	1.1261	8.5261	14.284
ln(CU)	512	0	1.2605	-2.0206	4.3041
ln(ED)	512	0	1.3575	-2.3100	4.5338
ln(NL)	512	0	1.0882	-2.1162	3.2718
HV	512	0	7.9380	-14.448	24.922
ln(TR)	512	0	1.1008	-2.1887	3.4337
ln(KP)	512	0	0.1027	-0.1662	0.0951
ln(LP)	512	0	0.0550	-0.0858	0.0626

Table 2. Model selection Vuong tests

a) Null Hypothesis: SIR=PCA

No Composites	Value	Model Accepted
1	0.688	SIR
2	4.414	SIR***
3	4.18	SIR***
4	3.546	SIR***
5	2.725	SIR***
6	1.45	SIR*
7	1.078	SIR
8	-0.12	PCA

b) Null Hypothesis: PIR=PCA

No Composites	Value	Model Accepted
1	0.298	PIR
2	3.933	PIR***
3	4.129	PIR***
4	3.534	PIR***
5	2.085	PIR***
6	0.917	PIR
7	0.333	PIR
8	-2.035	PCA**

c) Null Hypothesis: SIR=PIR

No Composites	Value	Model Accepted
1	0.553	SIR
2	1.724	SIR**
3	-0.324	PIR
4	-0.294	PIR
5	1.149	SIR
6	0.457	SIR
7	0.825	SIR
8	2.385	SIR***

Notes: The Vuong test statistic is distributed as a standard normal variable, N(0,1). * significant at 15%, ** significant at 10%, and *** significant at 5%. Critical values: 1.44 at 15%, 1.64 at 10%, and 1.96 at 5%.

Table 3. OLS parameter estimates of the cost function.

	No composites		SIR composites		PCA composites	
	Coef.	t-ratio	Coef.	t-ratio	Coef.	t-ratio
Intercept	10.3749	828.2	10.3449	678.3	10.3424	533.2
ln(CU)	0.5532	10.25	0.4832	8.86	0.5735	10.30
ln(ED)	0.0751	2.05	0.1262	3.57	0.0476	1.18
ln(NL)	0.1211	2.79	0.0078	0.16	-0.0569	-1.02
HV	0.0111	9.17	0.0069	4.56	0.0032	1.87
ln(TR)	0.1802	3.75	0.3200	5.52	0.2971	4.80
½·ln(CU) ^{2}	0.3182	0.80	-0.0528	-0.13	0.2708	0.65
½·ln(ED) ^{2}	0.5014	3.21	0.9494	6.35	0.4006	2.62
½·ln(NL) ^{2}	0.0114	0.03	-0.5087	-1.29	0.2175	0.54
½·HV^2	-0.0002	-0.93	-0.0003	-1.61	-0.0003	-1.50
½·ln(TR) ^{2}	0.8661	1.89	0.0346	0.07	0.0873	0.19
ln(CU)·ln(ED)	-0.3831	-1.69	-0.4389	-1.98	-0.4409	-1.98
ln(CU)·ln(NL)	0.4286	1.80	0.1924	0.83	0.2711	1.12
ln(CU)·HV	-0.0087	-1.16	-0.0392	-5.15	-0.0174	-2.09
ln(CU)·ln(TR)	-0.2735	-0.92	0.4922	1.61	0.1064	0.31
ln(ED)·ln(NL)	-0.0418	-0.25	0.0323	0.21	-0.1395	-0.87
ln(ED)·HV	0.0029	0.62	0.0119	2.58	-0.0069	-1.28
ln(ED)·ln(TR)	-0.1571	-0.73	-0.6952	-3.36	0.0952	0.44
ln(NL)*HV	-0.0009	-0.15	0.0127	2.14	0.0132	2.01
ln(NL)*ln(TR)	-0.4054	-1.03	0.2252	0.58	-0.3722	-0.97
HV·ln(TR)	0.0047	0.67	0.0135	1.90	0.0121	1.64
ln(KP)	0.0912	0.98	0.0901	1.15	0.0895	1.07
ln(LP)	0.3824	2.17	0.3983	2.67	0.4204	2.63
Composites	No		Yes		Yes	
R^2	0.985		0.990		0.988	
AIC	-3.904		-4.214		-4.083	
BIC	-3.713		-3.858		-3.727	

Table 4. ML estimates of the parameters describing the structure of the error term.

	No composites		SIR composites		PCA composites	
	Coef.	t-ratio	Coef.	t-ratio	Coef.	t-ratio
$\ln(\sigma_v)$	-2.1879	-22.38	-2.2252	-16.74	-2.1266	-2.29
$\ln(\sigma_u)$	-2.0638	-9.26	-3.0431	-1.64	-4.7849	-0.01
σ_v	0.1122		0.1080		0.1192	
σ_u	0.1270		0.0477		0.0084	
$\gamma = \sigma_u^2/(\sigma_v^2 + \sigma_u^2)$	0.5617		0.1630		0.0049	
log-likelihood	297.119		395.297		361.849	

Table 5. Descriptive statistics of annual efficiency scores.

Composites	Statistic	2001	2002	2003	2004
None	Mean	90.42	90.39	90.50	90.52
	St.dev.	4.23	4.08	3.77	3.70
	Max	95.82	96.72	96.43	96.84
	Min	75.56	75.27	78.44	76.55
SIR	Mean	96.26	96.27	96.27	96.28
	St.dev.	0.76	0.73	0.72	0.68
	Max	97.52	97.89	97.79	97.96
	Min	93.36	93.26	93.89	94.07
PCA	Mean	99.34	99.34	99.34	99.34
	St.dev.	0.02	0.02	0.02	0.02
	Max	99.38	99.40	99.39	99.40
	Min	99.28	99.28	99.28	99.29

ANNEX 1

Table A. Parameter estimates of the composite variables.

	SIR composites		PCA composites	
	Coef.	t-ratio	Coef.	t-ratio
c1	0.0424	3.01	-0.0015	-0.18
c2	0.0223	2.29	0.0308	3.53
c3	0.0552	3.27	0.0406	5.19
c4	0.0076	0.55	-0.0124	-1.85
c5	0.0276	2.61	0.0015	0.61
c1^2	-0.0584	-2.45	0.0077	2.59
c2^2	0.0260	1.58	0.0057	2.34
c3^2	-0.0657	-2.89	-0.0069	-3.91
c4^2	-0.0276	-1.51	-0.0029	-1.57
c5^2	0.0233	3.43	0.0053	4.64
c1*c2	-0.0317	-2.72	0.0068	3.30
c1*c3	-0.0982	-5.35	-0.0072	-4.21
c1*c4	0.0409	2.15	0.0064	5.85
c1*c5	-0.0145	-2.27	0.0054	4.46
c2*c3	-0.0200	-1.40	-0.0040	-2.56
c2*c4	0.0092	0.72	-0.0007	-0.44
c2*c5	0.0110	1.61	-0.0049	-4.03
c3*c4	0.0548	3.66	0.0007	0.62
c3*c5	-0.0483	-5.83	0.0016	1.73
c4*c5	0.0506	7.62	0.0005	0.71