WORKING PAPERS

N° 1640

Mai 2025

"Reconciling Engineers and Economists: the Case of a Cost Function for the Distribution of Gas"

Frédérique Fèbe, Jean-Pierre Florens et Léopold Simar

Toulouse School of Economics

Reconciling Engineers and Economists: the Case of a Cost Function for the Distribution of Gas

Frédérique FÈVE* Jean-Pierre FLORENS*
Léopold SIMAR*,§

May 2025

Abstract

The analysis of cost functions is an important topic in econometrics both for scientific studies and for industrial applications. The object of interest may be the cost of a firm or the cost of a specific production, in particular in case of a proposal to a procurement. Engineer methods evaluate the technical cost given the main characteristics of the output using the decomposition of the production process in elementary tasks and are based on physical laws. The error terms in these models may be viewed as idiosyncratic chocs. The economist usually observes ex post the cost and the characteristics of the product. The difference between theoretical cost and the observed one may be modeled by the inefficiency of the production process. In this case, econometric models are cost frontier models. In this paper we propose to take advantage of the situation where we have information from both approaches. We consider a system of two equations, one being a standard regression model (for the technical cost function) and one being a stochastic frontier model for the economic cost function where inefficiencies are explicitly introduced. We derive estimators of this joint model and derive its asymptotic properties. The models are presented in classical parametric approach, with few assumptions on the stochastic properties of the joint error terms. We suggest also a way to extend the model to a nonparametric approach, the latter provides an original way to model and estimate nonparametric stochastic frontier models. The techniques are illustrated in the case of the cost function for the distribution of gas in France.[1]

Key Words: Cost efficiency, Stochastic frontier models, Location-Scale efficiencies, Nonparametric stochastic frontier models.

JEL Classification: C10, C14, C51, D22.

*Toulouse School of Economics (TSE), Université Toulouse Capitole, Toulouse, France, frederique.feve@tse-fr.eu, and jean-pierre.florens@tse-fr.eu. F. Fève and J.P. Florens acknowledge funding from the French National Research Agency (ANR) under the Investments for the Future (Investissement d'Avenir), grant ANR-17-EURE-0010.

§Institut de Statistique, Biostatistique et Sciences Actuarielles (ISBA), LIDAM, UCLouvain, Louvain-la-Neuve, Belgium, leopold.simar@uclouvain.be

[1]The authors acknowledge Gaz Réseau Distribution de France (GRDF) for providing the anonymized data that illustrate our methodology. These data are available. The authors have no relevant financial or non-financial interests or other competing interests to disclose. The authors are solely responsible for the conclusions of the analysis.

1 Introduction and the Basic Model

Economic theory defines the cost function for a unit of production as the relationship between the cost of production and the level of production, considering the prices of the production factors. The level of production can be described using various variables. For example, we can consider the distribution of gas, electricity, mail services, or manufacturing telecommunications satellites or aircraft engines, etc. It is not a macroeconomic cost neither a total cost for the firm. The dimension of the product is described by a small number of variables X (quantity of gas consumed and length of the network, quantity of mail distributed and housing density, number of channels for transmission and weight of a satellite, power of the aircraft engine, etc.).

In order to empirically use the prices of production factors (labor costs, capital costs, etc.), we require samples in which these factors exhibit variation. This is often not the case in cross-sectional data (for example, in France, these quantities do not vary across the country). Therefore, we will only consider variables that explain the cost factors related to the product. We will represent this function as $\varphi(X)$, where X is the set of observable variables that characterize the production.

The analysis of production and cost functions is an important topic in econometrics both for scientific studies and for industrial applications. The object of interest may be the cost of a firm or the cost of a specific production, in particular in case of a proposal to a procurement. The requirement of a first evaluation of a production cost has motivated various approaches. The so called "engineer methods", or "normative methods", evaluate the cost given the main characteristics of the output using the decomposition of the production process in elementary tasks and are based on physical laws. The error terms in these models may be viewed as idiosyncratic chocs and the relation between the cost and the description of the output may be considered as a conditional expectation.

The economist usually observes ex post the cost and the characteristics of the product. The difference between theoretical cost and the observed one may be modeled by the inefficiency of production process. Then econometric models are cost frontier models, stochastic frontier analysis for example.

Even if the formalization of a joint model mixing engineer and economist approach seems original, the economists have been concerned by the importance of engineer approach in cost analysis, as described in e.g. Chenery (1949), Marsden et al. (1974), Wibe (1984) and Massol (2011).

- The engineering measure is obtained by aggregating theoretical costs necessary at each step of the production process. These step-wise costs are calculated using software tools

that take precise descriptions of the technical aspects of each step as inputs. These software programs are often fine-tuned or calibrated based on databases associated with the production sector and then adjusted according to the subjective knowledge of the engineers. It can be considered that this measure of "Engineering Cost" or "Technical Cost" (referred to as C_T) is closely related to the cost function $\varphi(X)$ with a random error having a zero mean. We express this as:

$$C_T = \varphi(X) + u, \text{ where } \mathbb{E}(u|X) = 0. \tag{1.1}$$

Additionally, it is commonly assumed that $u|X$ follows a normal distribution with mean 0 and variance σ_u^2. Note that below we will not need to assume the normality. It is important to note that C_T is a theoretical cost derived through a specific technique and is not directly observed. The error term u can be attributed to inaccuracies in the calculations made at each step of the process and is also influenced by the particular choice of the function $\varphi(\cdot)$. Note that the technical costs are generally expressed in technical units (like kcal, kwh, etc.).

- The economic measure is derived from the accounting process of the firm. This cost is observed and includes various factors that were not taken into account in the previous approach. It is reasonable to assume that to within some random error, this "Economic Cost" (referred to as C_E) will be greater than C_T, the engineering cost, and we define this difference as inefficiency. Of course the economic costs are evaluated in monetary units which are often different from the units for C_T. To solve this problem we normalize the two costs by their standard deviations: the two costs are "unit free". This does not affect the values of the elasticities that we derive below. Therefore, we can express the economic cost as a typical stochastic frontier model, common in Stochastic Frontier Analysis (SFA):

$$C_E = \varphi(X) + v + \eta, \text{ where } \mathbb{E}(v|X) = 0 \text{ and } \eta \geq 0, \tag{1.2}$$

with v representing the random error (here again we will not need the normality assumption) and $\eta \geq 0$ quantifying the inefficiency. It is usually assumed that $\eta|X$ is independent of the noise v (see e.g., Kumbhakar and Lovell, 2000). In our paper, we will consider for $\eta|X$ a quite flexible location-scale model defined as

$$\eta = \mu_\eta(X) + \sigma_\eta(X)\xi, \tag{1.3}$$

where $\mathbb{E}(\xi|X) = 0$ and $\mathbb{V}(\xi|X) = 1$ and the functions $\mu_\eta(\cdot)$ and $\sigma_\eta(\cdot)$ take only positive values. So we have

$$\mathbb{E}(\eta|X) = \mu_\eta(X), \text{ and } \mathbb{V}(\eta|X) = \sigma_\eta^2(X). \tag{1.4}$$

Remark 1.1. *Note that the location-scale specification for η is much more flexible than the popular specification in the literature on SFA where most of the approaches assume that $\eta|X \sim D_+(\sigma_\eta(X))$ where D_+ is a distribution over positive real numbers belonging to some one-parameter scale family with parameter $\sigma_\eta(X)$ that may be dependent on X.*[1] *In these models, it is implicit that $\mathbb{E}(\eta|X) = C\sqrt{\mathbb{V}(\eta|X)}$ for some constant C depending on the chosen family, imposing a strong restriction on the first two conditional moments of the efficiency term η.*

From an econometric point of view, this leads us to consider a system of two equations

$$C_T = \varphi(X) + u \tag{1.5}$$

$$C_E = \varphi(X) + v + \eta, \tag{1.6}$$

where we assume

$$\left(\begin{array}{c} u \\ v \end{array} \middle| X \right) \sim D_2 \left(\left(\begin{array}{c} 0 \\ 0 \end{array} \right), \left(\begin{array}{cc} \sigma_u^2 & \sigma_{uv} \\ \sigma_{uv} & \sigma_v^2 \end{array} \right) \right), \tag{1.7}$$

where D_2 is any bivariate distribution with mean zero and covariance matrix described in (1.7) and η follows the location-scale model defined in (1.3). So we expect some dependence between u and v but both error terms are independent of X and of $\eta|X$.

We can select some flexible parametric models for the functions $\varphi(X)$, $\mu_\eta(X)$ and $\sigma_\eta(X)$. It is important to note that since we will use least squares approaches for the estimation, the normality assumption in (1.7) is not necessary but the independence assumptions are required.

Additionally, it should be noted that without loss of generality, we could also include in the functions $\mu_\eta(X, Z)$ and $\sigma_\eta(X, Z)$ external or environmental factors Z, that might influence the inefficiencies and that are not present in the engineering model identifying the technical cost C_T, and thus Z is not present in the function $\varphi(X)$. Here we assume that both error terms u and v are independent of (X, Z) and of $\eta|X, Z$.

The paper is organized as follows. In the next section we introduce the basic model in a parametric setup and we derive its properties. We describe how a Non Linear Generalized Least-Squares (NLGLS) method allows for consistent estimation of parametric specifications for $\varphi(X)$, $\mu_\eta(X, Z)$ and $\sigma_\eta(X, Z)$. In Section 2.3 we apply the methodology with real data from the distribution of gas in France. Section 3 presents our nonparametric extension showing the identifiability of this general model, even in the case of single nonparametric stochastic cost function. In Section 3.2 we show how this approach is easy to implement in practice. Section 4 concludes summarizing the main contributions of our paper.

[1] A density $f(\cdot)$ belongs to a one-parameter scale family if it can be written as $f(\cdot) = (1/\sigma)\tilde{f}(\cdot/\sigma)$ for some $\sigma > 0$, where $\tilde{f}(\cdot)$ is any density on $\mathbb{R}_+$. Popular examples include the Half-Normal and Exponential distributions, but also the Gamma and Weibull distributions with fixed shape parameters.

2 A NLGLS Method for Parametric Specifications

2.1 The model and its estimation

Suppose we have chosen parametric specifications for the functions of interest, so that we write $\varphi(X) = \varphi(X;\beta)$, $\mu_\eta(W) = \mu_\eta(W;\gamma_\mu)$ and $\sigma_\eta(W) = \sigma_\eta(W;\gamma_\sigma)$ where $W = (X, Z)$ and the functions are known up to the parameters $(\beta, \gamma_\mu, \gamma_\sigma)$.

Due to the location-scale model for η in (1.3) and defining $\varepsilon = \eta - \mathbb{E}(\eta|W)$, the equation for the economic cost in (1.6) can be rewritten as

$$\begin{aligned} C_E &= \varphi(X;\beta) + \mu_\eta(W) + v + \varepsilon, \\ &= \varphi(X;\beta) + \mu_\eta(W) + v^* \end{aligned} \tag{2.1}$$

where $v^* = v + \varepsilon$, and due to the assumptions above we have

$$\mathbb{E}(v^*|X) = 0, \tag{2.2}$$

$$\mathbb{V}(v^*|W) = \mathbb{E}(v^{*2}|W) = \sigma^2_{v^*}(W) = \sigma^2_v + \sigma^2_\eta(W) \tag{2.3}$$

Due the independence between (u, v) and $\eta|W$, we have also $\sigma_{uv*} = \sigma_{uv}$.

The simultaneous model (1.5)–(1.6) can then be written as

$$\begin{pmatrix} C_T \\ C_E \end{pmatrix} = \begin{pmatrix} \varphi(X;\beta) \\ \varphi(X;\beta) + \mu_\eta(X,Z) \end{pmatrix} + \begin{pmatrix} u \\ v^* \end{pmatrix}, \tag{2.4}$$

where now

$$\mathbb{E}\left(\begin{pmatrix} u \\ v* \end{pmatrix} \middle| X, Z \right) = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \text{ and } \mathrm{Cov}(u, v^*|X, Z) = \Sigma_{X,Z} = \begin{pmatrix} \sigma^2_u & \sigma_{uv} \\ \sigma_{uv} & \sigma^2_{v^*}(X,Z) \end{pmatrix}. \tag{2.5}$$

With the chosen parametric specifications, the parameters of the model are therefore defined as $\theta = (\beta, \gamma_\mu)$ and $\lambda = (\sigma_u, \sigma_v, \sigma_{uv}, \gamma_\sigma)$, where the dimensions of $(\beta, \gamma_\mu, \gamma_\sigma)$ depend on the explanatory variables chosen in the respective models. Hence $\Sigma_{X,Z} = \Sigma_{X,Z,\lambda}$ and $\sigma^2_{v^*}(X, Z) = \sigma^2_{v^*}(X, Z, \sigma_v, \gamma_\sigma)$. For instance we might choose for the main part of the model $\varphi(X;\beta) = \beta' X$ and $\mu_\eta(W;\gamma_\mu) = \exp(\gamma'_\mu W)$ where the first component of X is 1 to allow for a constant in the linear parts of the model, but there is no constant in the vector Z.

Now we have a random sample of n units, providing observations $(C_{T,i}, C_{E,i}, X_i, Z_i)$ for $i = 1, \ldots, n$. Let $\boldsymbol{X}$ be the $n \times q$ matrix of selected regressors where the first column is $\boldsymbol{i}_n$ a n-vector of ones, and let $\boldsymbol{W} = (\boldsymbol{X}, \boldsymbol{Z})$ where there is no constants in the matrix $\boldsymbol{Z}$. Let also $\boldsymbol{C}$ being the n-vector of the costs C_i. Then we have the system of $2n$ equations

$$\begin{pmatrix} \boldsymbol{C}_T \\ \boldsymbol{C}_E \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}\beta \\ \boldsymbol{X}\beta + \exp(\boldsymbol{W}\gamma_\mu) \end{pmatrix} + \begin{pmatrix} \boldsymbol{u} \\ \boldsymbol{v}^* \end{pmatrix}, \tag{2.6}$$

where

$$\mathbb{E}\left(\begin{array}{c}\boldsymbol{u}\\ \boldsymbol{v}*\end{array}\middle|\boldsymbol{X},\boldsymbol{Z}\right)=\left(\begin{array}{c}\mathbf{0}_n\\ \mathbf{0}_n\end{array}\right) \text{ and } \mathrm{Cov}(\boldsymbol{u},\boldsymbol{v}^*|\boldsymbol{X},\boldsymbol{Z})=\boldsymbol{\Omega}_{X,Z,\lambda}=\left(\begin{array}{cc}\sigma_u^2\boldsymbol{I}_n & \sigma_{uv}\boldsymbol{I}_n\\ \sigma_{uv}\boldsymbol{I}_n & \boldsymbol{D}_{X,Z}\end{array}\right), \quad (2.7)$$

with $\boldsymbol{D}_{X,Z}=\mathrm{diag}(\boldsymbol{d}_{X,Z})$, where $\boldsymbol{d}_{X,Z}$ is a n-vector with element $d_{X_i,Z_i}=\sigma^2_{v^*}(X_i,Z_i,\sigma_v,\gamma_\sigma)$.

If $\boldsymbol{\Omega}_{X,Z,\lambda}$ were known, the NLGLS estimator of θ would be given by solving

$$\widehat{\theta}_n=\arg\min_\theta\left\{(\boldsymbol{u}' \quad \boldsymbol{v}^{*'})\boldsymbol{\Omega}^{-1}_{X,Z,\lambda}\left(\begin{array}{c}\boldsymbol{u}\\ \boldsymbol{v}^*\end{array}\right)\right\}, \quad (2.8)$$

where by (2.6), $\boldsymbol{u}=\boldsymbol{C}_T-\boldsymbol{X}\beta$ and $\boldsymbol{v}^*=\boldsymbol{C}_E-\boldsymbol{X}\beta-\exp(\boldsymbol{W}\gamma_\mu)$. The computations can be very fast when noting that, due to the bloc-diagonal structure of $\boldsymbol{\Omega}_{X,Z,\lambda}$, we can avoid the inversion of this $(2n\times 2n)$ matrix. Indeed we have after simple algebraic operations (by using properties of partitioned matrices, see e.g. Härdle and Simar, 2019):

$$\boldsymbol{\Omega}^{-1}_{X,Z,\lambda}=\left(\begin{array}{cc}\mathrm{diag}(\boldsymbol{a}_{X,Z}) & \mathrm{diag}(\boldsymbol{b}_{X,Z})\\ \mathrm{diag}(\boldsymbol{b}_{X,Z}) & \mathrm{diag}(\boldsymbol{c}_{X,Z})\end{array}\right), \quad (2.9)$$

where $\boldsymbol{a}_{X,Z},\boldsymbol{b}_{X,Z},\boldsymbol{c}_{X,Z}$ are n-vectors with elements for $i=1,\ldots,n$, respectively given by

$$a_{X_i,Z_i}=d_{X_i,Z_i}/(\sigma_u^2 d_{X_i,Z_i}-\sigma_{uv}^2), b_{X_i,Z_i}=-\sigma_{uv}/(\sigma_u^2 d_{X_i,Z_i}-\sigma_{uv}^2), \text{ and } c_{X_i,Z_i}=\sigma_u^2/(\sigma_u^2 d_{X_i,Z_i}-\sigma_{uv}^2). \quad (2.10)$$

Since $\boldsymbol{\Omega}_{X,Z,\lambda}$ depends on λ which is unknown, we may use an iterative method requiring to specify some parametric model for $\sigma^2_{v^*}(X_i,Z_i,\sigma_v,\gamma_\sigma)$, providing consistent estimates of λ. In the next section, we show that the solution $\widehat{\theta}_n$ of (2.8) benefits, under mild regularity conditions, form the properties of usual estimators of parametric models, i.e. $\sqrt{n}$-consistency and asymptotic normal distribution. In practice for inference about the values of θ, we use bootstrap techniques. We will illustrate this in the application below.

Practical details of the iterations

The iterations to solve (2.8) can be done as follows: at step $k=0$ we select as initial values for the covariance matrix $\boldsymbol{\Omega}^{(0)}_{X,Z,\lambda}=\boldsymbol{I}_{2n}$, the order-$2n$ identity matrix, then at step $k=1$ we compute $\alpha^{(k)},\beta^{(k)},\gamma_\mu^{(k)}$ solving

$$\arg\min_{\alpha,\beta,\gamma_\mu}\left\{(\boldsymbol{u}' \quad \boldsymbol{v}^{*'})(\boldsymbol{\Omega}^{(k-1)}_{X,Z,\lambda})^{-1}\left(\begin{array}{c}\boldsymbol{u}\\ \boldsymbol{v}^*\end{array}\right)\right\}, \quad (2.11)$$

which provides the n pairs of residuals $(\widehat{\boldsymbol{u}}^{(k)},\widehat{\boldsymbol{v}}^{*,(k)})$. The empirical variance of the n residuals $\widehat{\boldsymbol{u}}^{(k)}$ provides the new estimate $\widehat{\sigma}_u^{2,(k)}$ and the empirical covariance between $(\widehat{\boldsymbol{u}}^{(k)},\widehat{\boldsymbol{v}}^{*,(k)})$ provides $\widehat{\sigma}_{uv}^{(k)}$.

Since $d_{X,Z} = \mathbb{E}(v^{*2}|X,Z)$, the value of $d_{X,Z}$ is the conditional nonlinear regression of v^{*2} on (X,Z). As v^{*2} is not directly observed we regress its estimators $v^{*2,k}$ on (X,Z) by some selected parametric model, say

$$v^{*2,k} = \sigma^2_{v^*}(X,Z,\delta) + \zeta, \tag{2.12}$$

with $\mathbb{E}(\zeta|X,Z) = 0$, providing by (nonlinear) least squares a consistent estimator of δ and the fitted values $\widehat{\widehat{\boldsymbol{v}}}^{*2,(k)} = \sigma^2_{v^*}(\boldsymbol{X},\boldsymbol{Z},\widehat{\delta})$ (see White 1980). All of these provide the new $\boldsymbol{\Omega}^{(k)}_{X,Z,\lambda}$ by (2.7).

We redefine $k = k+1$ and iterate the process till convergence of the solutions to obtain $\widehat{\beta}_n, \widehat{\gamma}_{\mu,n}$. In our application below, we achieve convergence after a few iterations.

Regularization of the computation of the NLGLS in (2.8)

The optimal weighting matrix in equation (2.8) is $\boldsymbol{\Omega}^{-1}_{X,Z,\lambda}$. However, when the size of $\boldsymbol{\Omega}_{X,Z,\lambda}$ is large, the estimated matrix $\widehat{\boldsymbol{\Omega}}_{X,Z,\lambda}$ may be not invertible. Actually the smallest eigenvalues of $\boldsymbol{\Omega}_{X,Z,\lambda}$ will decline to zero when the dimension increases and the estimated values of these eigen values may be extremely close to zero. In that case, the inversion of $\boldsymbol{\Omega}_{X,Z,\lambda}$ should be regularized. A possible regularization is the Tikhonov regularization where $\boldsymbol{\Omega}_{X,Z,\lambda}$ is replaced by $\omega I_{2n} + \boldsymbol{\Omega}_{X,Z,\lambda}$, where the eigenvalues are bounded by $\omega > 0$. In our particular case, the regularization only occurs in the South West block $\boldsymbol{D}_{X,Z}$ and the regularization may be $\left[\omega \begin{pmatrix} \mathbf{0}_{n,n} & \mathbf{0}_{n,n} \\ \mathbf{0}_{n,n} & I_n \end{pmatrix} + \boldsymbol{\Omega}_{X,Z,\lambda}\right]$. For the Tikhonov regularization of the weighting matrix in GMM see Carasco and Florens (2000).

Remark 2.1. *Following the Remark 1.1, the model specification and the estimation procedure could easily be adapted the case of a simple SFA model with a single equation like* (1.2)–(1.3) *for modeling e.g. CE alone. Its estimation only requires the nonlinear least squares described above.*

2.2 Asymptotic properties

The model we describe above in (2.4) is a particular case of the general multivariate, nonlinear, heteroskedastic regression model defined by

$$Y = \phi(W;\theta) + U, \text{ with } \mathbb{E}(U|W) = 0 \text{ and } \mathbb{V}(U|W) = \Sigma_{W,\lambda}, \tag{2.13}$$

where $Y \in \mathbb{R}^g$, $X \in \mathbb{R}^{q_x}$, $W \in \mathbb{R}^{q_w}$, $\theta \in \mathbb{R}^p$ and $\lambda \in \mathbb{R}^\ell$, and where θ and λ are independent parameters (see e.g. Cameron and Trivedi, 2005, Section 6.10.3). In our case, $W = (X,Z)$ and $g = 2$ since $Y = (C_T\ C_E)'$, $U = (u\ v^*)'$ and $\Sigma_{W,\lambda}$ was introduced in (2.5). Note that $\theta =$

(β, γ_μ) and $\lambda = (\sigma_u, \sigma_v, \sigma_{uv}, \gamma_\sigma)$. The parameter θ is locally identified if rank $\left\{\mathbb{E}\left(\frac{\partial\phi'}{\partial\theta}\frac{\partial\phi}{\partial\theta'}\right)\right\} = p$ and λ is locally identified if $\Sigma_{X,Z,\lambda}$ is a one to one function of λ.

In our particular case, with the chosen models leading to (2.6), β is identified by the first equation in (2.4) if rank $\{\mathbb{E}(XX')\} = q_x$ (where we may have a constant in X). Then γ_μ is locally identified by the second equation in (2.4), if rank $\{\mathbb{E}(WW'\exp(2\gamma_\mu))\} = q_w$, i.e. if rank $\{\mathbb{E}(WW')\} = q_w$.

If $\lambda = (\sigma_u, \sigma_v, \sigma_{uv}, \gamma_\sigma)$ is given, the estimation of $\theta = (\alpha, \beta, \gamma_\mu)$ is obtained by

$$\widehat{\theta} = \arg\min_{\theta} \sum_{i=1}^{n} U_i' \Sigma_{X,Z,,\lambda}^{-1} U_i, \tag{2.14}$$

or by solving in θ the system of first order equations $\sum_{i=1}^{n} U_i' \Sigma_{X,Z,\lambda}^{-1} \frac{\partial U_i}{\partial\theta'} = 0$. The estimator $\widehat{\theta}$ is almost surely convergent to θ and

$$\sqrt{n}\left(\widehat{\theta} - \theta\right) \xrightarrow{\mathcal{L}} N_p\left(0, \left[\mathbb{E}\left(\frac{\partial\phi'}{\partial\theta}\Sigma_{X,Z,\lambda}^{-1}\frac{\partial\phi}{\partial\theta'}\right)\right]^{-1}\right). \tag{2.15}$$

If λ is unknown we can use an iterative method as described above to get consistent estimates of λ in an appropriate model for $\Sigma_{X,Z,\lambda}$ (see e.g. the model in (2.12)). The asymptotic properties of the estimator of θ remain identical as if λ was known.

The asymptotic normality of the estimator justifies the use of the multinomial bootstrap for the evaluation of confidence intervals (see e.g. Mammen, 1992).

2.3 Application to the distribution of gas in France

We show how our method works with a dataset coming from the company "Gaz Réseau de Distribution de France" (GRDF). We have information on 1218 distribution units for the year 2019. In this illustrative example, we will consider a cost variable C (the annual cost of the piping network) and two outputs X_1 and X_2 being the length of the piping network (in kms) and the gas consumption by the customers (in MWh/year). The managers at GRDF consider the former (X_1) as a cost driver which induces the cost and not as an input. For the cost we have the values of the "Technical cost" (C_T) prepared by the engineers and the observed "Economic cost" (C_E).

There are some outliers in the data, so we use robust techniques to eliminate these extreme points. We use the boxplots approach and eliminate the data points having a value for any of the variables outside the whiskers of the boxes, as described, e.g. in Section 1.1 of Härdle and Simar (2019). This reduces the sample size to 1020 units. As explained above, to avoid different units for the two costs, we divide each cost by its standard deviation.

The scale of the two outputs X being quite different, we divide also each X by its standard deviation ($s_{X_1} = 40.58$ and $s_{X_2} = 51.14 \times 10^3$).

For the location model for the inefficiencies, we will investigate the effect of an external environmental factor Z which is a measure of the density of the network (it is the ratio of the number of customers by the area covered by the unit). Basic summary statistics of the resulting variables are displayed in Table 1.

Table 1: Summary statistics on the variables with $n = 1020$. Q_a is the $a\%$ quantile of the variable.

variable	min	Q_{25}	Q_{50}	Q_{75}	max
C_T	0.0460	0.1625	0.3625	0.9790	4.9525
C_E	0.0043	0.3926	0.7468	1.4655	5.6876
scaled X_1	0.0041	0.4418	0.8293	1.5923	5.2403
scaled X_2	0.0041	0.2408	0.5606	1.2553	5.2739
Z	0.0010	0.0888	0.1749	0.3252	1.7901

The correlation matrix between the explanatory variables is given in Table 2, where we note that Z is positively, but slightly, correlated with the two outputs and that, as expected, the two outputs are positively correlated. The left panel of Figure 1 gives some insight on the distribution of the variable Z in our sample. The right panel gives an idea of the range of the n observed pairs (X_{1i}, X_{2i}) where we see, as expected that there are few observations in the NW and in the SE parts of the figure.

Table 2: Correlation matrix between X_1, X_2 and Z.

	X_1	X_2	Z
X_1	1.0000	0.7628	0.3535
X_2	0.7628	1.0000	0.4474
Z	0.3535	0.4474	1.0000

For the cost function $\varphi(X;\beta) = \beta' X$ we select a flexible quadratic model with $X' = (1 \quad X_1 \quad X_2 \quad X_1^2 \quad X_2^2 \quad X_1X_2)$, so $q = 6$. The location model for the inefficiencies may be written as $\mu_\eta(X, Z; \gamma_\mu) = \exp(\gamma_\mu'(X \quad Z))$ where here the dimension of $\gamma_\mu = q + 1 = 7$. In Table 3 we give the resulting estimator of the parameters and their 95% confidence intervals (obtained by 1000 bootstrap replications, using the symmetric bootstrap as in Hall (1988)) are displayed in the following tables.

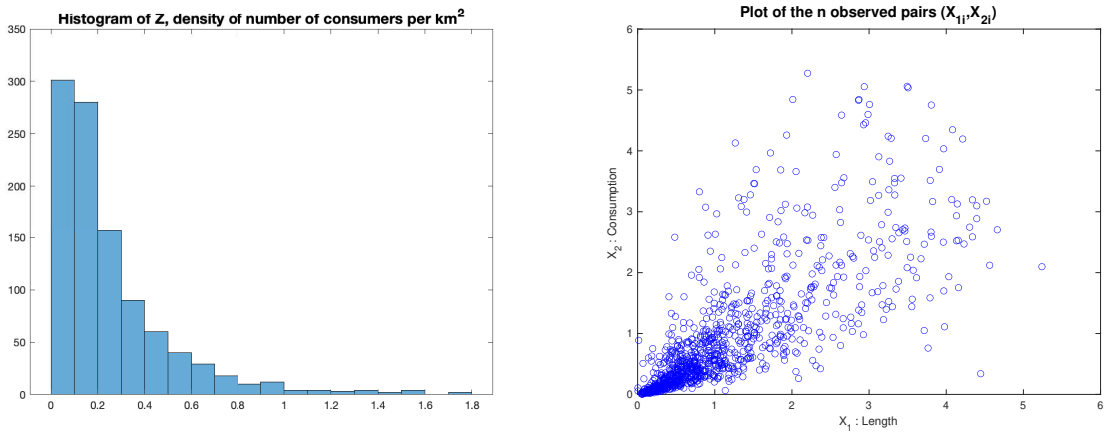

Figure 1: *Left panel: Histogram of the values of Z. Right panel: Plot of the n observed outputs (X_{1i}, X_{2i}).*

Table 3: Point estimates and 95% confidence intervals of θ, for the cost and inefficiency functions.

variable	$\widehat{\theta}_k$	lower bound	upper bound
Cost function $\varphi(\cdot)$			
Cst	-0.0195	-0.0243	-0.0147
X_1	0.2651	0.2529	0.2774
X_2	0.1231	0.1127	0.1336
X_1^2	0.0529	0.0405	0.0654
X_2^2	-0.0496	-0.0600	-0.0391
X_1X_2	0.1924	0.1699	0.2149
Expected Inefficiency $\mu_\eta(\cdot)$			
Cst	-2.1722	-2.3195	-2.0248
X_1	1.7311	1.3302	2.1319
X_2	0.3123	-0.1071	0.7317
X_1^2	-0.1866	-0.3882	0.0150
X_2^2	0.0583	-0.0276	0.1443
X_1X_2	-0.6264	-0.9582	-0.2946
Z	-0.1867	-0.5351	0.1617

We see that most of the selected variables are significant for the cost function and also for the expected inefficiencies model, where the effect of Z seems mostly negative but not in a fully significant way. Probably most of the variations of the inefficiencies are controlled by our flexible model for the X variables. This will be confirmed in Figure 5 below where we will see that the effect of Z is dominated by the effects of the X components.

We observe that the sign of the coefficients of squares of the Xs are both very close to zero, but are still have a significant sign (positive for X_1^2 and negative for X_2^2), but the interaction term is clearly positive, which complicates the analysis. This will be confirmed in Figure 2 below where the surface of the cost function is twisted and no clear convexity or concavity in the direction of X_1 or of X_2 is detected. The plot of the elasticities below in Figure 6 will give more insight on the possible scale economies.

As explained in equation (2.12), we select a simple model for the variance of v^*. We define

$$v^{*2} = (\delta_1 + \delta_2 Z)^2 + \zeta, \tag{2.16}$$

where $\mathbb{E}(\zeta|Z) = 0$. We tried several other specifications, a.o., to include or not the X variables in the model , but without introducing significant results and with almost no effect on the estimation of the cost functions and the location model for the inefficiencies. This simple model can also be written as

$$v^{*2} = \delta_1^2 + 2\delta_1\delta_2 Z + \delta_2^2 Z^2 + \zeta, \tag{2.17}$$

and the results are given in Table 4.

Table 4: Point estimates and 95% confidence intervals for the parameters of the variance function in (2.17).

parameter	Estimates	lower bound	upper bound
δ_1^2	0.0512	0.0337	0.0687
$2\delta_1\delta_2$	0.1157	0.0642	0.1671
δ_2^2	0.0653	-0.0182	0.1488

Finally to illustrate the results of our estimated model we provide some appealing figures. Figure 2 displays the data points and the fitted cost function for both the technical costs C_T and for the economic cost C_E. We see how well the technical costs C_T is fitted by our model and we also note that most of the data points for the economic cost C_E lies above the cost frontier, due to inefficiencies.

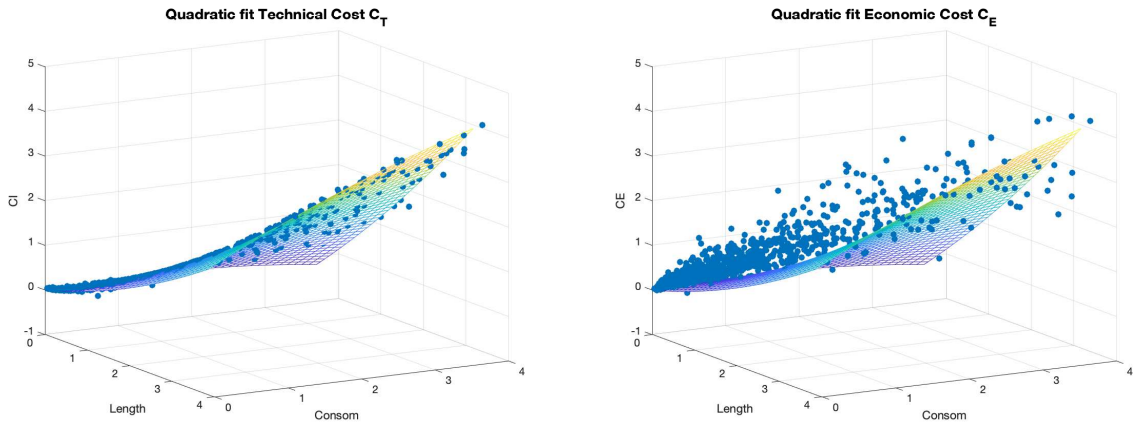

Figure 2: *Left panel, data points of the observed technical costs C_T and the fitted cost frontier $\varphi(X_i)$. Right panel, data points of the observed economic costs C_E and the same fitted cost frontier.*

Note that the global quality of parametric model can be appreciated through its R^2: we have for C_T, $R^2_{C_T} = 0.9991$ and for the model in C_E, we have $R^2_{C_E} = 0.9114$. This may also be viewed in Figures 3.

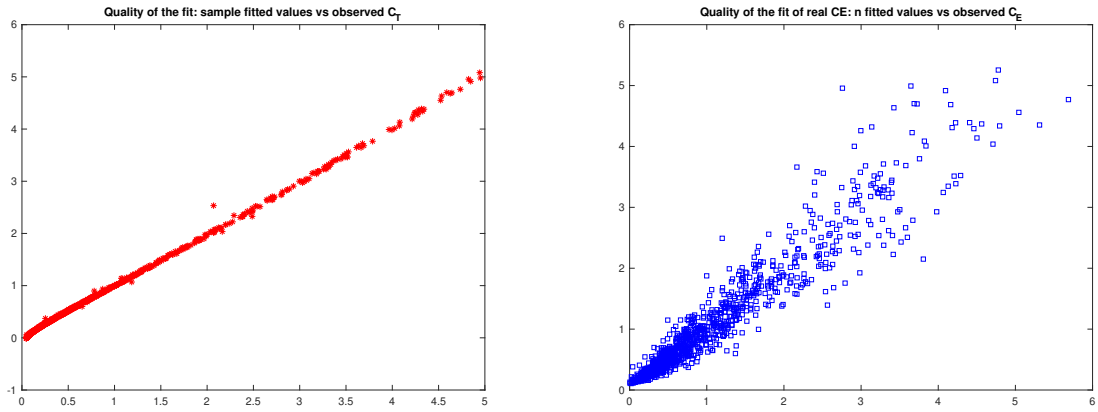

Figure 3: *Left panel, fitted values of technical costs $\widehat{C}_T = \widehat{\varphi}(X_i, \beta)$ versus observed C_T. Right panel, fitted values of economic costs $\widehat{C}_E = \widehat{\varphi}(X_i, \beta) + \widehat{\mu}_\eta(X_i, Z_i)$, versus observed C_E.*

Figure 4 shows the distribution of the expected inefficiencies $\widehat{\mu}_\eta(X_i, Z_i)$ obtained by our model for $i = 1, \ldots, n$.

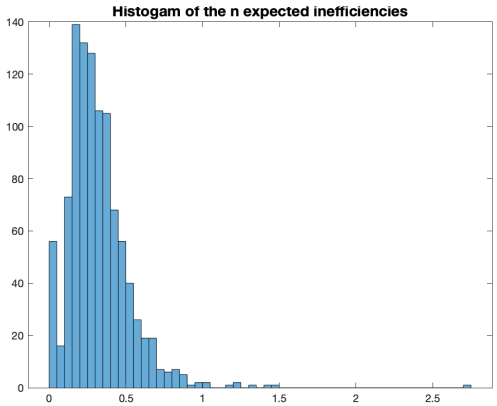

Figure 4: *Histogram of the n expected inefficiencies $\widehat{\mu}_\eta(X_i, Z_i)$.*

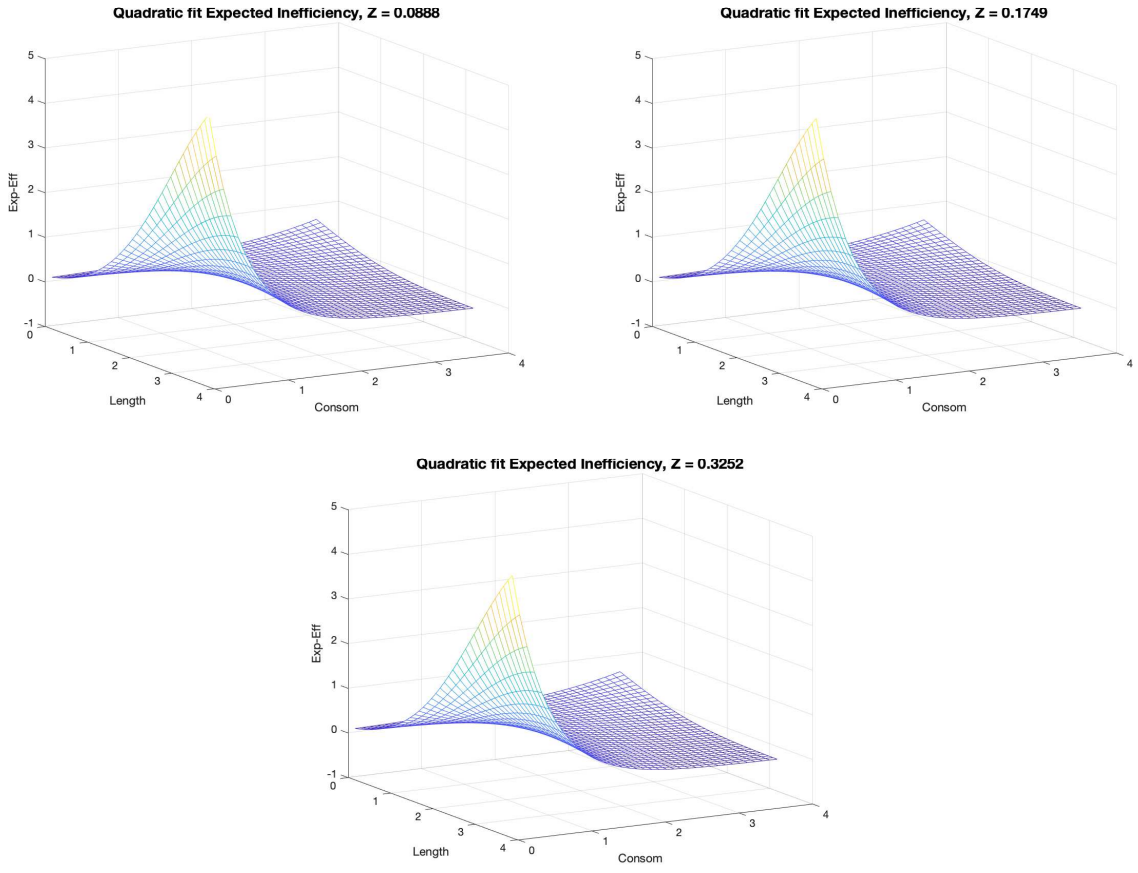

Figure 5: *Fitted values $\widehat{\mu}_\eta(X, Z)$ as a function of X for fixed values of Z at its 3 quartiles.*

In order to analyze the effect of the X on the expected inefficiencies for various values of Z. We select 3 values of Z (its 3 quartiles). The results are displayed in Figure 5 for a

grid of values for X. This confirm that there is a slight shift effect due to Z by the scaling $\exp(\widehat{\gamma}_z Z)$: with the value $\widehat{\gamma}_z = -01867$ this scaling factor takes the values 0.9836, 0.9679 and 0.9416, respectively for the 3 quartiles of Z. These small changes are not visible in Figure 5.

It may also be useful to analyze the estimates elasticities of the cost function with respect to the two outputs. Since the model is quadratic we have one estimated value for each data point. The boxplots in Figure 6 show that these cost elasticities are higher when computed relative to changes in the length of the network than relative to the level of consumption in the network. It seems we have mainly diseconomies of scale relative to X_1, the length of the network but scale of economies relative to X_2 the gas consumption of the customers.

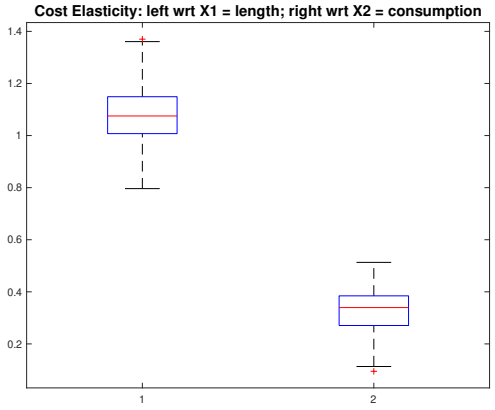

Figure 6: *Elasticities of cost with respect to X_1, the length of the network and with respect to X_2, the consumption in the network. Some outlying values (less than 5%) have been dropped out.*

Finally, it could be of interest for the practitioner and the manager of the network to have a look on the individual measures of expected inefficiencies to identify other sources of inefficiencies. These are provided for each data point by $\widehat{\mu}_\eta(X_i, Z_i)$ for $i = 1, \ldots, n$.

3 Extension to A Nonparametric Model

3.1 The Nonparametric Model and its Properties

The results of the preceding sections indicate that in our application, the chosen parametric model seems to be adequate to fit the data. But this is not always the case and then it is appealing to consider more flexible models and thus investigate how to extended our model to nonparametric models. We will propose nonparametric model for the cost function (in place of $\varphi(X, \beta)$) and for the expected inefficiency (in place of $\mu_\eta(X, Z, \gamma_\mu)$).

For the cost function, say $m(X)$, we can use local linear approaches (see e.g. Fan and Gijbels, 1996) and we can select local exponential estimators for the location model, say $\exp(g(X,Z))$, in order to satisfy the sign constraint (see Ziegelmann, 2002). In a nutshell, the idea is to follow the approach described above but by localizing the estimation in (2.8) near each data point by using an appropriate kernel function. But, as explained below, due to the particularity of our setup, we will not localize in Z to obtain an estimate of the engineering cost C_T that does not depend on the values of Z_i.

One might be tempted to extend the model (2.4) as follows. We assume that $X \in \mathbb{R}^d$ is the vector of outputs and $W = (X, Z) \in \mathbb{R}^{q_w}$

$$\begin{pmatrix} C_T \\ C_E \end{pmatrix} = \begin{pmatrix} m(X) \\ m(X) + \exp(g(X,Z)) \end{pmatrix} + \begin{pmatrix} u \\ v^* \end{pmatrix}, \tag{3.1}$$

where $\mathbb{E}(u|X,Z) = 0$ and $\mathbb{E}(v^*|X,Z) = 0$. Now, for any $w = (x,z)$ in a neighborhood of a point $w_0 = (x_0, z_0)$, we have the local linear and local exponential approximations

$$m(x) = m_0 + m_1'(x - x_0) \tag{3.2}$$

$$\mu_\eta(x,z) = \exp\left\{g_0 + g_1'(w - w_0)\right\}. \tag{3.3}$$

We could then obtain for any $w_0 = (x_0, z_0)$ the local estimators of the model by using weighted (by some kernel functions) (non-linear) least squares methods. By applying this approach, clearly the estimates of the cost function in C_T would depend on the local value z_0. As pointed above, in our setup, the engineering cost is not defined as dependent on any external factors Z so this approach is not valid.

We will rather use a model that localize only on the variables X and we will select a linear model for the impact of Z (any other parametric specification could be chosen) on the expected inefficiency. We have

$$\begin{pmatrix} C_T \\ C_E \end{pmatrix} = \begin{pmatrix} m(X) \\ m(X) + \exp\left\{g(X) + \gamma' Z\right\} \end{pmatrix} + \begin{pmatrix} u \\ v^* \end{pmatrix}. \tag{3.4}$$

Identification

Clearly, the function $m(X)$ is identified by the first equation of (3.4), but still even without this first equation, the function is also identified by only the second equation of (3.4).

The argument is going along the next lines. $\mathbb{E}(C_E|X,Z) = f(m,g,\gamma)$ is identified if $f(\bar{m},\bar{g},\bar{\gamma}) = f(m,g,\gamma)$ implies $(m,g,\gamma) = (\bar{m},\bar{g},\bar{\gamma})$. We need the quite mild following assumption for our argument.

Assumption 3.1. *X and Z are variational independent, i.e., for any function a of X and for all j,*

$$\frac{\partial}{\partial Z_j} a(X) = 0. \tag{3.5}$$

This assumption implies, a.o., that the support of the variables X does not depend on Z. Suppose, without loss of generality that γ_1, the coefficient of Z_1 is different from zero. We can compute

$$\frac{\partial}{\partial Z_1}\mathbb{E}(C_E|X,Z) = \gamma_1 \exp\left\{g(X) + \gamma' Z\right\}. \tag{3.6}$$

We verify now that this derivative is identified. Consider two values of the parameters (g,γ) and $(\bar{g},\bar{\gamma})$. If we have

$$\gamma_1 \exp\left\{g(X) + \gamma' Z\right\} = \bar{\gamma}_1 \exp\left\{\bar{g}(X) + \bar{\gamma}' Z\right\}, \tag{3.7}$$

we have also

$$\frac{\bar{\gamma}_1}{\gamma_1} \exp\left\{(\bar{\gamma}-\gamma)' Z\right\} = \frac{g(X)}{\bar{g}(X)} \tag{3.8}$$

If we derive with respect to Z we have

$$\frac{\bar{\gamma}_1}{\gamma_1}(\bar{\gamma}-\gamma) \exp\left\{(\bar{\gamma}-\gamma)' Z\right\} = 0, \tag{3.9}$$

which implies that we must have $\gamma = \bar{\gamma}$. So, by (3.8), we have also $g = \bar{g}$ and therefore necessarily $m = \bar{m}$. So we see that even with only the equation on C_E, our nonparametric stochastic cost function is identified (under Assumption 3.1).[2]

Estimation by local linearisation

Now we have for any $w = (x,z)$ in a neighborhood of a point $w_0 = (x_0,z)$, the local linear and local exponential approximations

$$m(x) = m_0 + m_1'(x - x_0) \tag{3.10}$$

$$\mu_\eta(x,z) = \exp\left\{g_0 + g_1'(x - x_0) + \gamma' z\right\}. \tag{3.11}$$

The estimation of the model can go along the following lines. Define for each observation $i = 1,\ldots,n$, the residuals

$$u_i = C_{T,i} - \left[m_0 + m_1'(X_i - x_0)\right] \tag{3.12}$$

$$v_i = C_{E,i} - \left[\left\{m_0 + m_1'(X_i - x_0)\right\} + \exp\left\{g_0 + g_1'(X_i - x_0) + \gamma' Z_i\right\}\right]. \tag{3.13}$$

[2]Interestingly, we could thus apply the procedure described in this section to an usual stochastic frontier model (here in the cost orientation), where we only have data on the economic costs, without the knowledge of the engineering costs. This means a stochastic cost frontier model ignoring the first equation in model (3.4). To the best of our knowledge this provides an original way to handle nonparametric stochastic frontier models.

where $m_0, g_0 \in \mathbb{R}$ and $m_1, g_1 \in \mathbb{R}^d$. Then a simple local estimators could be obtained by solving in $m = (m_0, m_1) \in \mathbb{R}^{d+1}$, $g = (g_0, g_1) \in \mathbb{R}^{d+1}$ and in $\gamma \in \mathbb{R}$ the following weighted least squares problem

$$\left(\widehat{m}(x_0), \widehat{g}(x_0), \widehat{\gamma}(x_0)\right) = \arg\min_{m,g,\gamma} \left\{ (\boldsymbol{u}' \quad \boldsymbol{v}') \boldsymbol{\mathcal{W}} \begin{pmatrix} \boldsymbol{u} \\ \boldsymbol{v} \end{pmatrix} \right\}, \tag{3.14}$$

where $\boldsymbol{u}' = (u_1, \ldots, u_n)$, $\boldsymbol{v}' = (v_1, \ldots, v_n)$ and $\boldsymbol{\mathcal{W}}$ is the $(2n \times 2n)$ matrix of weights

$$\boldsymbol{\mathcal{W}} = \begin{pmatrix} \boldsymbol{\mathcal{W}}_n & 0_{n,n} \\ 0_{n,n} & \boldsymbol{\mathcal{W}}_n \end{pmatrix}, \tag{3.15}$$

with $0_{n,n}$ being a matrix of zeros and $\boldsymbol{\mathcal{W}}_n = \text{diag}(K_h(X_i - x_0))$ is the standard $n \times n$ diagonal matrix of the kernel weights (in practice we will use product kernels). Note that the solution in γ of (3.14) will depend on the chosen local value of x_0. The bandwidths $h = (h_1, \ldots, h_d)$ may be computed by leave-one-out cross validation (here on the pairs $(C_{T,i}, C_{E,i})$, taking for each i, the sum of the two). In practice we compute these local estimates for x_0 being one of the observations X_j for $j = 1, \ldots, n$.

The above approach in (3.14)–(3.15) may work but it assumes implicitly that the two functions $m(x)$ and $g(x)$ have similar shape, allowing a common bandwidth $h \in \mathbb{R}^d$ for localizing both parts of the models. There is no reasons why it should always be the case, so it is more appropriate to allow different shapes of the two functions, and so to use two different bandwidths, say $h_m \in \mathbb{R}^d$ and $h_g \in \mathbb{R}^d$ for the two parts of the model.

This can be achieved in an iteration process in the spirit of backfitting (see e.g. Mammen and Park, 2005). We could start the process with initial values of the estimators obtained by the joint localisation with a common bandwidth $h \in \mathbb{R}^d$ described above in (3.14)–(3.15). Practically the initial step, with $k = 0$, is given by the solution for $i = 1, \ldots, n$ of

$$\begin{aligned}(\widehat{m}^{(0)}(X_i), \widehat{g}^{(0)}(X_i), \widehat{\gamma}^{(0)}(X_i)) = \arg\min_{m,g,\gamma} \sum_{j=1}^{n} \Big\{ &\left[(C_{T,j} - (m_0 + m_1'(X_j - X_i)) \right]^2 K_h(X_j - X_i) \\ &+ \left[C_{E,j} - (m_0 + m_1'(X_j - X_i) + \exp(g_0 + g_1'(X_j - X_i) + \gamma' Z_j) \right]^2 K_h(X_j - X_i) \Big\},\end{aligned} \tag{3.16}$$

where the common h may be computed by the usual least squares leave-one-out cross validation (LSCV). We will now compute two different localisations for the two parts of the model each with its own vector of bandwidths. The iteration goes then as follows:

[1] Set $k = k + 1$. We first start with the expected inefficiency term and solve for $i =$

$1, \ldots, n$ the problems

$$(\widehat{g}^{(k)}(X_i), \widehat{\gamma}^{(k)}(X_i)) = \arg\min_{g,\gamma} \sum_{j=1}^{n} \left\{ \left[C_{E,j} - m_0^{(k-1)}(X_j) - \exp(g_0 + g_1'(X_j - X_i) + \gamma' Z_j) \right]^2 K_{h_g}(X_j - X_i) \right\}, \tag{3.17}$$

where now a specific bandwidth $h_g \mathbb{R}^d$ can be specified to the usual LSCV.

[2] Having the solutions $(g^{(k)}, \gamma^{(k)})$, the second step of the iteration k is given by the solution for $i = 1, \ldots, n$ tof the problems

$$\widehat{m}^{(k)}(X_i) = \arg\min_{m} \sum_{j=1}^{n} \left\{ \left[C_{T,j} - (m_0 + m_1'(X_j - X_i)) \right]^2 K_{h_m}(X_j - X_i) + \left[C_{E,j} - \exp(g_0^{(k)}(X_j) + \gamma^{(k)\prime}(X_j) Z_j) - (m_0 + m_1'(X_j - X_i)) \right]^2 K_{h_m}(X_j - X_i) \right\}, \tag{3.18}$$

where the bandwidths $h_m \in \mathbb{R}^d$ can be selected by LSCV.

As in the backfitting algorithm, we iterate till convergence of the solutions.

The asymptotic properties of backfitting methods have been investigated by Fan et al. (1998) and Mammen et al. (1999). Roughly speaking the estimates shares the usual properties of the nonparametric estimators (local linear, Fan and Gijbels, 1996 or local exponential, Ziegelman, 2002), with the rate of convergence $\sqrt{nh^d}$ driven by the number of variables (d) used in each step of the backfitting algorithm.

3.2 Nonparametric Approach for our Application

We use the same data set on the distribution of gas and estimate the model (3.4) with the algorithm described above. The iterations were mostly stabilized after 5 or 6 iterations (for each of the fitted values of the cost and of the expected inefficiency, a change of value less than 10^{-6} from one iteration to the next one). We give the final results after 10 iterations.[3]

As above we have 2 outputs ($X_1 = L$, the length of the network and $X_2 = C$, the consumption of gas of the network). The external variable Z is as above the density of the network, and to illustrate the flexibility of our approach, we use a quadratic model $\gamma_1' Z + \gamma_2' Z^2$ for the effect of Z on the expected inefficiencies.[4] The model can thus be written as

$$\begin{pmatrix} C_T \\ C_E \end{pmatrix} = \begin{pmatrix} m(X) \\ m(X) + \exp\{g(X) + \gamma_1' Z + \gamma_2' Z^2\} \end{pmatrix} + \begin{pmatrix} u \\ v^* \end{pmatrix}.$$

[3] For our large sample of $n = 1021$ data, it took less than 1h30 on a PC, with a processor Intel(R), 3.10Ghz with 6 cores, 16G Ram. The iteration 0, took 14 minutes.

[4] We did the exercise without the variable Z^2 and we obtained quite similar results.

with $\mathbb{E}(u|X,Z)=0$ and $\mathbb{E}(v^*|X,Z)=0$

The results can be summarized as follows: the starting value (iteration $k=0$) provided a common bandwidth $h' = (h_L \ h_C) = (0.1999 \ 0.2314)$ whereas after the 10 iterations we obtain the values $h'_m = (3.9389 \ 0.3045)$ for estimating $m(X)$ and $h'_g = (0.5520 \ 3.5671$ for estimating $\exp\{g(X)+\gamma'_1 Z+\gamma'_2 Z^2\}$. Clearly quite different values of the bandwidths for each component.

Figure 7 displays the data points and the fitted cost function for both C_T and C_E on a grid of values for (X_1, X_2).

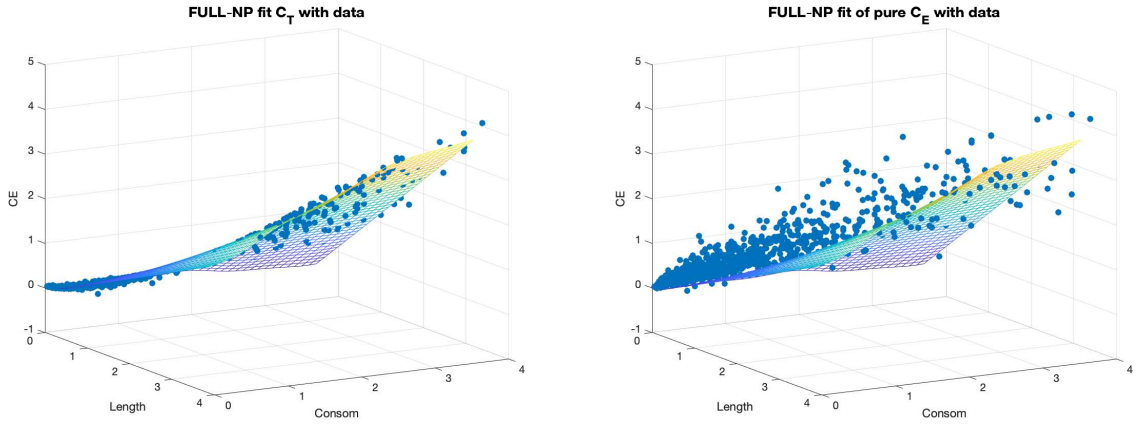

Figure 7: *Nonparametric approach: left panel, data points of the observed technical costs C_T and the fitted cost frontier $\widehat{m}_0(X_i)$. Right panel, data points of the observed economic costs C_E and the same fitted cost frontier.*

The final fit is very good and, as expected, quite similar to what we obtain for the quite flexible parametric model used in Section 2.3, see Figure 2. The same comments apply for both pictures. Here again, the global quality of nonparametric model can be appreciated through its R^2. We have $R^2_{C_T} = 0.9942$ and $R^2_{C_E} = 0.9305$. This may also be viewed in Figure 8, which slightly improve the fits when compared to its parametric counterparts in Figure 3.

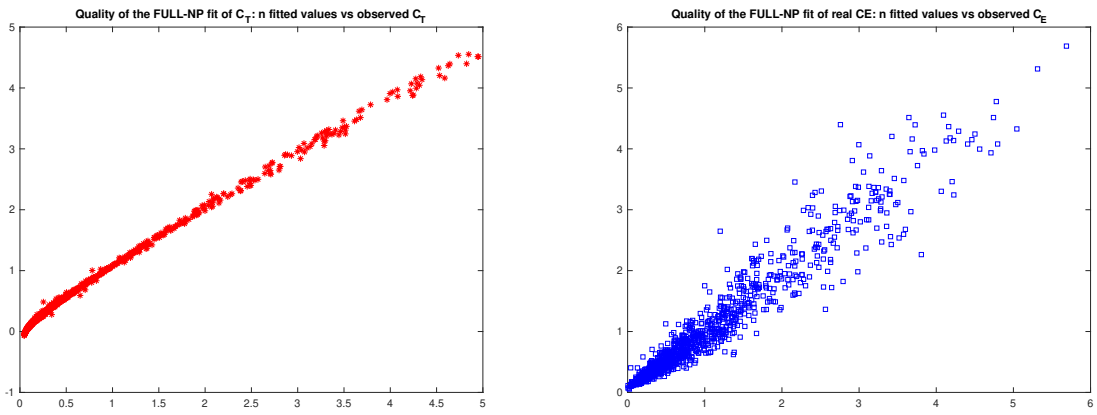

Figure 8: *Nonparametric approach: left panel, fitted values of technical costs* $\widehat{C}_T = \widehat{m}_0(X_i)$ *versus observed* C_T*. Right panel, fitted values of economic costs* $\widehat{C}_E = \widehat{m}_0(X_i) + \exp\{\widehat{g}_0(X_i) + \widehat{\gamma}_1(X_i)Z_i + \widehat{\gamma}_2(X_i)Z_i^2\}$*, versus observed* C_E*.*

Figure 9 shows the distribution of the expected inefficiencies

$$\widehat{\mu}_\eta(X_i, Z_i) = \exp\left\{\widehat{g}_0(X_i) + \widehat{\gamma}_1(X_i)Z_i + \widehat{\gamma}_2(X_i)Z_i^2\right\} \tag{3.19}$$

obtained for $i = 1, \ldots, n$, in the nonparametric approach. It is again quite similar to the one obtained in our parametric approach, in Figure 4, slightly more concentrated below the value 1 (less inefficiencies in C_E).

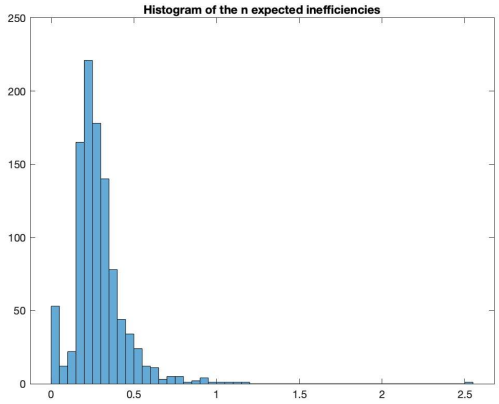

Figure 9: *Nonparametric approach. Histogram of the* n *expected inefficiencies* $\widehat{\mu}_\eta(X_i, Z_i)$*.*

Here again, we can analyze the effect of the two outputs on the expected inefficiencies for 3 fixed values of Z (its 3 quartiles). We can see in Figure 10 that Z has very low effect

on the level of the expected inefficiencies. Compared with Figure 5 of the parametric case, we see that the peaks for small values of $X_2 = C$ and large values of $X_1 = L$ is much less important here. This is probably due to the higher flexibility of the nonparametric model.

Interestingly we provide in Figure the "scaling effect" of Z on the expected inefficiencies (i.e. $\exp\{\widehat{\gamma}_1(X_i)Z_i + \widehat{\gamma}_2(X_i)Z_i^2\}$) as a function of Z. We see the global decreasing effect of Z, with a lot of variations, with a slight positive upward curvature, probably introduced by the term Z^2. This is confirmed by Figure 12 which displays the boxplots of n estimated values of $\widehat{\gamma}_1(X_i)$ and $\widehat{\gamma}_2(X_i)$.

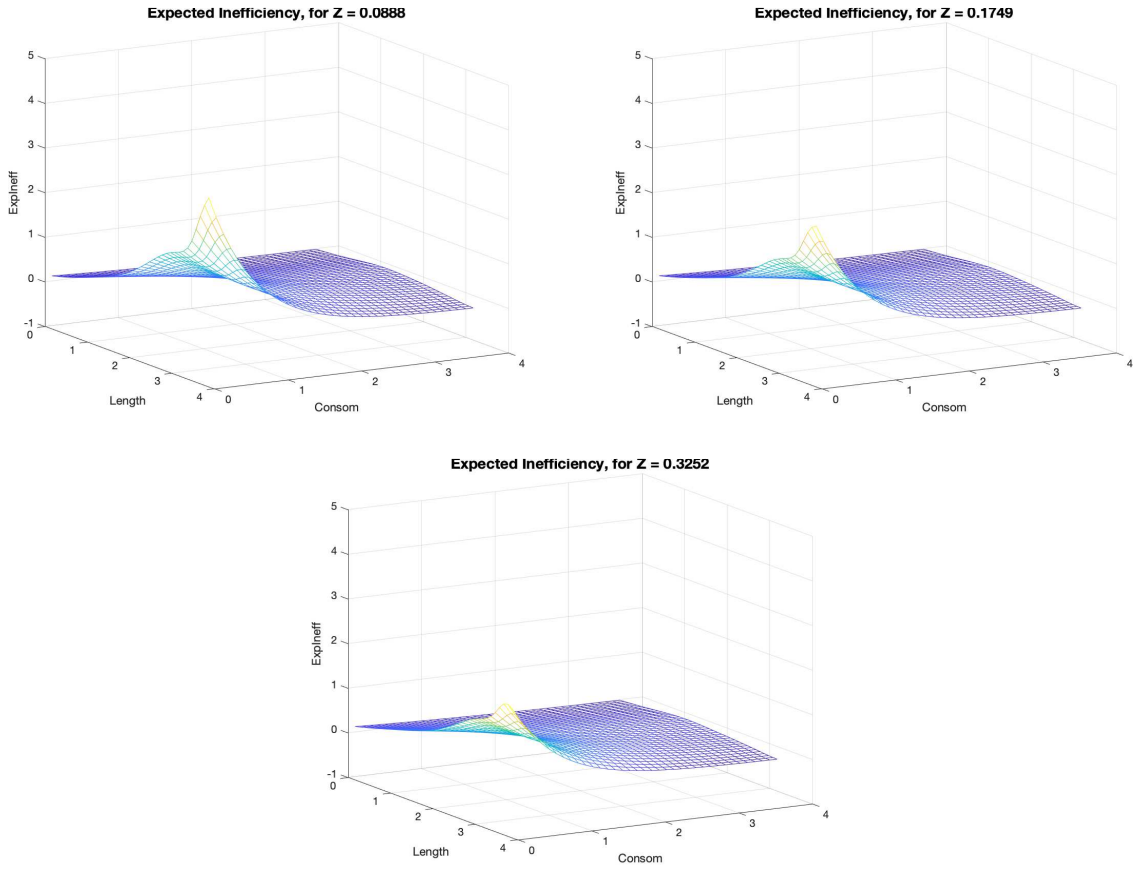

Figure 10: *Nonparametric approach: fitted values* $\widehat{\mu}_\eta(X, Z)$ *for a grid of values for* X *and for a fixed value of* Z *at its 3 quartiles.*

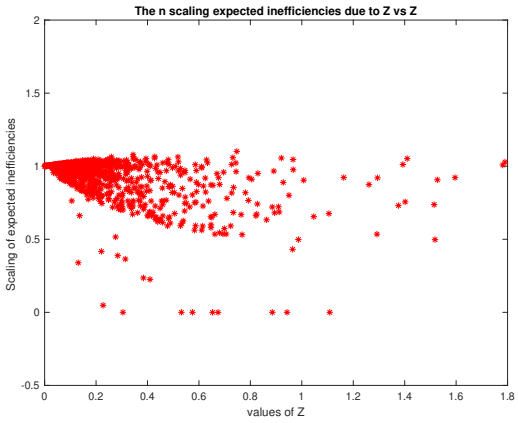

Figure 11: *Nonparametric approach. Scaling effect of Z on the expected inefficiencies $\widehat{\mu}_\eta(X_i, Z_i)$. A couple of numerical outliers (huge values) have been eliminated.*

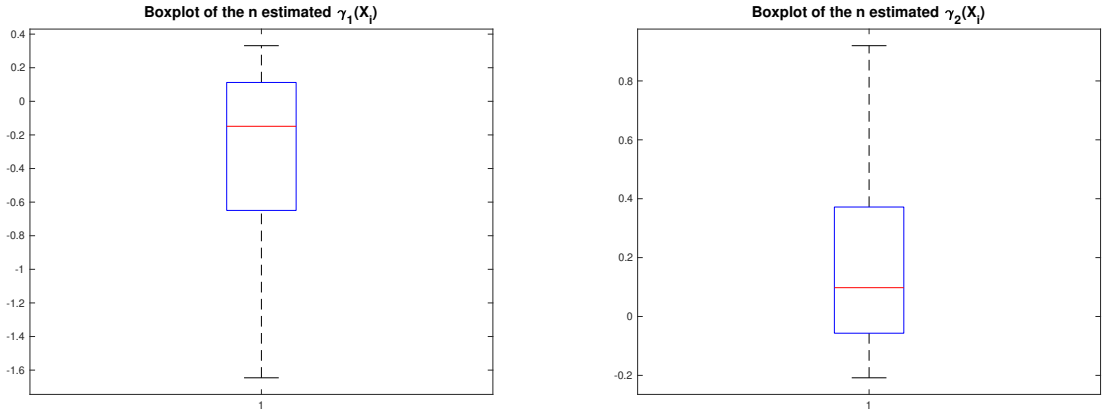

Figure 12: *Nonparametric approach: boxplots of n estimated values of $\widehat{\gamma}_1(X_i)$ and $\widehat{\gamma}_2(X_i)$. A few outlying values (less than 6%) have been eliminated.*

Finally, the nonparametric approach gives more additional results, for instance we can reproduce the boxplots of the first derivatives of the cost function with respect to the two outputs, i.e. $\widehat{m}_{1,X_1}(X_i)$ and $\widehat{m}_{1,X_2}(X_i)$. We have for each derivatives n values and their boxplots are provided in Figure 13.

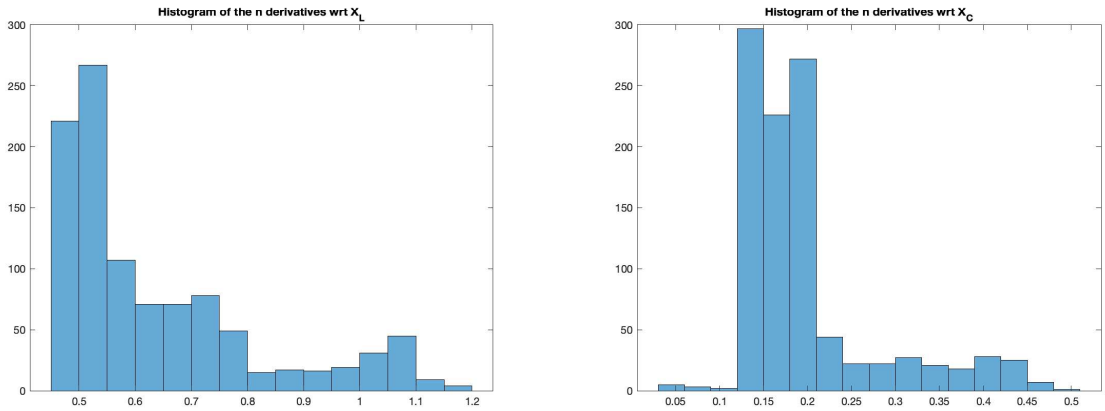

Figure 13: *Nonparametric approach: boxplots of the n estimated derivatives* $\widehat{m}_{1,X_1}(X_i)$ *and* $\widehat{m}_{1,X_2}(X_i)$.

In fact we can also provide the boxplots of the estimated elasticities as we did for our parametric model in Figure 6. The results are shown in Figure 14. Again, as in the parametric case we see that we have mainly diseconomies of scale relative to X_1, the length of the network but scale of economies relative to X_2 the gas consumption of the customers.

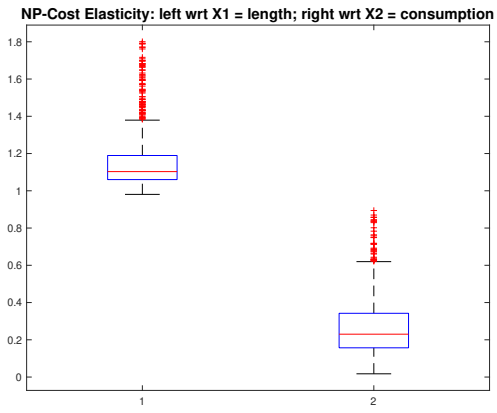

Figure 14: *Elasticities of cost with respect to* X_1*, the length of the network and with respect to* X_2*, the consumption in the network. Some outlying values (less than 10%) have been dropped out.*

4 Conclusions

This paper analyzes situations in which cost evaluations for producing goods can be calculated using various "normative," "technical," or "engineering" methods by decomposing the

process into elementary tasks based on the characteristics of the outputs. Some error terms may be interpreted as idiosyncratic approximation errors. Conversely, economists rely on an accounting measure of the real cost; however, this includes the potential for inefficiencies in the production process.

Our first original contribution is the proposal of a model that reconciles or combines both types of cost measures. This results in a two-equation model: one regression for the technical cost and one stochastic cost frontier for the economic cost. We develop the methodology for managing this combination, initially within a purely parametric framework. Since our model accommodates least-squares methods, there is no need to specify any stochastic assumptions regarding the error terms or the inefficiency distribution.

Our second original contribution is a nonparametric extension of the model and we prove the identifiability of the model. By doing so, we present also a novel approach to modeling nonparametric stochastic frontier models, specifically in a cost-oriented context.

These methodologies were inspired by the analysis of the cost function for gas distribution. In this specific dataset, the flexible parametric model we selected produces results comparable to those of the more flexible nonparametric model. However, this finding is not universally applicable, as the nonparametric approach offers a considerably less restrictive framework for modeling cost functions. In summary, our application reveals diseconomies of scale concerning the cost relative to the length of the network and economies of scale regarding the gas consumption of customers. We have also observed some impact of these two outputs on expected inefficiency. However, the only available external factor that might influence these inefficiencies–the density of the network–appears to have a positive effect on efficiency, albeit on a smaller scale. One way to improve the model would be to explore additional external factors that could enhance the fit of expected inefficiencies and, consequently, the observed economic costs.

References

[1] Cameron, A.C. and P.K. Trivedi (2005), *Microeconometrics, Methods and Applications*, Cambridge University Press, New York.

[2] Carrasco, M. and J.P. Florens (2000), Generalization of GMM to a Continuum of Moment Conditions. *Econometric Theory*, 16, 797-834.

[3] Chenery, H.B. (1949). Engineering production functions, *Quarterly Journal of Economics*, 63, 507-531.

[4] Fan, J. and I. Gijbels (1996), *Local Polynomial Modelling and Its Applications*. Chapman and Hall, London.

[5] Fan, J., Härdle, W. and E. Mammen (1998), Direct Estimation of Low-Dimensional Components in Additive Models, *The Annals of Statistics*, 26, 3, 943–971.

[6] Hall, P. (1988). On Symmetric Bootstrap Confidence Intervals, *Journal of the Royal Statistical Society, B*, 50, 1, 35–45.

[7] Härdle, W.K. and L. Simar (2019), *Applied Multivariate Statistical Analysis*, Fifth Edition, Springer Nature, Switzerland.

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

[9] Mammen, E. (1992). *When Does Bootstrap Work? Asymptotic Results and Simulation*, Springer-Verlag, New York.

[10] Mammen, E., Linton, O. and J. Nielsen (1999), The Existence and Asymptotic Properties of a Backfitting Projection Algorithm under Weak Conditions, *The Annals of Statistics*, 27, 5, 1443–1490.

[11] Mammen, E. and B. U. Park (2005), Bandwidth Selection for Smooth Backfitting in Additive Models, *The Annals of Statistics*, 33, 3, 1260–1294.

[12] Marsden J., Pingry D. and Whinston A (1974).). Engineering Foundations of Production Functions, *Journal of Economic Theory*, 9, 124–140.

[13] Massol O., (2011). A cost function for the natural gas transmission industry: further considerations. *The Engineering Economist*, 56(2), 95-122.

[14] White H. (1980), A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroscedasticity. *Econometrica*, 48, 817-838.

[15] Wibe, S. (1984). Engineering Production Functions: A survey. *Economica* published by Wiley on behalf of the London School of Economics, 51, 401-411.

[16] Ziegelman , F.A. (2002), Nonparametric Estimation of Volatility Functions: The Local Exponential Estimator. *Econometric Theory*,18, 4, 985–991.