Coûts et économies d'échelle dans la production d'eau en Côte d'Ivoire : une approche semiparamétrique.

(Version très préliminaire)

Daouda DIAKITÉ*

Décembre 2011

Résumé. Dans cet article, nous estimons différentes spécifications paramétriques et non paramétriques de fonctions de coûts de productions d'eau potable des villes de Côte d'Ivoire. Des tests sont menées pour comparer ces différentes spécifications et elles montrent qu'une fonction de coûts de production de type Translog homothétique est la mieux adaptée aux données ivoiriennes.

JEL Classification : C33, D24, Q25.

Môts clés : Service d'eau potable, Fonction de coût translog multi-produits, Rendements d'échelle, Données de panel.

*CÉMOI, Université de la Réunion. Mail : ddiakite@univ-reunion.fr

Introduction

L'analyse de la structure des coûts de production est essentielle pour toute industrie afin de réaliser une gestion optimale des ressources productives. Cette analyse permet entre autre de mettre en évidence la présence et/ou l'absence d'économies d'échelle lors du processus de production. Une bonne analyse des coûts de production passe nécessairement par une bonne connaissance de la relation qui lie les facteurs de production à la quantité produite. En d'autres termes, il est impératif de bien spécifier une fonction de coût de production. En statistique, et plus particulièrement en économétrie, la connaissance de la relation entre différentes variables pour lesquelles on dispose d'un jeu de données passe par une étape d'estimation de la relation en question. Les premières méthodes développées et largement utilisées de nos jours, les modèles dits purement paramétriques, font appel à des postulats contraignants qui réduisent considérablement la précision et le pouvoir de prédiction de tels modèles.

Plus concrètement, lorsqu'on désire analyser la relation entre une variable dépendante Y et une série de variables explicatives $X_1, \cdots, X_n$, la méthode paramétrique classique généralement utilisée est la régression linéaire. Celle-ci est très pratique puisqu'elle suppose un modèle simple, de la forme

$$y_i = \alpha_0 + \alpha_1 x_{1i} + \cdots + \alpha_n x_{ni} + \varepsilon_i$$

Ce modèle a l'avantage d'être simple à calculer et ses paramètres estimés faciles à interpréter. En outre, il repose sur un ensemble d'hypothèses notamment sur les résidus et celles-ci permettent d'opérer de nombreux tests statistiques. Cependant, la régression linéaire repose sur un postulat très restrictif à savoir la linéarité de la relation. Or dans bien des cas, cette linéarité est loin d'être justifiée et l'on est obligé de rechercher d'autres formes fonctionnelles à même de mieux décrire les données. La recherche de la bonne forme fonctionnelle (forme quadratique, logarithmique, forme non linéaire, transformation des données,etc...) devient très vite fastidieux et l'on reste dans tous les cas "prisonnier" de la forme retenue.

Ainsi, de nos jours, on assiste à l'apparition de nouvelles techniques de régression plus souples, qui laissent au jeu de données lui-même le soin de choisir la forme fonctionnelle entre les variables. Ces méthodes sont connues sous le nom de *régression non paramétrique*. Le principal avantage des méthodes non paramétriques est qu'elles ne supposent aucune forme fonctionnelle particulière, ce qui leur donne beaucoup plus de flexibilité. Du coup, la régression non paramétrique devient très utile dans les cas où les modèles paramétriques sont inopérants. Mieux, opposées dans un premier temps, les deux méthodes apparaissent in fine en pratique très complémentaires, la régression non paramétrique pouvant justifier le choix d'un modèle paramétrique.

Il existe une foisonnante littérature en régression non paramétrique avec différentes méthodes et divers estimateurs associés. Les méthodes les plus couramment utilisées sont

la méthode du noyau, les polynomes locaux, les polynomes d'ajustement, les splines de regression et les splines de lissage. Cependant, les méthodes non paramétriques présentent elles aussi certaines limites. D'abord, la flexibilité des différents estimateurs a un coût et les méthodes doivent arbitrer entre biais et variance. Plus on désire suivre fidèlement les données, plus on augmente la variance et moins l'estimateur est lisse. A l'opposé, l'utilisation d'un estimateur plus lisse, baisse la variance mais augmente le biais et l'on s'éloigne des données. Ensuite, il y a le problème lié d'une part au choix du paramètre de lissage et d'autre part le niveau technique de certains calculs. Ces derniers rendent ces méthodes non paramétriques difficiles d'accès pour tous. Enfin, les estimations sous contraintes et les tests d'hypothèses ne sont pas du tout aisé à mettre en oeuvre dans le contexte de ces modèles surtout pour les économistes. Cependant, le rythme de développement des packages des différents logiciels commencent à résoudre la plupart de ces difficultés.

L'objectif de ce travail est l'estimation non paramétrique de fonctions de coût de production dans l'industrie d'eau potable en Côte d'Ivoire. Plus concrètement, il s'agira d'estimer une fonction de coûts sous différentes spécifications mais dans le cadre de modèles semi-paramétriques (modèles à la frontière entre modèles purement paramétriques et modèles purement non paramétriques). La principale spécification reste toutefois une fonction de coût translog où les volumes d'eau produits apparaissent sous forme non paramétrique tandis que les autres variables apparaissent sous forme paramétrique. Pour ce faire nous disposons de données de panel sur 145 services d'eau en milieu urbain observés sur 5 années (1998 à 2002). but de ce mémoire est de se familiariser avec ces techniques. Ainsi dans un premier chapitre, nous ferons une présentation succinte de certains des estimateurs susmentionnés. Dans ce qui suit, nous présentons tout d'abord dans la section 2, les modèles non paramétriques, ensuite les fonctions de coûts semi-paramétriques sont dérivées dansla section 3. La section 4 est consacrée à la présentation des données utilisées, l'estimation des fonctions de coûts ainsi que les résultats des estimations. La section 5 donne enfin les élements de conclusion.

1 Typologie des modèles

Dans cette section, nous décrivons succinctement les deux grandes familles de modèles en régression non paramétrique. Les modèles *purement non paramétriques* et les modèles *semi-paramétriques*.

1.1 Les modèles purement non paramétriques

Dans sa formulation très générale, les modèles de régression purement non paramétrique se présentent sous la forme suivante :

$$y_i = r(x_{1i}, x_{2i}, \ldots, x_{pi}) + \epsilon_i, \quad i = 1, \ldots, n \tag{1.1}$$

ou sous sa forme matricelle

$$Y = r(X) + \epsilon$$

Cette formulation ne suppose aucune forme précise pour la relation entre la variable à expliquer et les variables explicatives. L'objectif de la régression non paramétrique est alors d'obtenir une estimation de la fonction $r(X)$ du modèle (1.1), ce qui est loin d'être une évidence. Plusieurs modèles non paramétriques ont été proposés au cours du temps et ils possèdent tous leurs avantages et leurs inconvénients. Nous en présentons les plus utilisés.

Le premier de ces modèles est le *modèle additif généralisé (GAM : generalized additive model)* développé par Hastie et Tibshirani (1990). Ce modèle est la version non paramétrique du modèle linéaire généralisé (GLM). Ainsi, pour obtenir une estimation de la fonction de régression $r(X)$, on suppose d'abord que le modèle (1.1) peut s'écrire sous la forme :

$$y_i = \alpha + \sum\nolimits_{j=1}^{p} r_j(x_{ji}) + \epsilon_i, \quad i = 1, \ldots, n \tag{1.2}$$

où les erreurs ϵ_i sont supposées non corrélées entre elles, de moyenne nulle et de variance σ^2. Les r_j sont des fonctions arbitraires pour lesquelles ont suppose que $E\left[r_j\left(X_j\right)\right] = 0$ où cette espérance est prise par rapport à la distribution marginale de X_j. On suppose ici donc que la relation entre la variable dépendante Y et les variables explicatives est strictement additive. Cette additivité ajoutée à la séparabilité du modèle constitue un avantage important du point de vue de l'interprétation et de la visualisation de la fonction de régression, puisque l'on peut alors analyser chaque variable explicative séparément. Par contre, dans le cas où la relation entre les variables comporte des interactions importantes, le modèle additif peut conduire à des estimations erronées, à moins que l'on ait déjà une bonne idée de ces dernières et que l'on modifie le modèle en conséquence.

Pour pallier les insuffisances des modèles GAM, des variantes ont été proposées. La première de ces variantes consiste à inclure dans le modèle des termes d'interactions du type $r_{jk}\left(X_j, X_k\right)$ et/ou de façon plus générale du type $r_{j\ldots k}\left(X_j, \ldots, X_k\right)$ entre variables explicatives selon le même principe que dans la régression linéaire multiple. Cependant au delà des interactions du niveau 2, les autres interactions restent difficiles à analyser surtout graphiquement.

A côté des modèles GAM, on trouve les modèles dits *additifs mixtes*. Ces modèles se présentent sous la forme suivante :

$$y_i = U_i + \sum_{j=1}^{p} r_j(x_{ji}) + \epsilon_{ij}, \quad j = 1, \ldots, m_i, \quad i = 1, \ldots, n \tag{1.3}$$

où les $U_i \sim N\left(0, \sigma_U^2\right)$ sont des constantes aléatoires pour chaque arbre et les ϵ_{ij} les termes d'erreurs.

Des variantes de ce modèle peuvent être obtenues comme dans les modèles de type GAM, en y rajoutant des termes d'interaction.

1.2 Les modèles semi-paramétriques

De nos jours, avec le développement rapide des packages des différents logiciels, on arrive à estimer la plupart des modèles purement paramétriques. Toutefois, ces modèles restent encore difficiles à mamier et surtout l'interprétation des résultats n'est pas ce qu'il y a de plus aisé. Ainsi par souci de parsimonie, des modèles hybrides à cheval entre modèles purement paramétriques et modèles purement non paramétriques ont été élaborés; ce sont les modèles semiparamétriques.

Le plus simple de ces modèles est *le modèle partiellement linéaire (partial linear model)* qui sous sa forme la plus simple s'écrit comme suit :

$$Y = Z\beta + r(X) + \epsilon$$

où $E\left(\epsilon/Z, X\right) = \sigma_\epsilon^2$. Pour simplifier, toutes les variables sont supposées être des scalaires et Z est une fonction non paramétrique de X ($Z = g\left(X\right) + u$). Ainsi on suppose que $E\left(\epsilon/Z, X\right) = g\left(X\right)$ et $Var(Z/X) = \sigma_u^2$.

La généralisation des modèles partiellement linéaires donne *les modèles partiellement paramétriques (partial parametric model*) qui se présentent sous la forme suivante :

$$Y = f\left(Z; \beta\right) + r(X) + \epsilon$$

où $f(.)$ est une fonction connue.

Une autre forme de modèles semi-paramétriques est connue sous le nom de *modèle d'indice (Index model)* qui est donné par :

$$Y = r(X\beta) + \epsilon$$

où pour toute valeur fixée de l'indice $X\beta$, la fonction $r(X\beta)$ est constante.

Une généralisation naturelle de ces modèles d'indices est connue sous le nom de *modèle d'indice partiellement linéaire* donné par :

$$Y = r(X\beta) + Z\delta + \epsilon$$

Une variante de ce dernier modèle est donnée par le modèle suivant :

$$Y = r\left[f\left(X;\beta\right)\right] + Z\delta + \epsilon$$

où $f(.)$ est une fonction connue.

2 Les fonctions de coûts semi-paramétriques

Il existe différentes formes fonctionnelles pour estimer les fonctions de coûts et elles peuvent être regroupées en deux catégories : les formes fonctionnelles simples et les formes fonctionnelles flexibles.

2.1 Les formes fonctionnelles simples

Cette catégorie comprend entre autres les fonctions de type Cobb-Douglas, Léontief et CES (Constant Elasticity of substitution). Ces fonctions ont la particularité d'imposer la constance de l'élasticité de substitution entre facteurs de production; restriction en somme assez forte. En effet, la fonction Cobb-Douglas suppose que l'élasticité de substitution est égale à l'unité (substitution parfaite entre facteurs), la fonction Léontief suppose une élasticité nulle (stricte complémentarité entre facteurs) tandis que la fonction CES, en fait une généralisation des deux premières[1], suppose une élasticité constante sans toutefois fixer une valeur.

Dans sa forme purement paramétrique, la fonction de coût de production Cobb-Douglas se présente comme suit :

$$CV_h = A_0 \prod_{j=1}^{p} \left(y_{jh}\right)^{\alpha_j} \prod_{i=1}^{n} \left(w_{ih}\right)^{\beta_i} \prod_{k=1}^{K} \left(z_{ih}\right)^{\delta_k} \nu_h$$

où l'indice h désigne l'entreprise, w_i le prix des facteurs de production, y_i les quantités des différents produits et z_i les facteurs techniques. Il est plus pratique de prendre le logarithme de sorte à obtenir une forme linéaire :

$$ln(CV_h) = A_0 + \sum\nolimits_{j=1}^{p} \alpha_j ln(y_{jh}) + \sum\nolimits_{i=1}^{n} \beta_i ln(w_{ih}) + \sum\nolimits_{k=1}^{K} \delta_k ln(z_{kh}) + \epsilon_h$$

[1]Lorsque l'élasticité de substitution d'une fonction CES est égale à l'unité, on montre qu'elle tend à se confondre avec une fontion Cobb-Douglas. A contrario, lorsque son élasticité est nulle, on montre qu'elle se confond avec une fonction de type Léontief.

On en déduit aisément la *version semi-paramétrique* de cette fonction qui s'écrit comme suit :

$$ln(CV_h) = A_0 + \sum_j r_j \left[ln(y_{jh})\right] + \sum_i \beta_i ln(w_{ih}) + \sum_k \delta_k ln(z_{kh}) + \epsilon_h$$

La fonction de production CES paramétrique linéarisée se présente sous la forme :

$$ln(CV_h) = A_0 + \sum_j \alpha_j ln(y_{jh}) + \frac{1}{\rho} ln \left[\sum_i^{n-1} \beta_i \left[w_{ih}\right]^\rho + \left(1 - \sum_i^{n-1} \beta_i\right) \left[w_{nh}\right]^\rho \right] + \sum_k \delta_k ln(z_{kh}) + \epsilon_h$$

d'où la *version semi-paramétrique* suivante :

$$ln(CV_h) = A_0 + \sum_j r_j \left[ln(y_{jh})\right] + \frac{1}{\rho} ln \left[\sum_i^{n-1} \beta_i \left[w_{ih}\right]^\rho + \left(1 - \sum_i^{n-1} \beta_i\right) \left[w_{nh}\right]^\rho \right] + \sum_k \delta_k ln(z_{kh}) + \epsilon_h$$

2.2 Les formes fonctionnelles flexibles

Pour lever l'hypothèse forte imposée aux premières (constance de l'élasticité de substitution), des formes fonctionnelles dites flexibles ont été introduites dans la littérature. Ce sont les fonctions de type Translog (transcendental logarithmic), Léontieff généralisée, McFadden Généralisée et Barnett Généralisée. De nos jours, ces spécifications sont les plus utilisées dans la littérature et en particulier la fonction de coût Translog introduite par Christensen, Jorgensen et Lau (1973). En effet, cette dernière présente de nombreux avantages théoriques et pratiques. D'une part, elle est basée sur un modèle économique; ce qui permet d'introduire de manière explicite la théorie économique dans la modélisation. D'autre part, elle impose peu de restrictions a priori sur les caractéristiques de la technologie de production et satisfait l'hypothèse d'homogénéité en prix à travers un ensemble de restrictions linéaires sur les paramètres. Cependant, la fonction Translog présente deux faiblesses majeures. En effet, la plupart des travaux empiriques sur les coûts relèvent que la fonction ne satisfait pas aux conditions de concavité[2] d'une part et qu'elle reste indéfinie pour un niveau de production nul d'autre part. Pour pallier ces insuffisances, en particulier la non concavité, les fonctions Léontieff généralisée, McFadden Généralisée et Barnett Généralisée ont été proposées. Nonobstant ses limites, la fonction Translog demeure la plus utilisée. Elle est en outre bien adaptée au cas multi-produits, et les équations de parts de coût dérivées sont linéaires dans les paramètres et peuvent être

[2]Elle peut cependant être imposée (Diewert et Wales, 1987)

estimées conjointement à la fonction de coût. Nous nous limiterons à cette fonction parmi les formes flexibles pour notre étude.

La forme Translog est une approximation de second ordre de toute fonction deux fois différentiable, inconnue a priori, et s'écrit (en utilisation tous les termes possibles) :

$$\begin{aligned} ln(CV_h) = A_0 &+ \sum_j \alpha_j ln(y_{jh}) + \sum_i \beta_i ln(w_{ih}) + \sum_k \delta_k ln(z_{kh}) \\ &+ \frac{1}{2}\sum_j \sum_k \lambda_{jk} ln(y_{jh}) ln(y_{kh}) + \frac{1}{2}\sum_i \sum_q \gamma_{iq} ln(w_{ih}) ln(w_{qh}) \\ &+ \frac{1}{2}\sum_r \sum_s \zeta_{rs} ln(z_{rh}) ln(z_{sh}) + \sum_i \sum_j \eta_{ij} ln(w_{ih}) ln(y_{jh}) \\ &+ \sum_i \sum_r \theta_{ir} ln(w_{ih}) ln(z_{rh}) + \sum_j \sum_r \xi_{jr} ln(y_{jh}) ln(z_{rh}) + \epsilon_h \end{aligned} \tag{2.1}$$

La *forme semi-paramétrique* s'en déduit aisément et est donnée par :

$$\begin{aligned} ln(CV_h) = A_0 &+ \sum_j r_j \left[ln(y_{jh})\right] + \sum_i \beta_i ln(w_{ih}) + \sum_k \delta_k ln(z_{kh}) \\ &+ \frac{1}{2}\sum_j \sum_k \lambda_{jk} ln(y_{jh}) ln(y_{kh}) + \frac{1}{2}\sum_i \sum_q \gamma_{iq} ln(w_{ih}) ln(w_{qh}) \\ &+ \frac{1}{2}\sum_r \sum_s \zeta_{rs} ln(z_{rh}) ln(z_{sh}) + \sum_i \sum_j \eta_{ij} ln(w_{ih}) ln(y_{jh}) \\ &+ \sum_i \sum_r \theta_{ir} ln(w_{ih}) ln(z_{rh}) + \sum_j \sum_r \xi_{jr} ln(y_{jh}) ln(z_{rh}) + \epsilon_h \end{aligned} \tag{2.2}$$

3 Données de l'étude et estimations

3.1 Données sur les coûts de production

Nos données concernent les services d'eau potable de Côte d'Ivoire. Dans ce pays, la production de l'eau potable, à l'instar de sa distribution dans les zones urbaines, est assurée par un même et unique opérateur, la SODECI (Société de Distribution d'Eau de Côte d'Ivoire). Il n'y a cependant pas un réseau unique d'eau potable car chaque grande localité du pays dispose et de son centre de production et de son réseau de distribution. Mieux, ces différents centres de production sont indépendants les uns des autres car l'eau produite dans un centre de production donné est distribuée dans une ou des localités bien précises.

Table 1: Statistiques Descriptives

Variables	Unités	Obs	Moy	Ecart type	Min	Max
V_{prod}	m^3	147	933 320	2 801 644	4731	2,38 10^7
V_f	m^3	147	763 700	2 118 272	4021	1,62 10^7
V_p	m^3	147	169 620	779 548	56	9 122 616
$rendt$	%	147	0,87	0,12	0,36	0,91
$abon$		147	2 977	6 417	35	41 437
$long$	km	147	67	180	4	2 012
$prod$	m^3/j	147	185	524	3	4 410
$stoc$	m^3	147	1107	2 860	1	20 000
com		147	2	4	1	34
cv	$F\ CFA$	147	8,82 10^7	1,80 10^8	850	1,28 10^9
s_e		147	0,22	0,09	0,01	0,58
s_l		147	0,21	0,09	0,04	0,54
s_m		147	0,57	0,10	0,29	0,98
w_e	F/Kwh	147	63,61	26,65	22	367,48
w_l	F/m^3	147	65,93	95,75	1,98	1443,32
w_m	F/m^3	147	177,91	277,46	12,47	3161,65

Pour cette étude, les données ont été collectées essentiellement auprès de la SODECI. Nous disposons des données sur la période 1998-2002 pour l'ensemble des centres de production du pays. Sur l'ensemble, nous avons retenu dans notre échantillon les 147 localités qui avaient un réseau de production et de distribution d'eau potable avant le début de la période d'étude (1998). Ceci nous donne un échantillon en panel cylindré de 735 observations. Cependant, pour ce mémoire, nous utiliserons que les données de l'année 2002 soit 147 observations. Nous prendons en compte la structure en panel de ces données dans des versions ultérieures de ce travail.

V_{prod} désigne les volumes d'eau produits par les services d'eau de la localité. La différence entre volumes produits et volumes facturés (V_f) donne les volumes d'eau perdus (V_p) tandis que leur rapport (volumes facturés/volumes produits) donne le rendement du réseau ($Rendt$). Les différents volumes sont exprimés en m^3. Les volumes produits vont de $4700\ m^3$ à $24\ millions$ de m^3 et les volumes facturés y représentent en moyenne 82% contre 18% pour les volumes perdus.

Le coût variable d'exploitation (CV) pour chaque service de l'échantillon est donc la somme des dépenses en main-d'oeuvre (L), des dépenses en électricité (E) et de toutes les autres dépenses que nous regroupons sous le nom de matériel (M).

Le prix de l'électricité (w_e) est obtenu en divisant les dépenses totales en énergie par sa consommation totale pour chaque service et il est exprimé en $F\ CFA/Kwh$. Pour le

prix du travail (w_l), nous n'avons malheureusement pas le nombre exact de travailleurs pour chaque service, encore moins le nombre d'heures travaillées. Ainsi pour le prix du travail ainsi que celui du matériel (w_m), nous avons simplement divisé les dépenses totales respectives par le volume d'eau total correspondant produit par service. Ces deux prix sont donc exprimés en $F\ CFA/m^3$.

Les variables techniques de l'étude sont le nombre d'abonnés ($Abon$), le nombre de localités désservies (com), la longueur du réseau ($Long$), la capacité de stockage ($Stoc$) et la capacité de production ($Prod$). Les statistiques descriptives relatives à l'ensemble de ces variables utilisées dans l'étude sont données dans le tableau (1).

3.2 Estimations

Pour estimer une fonction de coût translog, il est indispensable de choisir un point de référence autour duquel se fait l'approximation. Différents points d'approximation sont proposés dans la littérature notamment l'origine du repère. Ici, nous prendrons comme point d'approximation la moyenne du logarithme des variables explicatives. Le logarithme du coût variable étant une fonction des variables transformées logarithmiquement, cela revient à diviser toutes nos variables explicatives par leur moyenne géométrique respective. Le choix d'un tel point permet de minimiser les erreurs d'estimations des élasticités évaluées à la moyenne géométrique des variables.

La fonction de coût doit vérifier certaines conditions de régularité afin de correspondre à une structure de production "*well-behaved*". Avant toute estimation, nous imposons la condition d'homogénéité de degré 1 dans le prix des facteurs (en divisant le coût variable et le prix des autres inputs par le prix du travail). L'imposition et/ou la vérification des autres conditions notamment la monotonicité, la concavité et la symétrie (beaucoup plus complexes) ne seront pas abordées dans le cadre de ce mémoire.

Nous allons donc estimer successivement trois fonctions de coût purement paramétriques ainsi que leur équivalent semi-paramétrique et à titre de comparaison.

3.2.1 Modèles paramétriques.

La première fonction considérée est de type Cobb-Douglas. Elle s'écrit :

$$\begin{aligned} ln(cv_i) &= A_0 + ln(yf_i) + \beta_1 ln(yl_i) + \beta_2 ln(we_i) + \beta_3 ln(wm_i) \\ &\quad + \beta_4 ln(prod_i) + \beta_5 ln(stoc_i) + \beta_6 ln(long_i) + \beta_7 ln(aboon_i) + \epsilon_i \end{aligned} \quad (3.1)$$

où cv est le coût variable moyen de production obtenu en divisant le coût variable (CV) par la quantité totale d'eau produite (V_{prod}). Dans cette fonction, le volume d'eau facturé (yf) intervient de façon non pramétrique tandis que toutes les autres variables sont paramétriques.

Les résultats de l'estimation avec le package *SemiPar* se subdivisent en *composante linéaire* et *composante non paramétrique*. Ces résultats se présentent comme suit

Les variables de prix de facteur sont significatives tandis que les variables techniques ne le sont pas. L'estimation de la partie non paramétrique a nécessité l'usage de 35 noeuds, un paramètre de lissage $\lambda = 3,8$ et 4 degré de libertés. Cette partie non paramétrique peut s'interpréter comme un effet d'échelle en particulier des rendements d'échelle. La courbe est dans l'ensemble constante jusqu'au niveau de production de $yf = -2$ (environ $103963\ m^3/an$). Cela correspond à des rendements d'échelle constants. Pour $yf \in [-2, 1]$ (entre $103963\ m^3/an$ et $2,09.10^6\ m^3/an$), la courbe décroît, ce qui correspond à des rendements d'échelle croissants. A partir de $yf = 2,09.10^6\ m^3/an$, la courbe est croissante, ce qui correspond à des rendements d'échelle décroissants.

La seconde fonction est de type translog où nous n'intégrons pas tous les termes croisés possibles. Elle s'écrit :

$$\begin{aligned} ln(cv_i) &= A_0 + ln(yf_i) + \beta_2 ln(we_i) + \beta_3 ln(wm_i) \\ &\quad + \frac{1}{2}\beta_{22}\left[ln(we_i)\right]^2 + \frac{1}{2}\beta_{33}\left[ln(wm_i)\right]^2 + \beta_{24} ln(we_i) ln(yf_i) \\ &\quad + \beta_{34} ln(wm_i) ln(yf_i) + \beta_{23} ln(wm_i) ln(we_i) \\ &\quad + \beta_4 ln(prod_i) + \beta_5 ln(stoc_i) + \beta_6 ln(long_i) + \beta_7 ln(aboon_i) + \epsilon_i \end{aligned} \tag{3.2}$$

La courbe de l'estimation de la partie non paramétrique est donnée par la figure

Ici également, les variables de prix ainsi que leur carré respectif sont significatifs. L'estimation de la partie non paramétrique a nécessité l'usage de 35 noeuds, un paramètre de lissage $\lambda = 3,15$ et 4 degré de libertés. Ici par contre, les rendements d'échelle sont d'abord faiblement décroissants jusqu'à $103963\ m^3/an$, ensuite croissants jusqu'à $2,09.10^6\ m^3/an$ et au delà les rendements deviennent fortement décroissants.

La troisième fonction est la fonction translog où l'on impose l'hypothèse d'homothétie[3] à la fonction de coût. Cela revient à imposer les restrictions suivantes ($\beta_{24} = \beta_{34} = 0$). Ainsi cette fonction s'écrit :

$$\begin{aligned} ln(cv_i) &= A_0 + ln(yf_i) + \beta_2 ln(we_i) + \beta_3 ln(wm_i) \\ &\quad + \frac{1}{2}\beta_{22}\left[ln(we_i)\right]^2 + \frac{1}{2}\beta_{33}\left[ln(wm_i)\right]^2 + \beta_{23} ln(wm_i) ln(we_i) \\ &\quad + \beta_4 ln(prod_i) + \beta_5 ln(stoc_i) + \beta_6 ln(long_i) + \beta_7 ln(aboon_i) + \epsilon_i \end{aligned} \tag{3.3}$$

La courbe de la partie non paramétrique est donnée par la figure

[3]Une technologie est homothétique si les ratios des facteurs sont constants et indépendants du niveau de production et des facteurs fixes. En d'autres termes, si les paniers d'inputs x et $x^{'}$ vproduisent le même niveau de produit alors les paniers tx et $tx^{'}$ produisent également le même niveau d'output (Salvenes et Tjotta, 1994). Cette hypothèse est cruciale pour dériver des rendements d'échelle de long terme.

Table 2: Estimations paramétriques de fonctions de coût

	Modèle Translog			Modèle Translog Homothétique			Modèle Log-lineaire		
Variables	Coef.	e. t	Signif.	Coef.	e. t	Signif.	Coef.	e. t	Signif.
yf	-0,0340	0,0413		-0,0575	0,0408		-0,1068763	0,0183766	***
we	0,1062	0,0152	***	0,1166	0,0128	***	0,1456224	0,0083005	***
wm	0,6700	0,0174	***	0,6463	0,0136	***	0,6976068	0,0091753	***
abon	0,0331	0,0157	**	0,0331	0,0156	**	0,0275698	0,0189514	
prod	—	—		—	—		—	—	
stoc	—	—		—	—		—	—	
com	—	—		—	—		—	—	
yf*yf	-0,0808	0,0273	***	-0,1019	0,0269	***			
we*we	0,0504	0,0099	***	0,0522	0,0097	***			
wm*wm	0,1506	0,0128	***	0,1398	0,0122	***			
yf*yl	0,0018	0,0012		0,0020	0,0010	*			
yf*we	-0,0123	0,0063	**	—	—				
yf*wm	0,0264	0,0070	***	—	—				
yl*we	0,0055	0,0037		—	—				
yl*wm	-0,0058	0,0046		—	—				
we*wm	-0,0390	0,0101	***	-0,0355	0,0096	***			
yf*prod	0,0884	0,0316	***	0,0996	0,0317	***			
we*stoc	-0,0025	0,0057		-0,0030	0,0049				
we*com	-0,0123	0,0071	*	-0,0121	0,0067	*			
wm*stoc	-0,0218	0,0060	***	-0,0092	0,0051	*			
SCR	2,7285			1.6466			1.6938		

Significativité : à $1\%^{(***)}$, à $5\%^{(**)}$, à $10\%^{(*)}$ // SCR : Somme des carrés résiduels

On retrouve quasiment les mêmes résultats qu'avec la fonction de coût translog complet.

Ici également les variables de prix sont significatives ainsi que leur carré respectif.

La question est maintenant de savoir lequel de ces modèles est le plus approprié pour décrire les données. Cela implique de faire des tests de spécification.

3.2.2 Modèles semi-paramétriques

Nous estimons ici, les équivalent semiparamétrique des trois précédentes fonctions. La première considérée est la version semi-paramétrique d'une fonction de type Cobb-Douglas.

Elle s'écrit :

$$\begin{aligned} ln(cv_i) &= A_0 + r\left[ln(yf_i)\right] + \beta_1 ln(yl_i) + \beta_2 ln(we_i) + \beta_3 ln(wm_i) \\ &\quad + \beta_4 ln(prod_i) + \beta_5 ln(stoc_i) + \beta_6 ln(long_i) + \beta_7 ln(aboon_i) + \epsilon_i \end{aligned} \quad (3.4)$$

où cv est le coût variable moyen de production obtenu en divisant le coût variable (CV) par la quantité totale d'eau produite (V_{prod}). Dans cette fonction, le volume d'eau facturé (yf) intervient de façon non pramétrique tandis que toutes les autres variables sont paramétriques.

Les résultats de l'estimation avec le package *SemiPar* se subdivisent en *composante linéaire* et *composante non paramétrique*. Ces résultats se présentent comme suit

Les variables de prix de facteur sont significatives tandis que les variables techniques ne le sont pas. L'estimation de la partie non paramétrique a nécessité l'usage de 35 noeuds, un paramètre de lissage $\lambda = 3,8$ et 4 degré de libertés. Cette partie non paramétrique peut s'interpréter comme un effet d'échelle en particulier des rendements d'échelle. La courbe est dans l'ensemble constante jusqu'au niveau de production de $yf = -2$ (environ $103963\ m^3/an$). Cela correspond à des rendements d'échelle constants. Pour $yf \in [-2, 1]$ (entre $103963\ m^3/an$ et $2,09.10^6\ m^3/an$), la courbe décroît, ce qui correspond à des rendements d'échelle croissants. A partir de $yf = 2,09.10^6\ m^3/an$, la courbe est croissante, ce qui correspond à des rendements d'échelle décroissants.

La seconde est la version semi-paramétrique d'une fonction de type translog où nous n'intégrons pas tous les termes croisés possibles. Elle s'écrit :

$$\begin{aligned} ln(cv_i) &= A_0 + r\left[ln(yf_i)\right] + \beta_2 ln(we_i) + \beta_3 ln(wm_i) \\ &\quad + \frac{1}{2}\beta_{22}\left[ln(we_i)\right]^2 + \frac{1}{2}\beta_{33}\left[ln(wm_i)\right]^2 + \beta_{24} ln(we_i) ln(yf_i) \\ &\quad + \beta_{34} ln(wm_i) ln(yf_i) + \beta_{23} ln(wm_i) ln(we_i) \\ &\quad + \beta_4 ln(prod_i) + \beta_5 ln(stoc_i) + \beta_6 ln(long_i) + \beta_7 ln(aboon_i) + \epsilon_i \end{aligned} \quad (3.5)$$

La courbe de l'estimation de la partie non paramétrique est donnée par la figure

Ici également, les variables de prix ainsi que leur carré respectif sont significatifs. L'estimation de la partie non paramétrique a nécessité l'usage de 35 noeuds, un paramètre de lissage $\lambda = 3,15$ et 4 degré de libertés. Ici par contre, les rendements d'échelle sont d'abord faiblement décroissants jusqu'à $103963\ m^3/an$, ensuite croissants jusqu'à $2,09.10^6\ m^3/an$ et au delà les rendements deviennent fortement décroissants.

La troisième est la fonction semi-paramétrique translog où l'on impose l'hypothèse d'homothétie à la fonction de coût. Cela revient à imposer les restrictions suivantes ($\beta_{24} = \beta_{34} = 0$). Ainsi cette fonction s'écrit :

$$\begin{aligned} ln(cv_i) &= A_0 + r\left[ln(yf_i)\right] + \beta_2 ln(we_i) + \beta_3 ln(wm_i) \\ &\quad + \frac{1}{2}\beta_{22}\left[ln(we_i)\right]^2 + \frac{1}{2}\beta_{33}\left[ln(wm_i)\right]^2 + \beta_{23} ln(wm_i) ln(we_i) \\ &\quad + \beta_4 ln(prod_i) + \beta_5 ln(stoc_i) + \beta_6 ln(long_i) + \beta_7 ln(aboon_i) + \epsilon_i \end{aligned} \quad (3.6)$$

La courbe de la partie non paramétrique est donnée par la figure

On retrouve quasiment les mêmes résultats qu'avec la fonction de coût translog complet.

La dernière fonction[4] que nous avions estimé est une fonction de coût quadratique qui s'écrit

$$\begin{aligned} ln(cv_i) &= A_0 + \beta_{11} ln(yf_i) + \beta_{12} \left[ln(yf_i)\right]^2 + \beta_1 ln(yl_i) + \beta_2 ln(we_i) + \beta_3 ln(wm_i) \\ &\quad + \beta_4 ln(prod_i) + \beta_5 ln(stoc_i) + \beta_6 ln(long_i) + \beta_7 ln(aboon_i) + \epsilon_i \end{aligned} \tag{3.7}$$

Cette dernière fonction est motivée par la forme en parabole que suggère la partie non paramétrique des estimations précédentes.

Ici également les variables de prix sont significatives ainsi que leur carré respectif.

La question est maintenant de savoir lequel de ces modèles est le plus approprié pour décrire les données. Cela implique de faire des tests de spécification.

3.3 Tests de spécification

En économétrie paramétrique, il existe un éventail de tests pour sélectionner le meilleur modèle (le test du ratio de vraisemblance, le test du multiplicateur de Lagrange, le test de Fisher, etc...). Sur la base de ce dernier test, Yatchew (2003) propose des tests adaptés au cas non paramétrique.

L'idée du premier test est la suivante. Supposons que l'on veuille tester l'hypothèse nulle selon laquelle nos fonctions de coûts sont purement paramétriques. Désignons par s^2_{res} le carré des résidus de cette régression paramétrique. Si cette spécification est correcte, s^2_{res} sera une bonne approximation de σ^2_ϵ. Sinon, s^2_ϵ (le carré des résidus de la régression semi-paramétrique) demeurera le meilleur estimateur et formera ainsi la base du test. Soit la statistique suivante[5]:

$$V = \sqrt{n} \frac{\left(s^2_{res} - s^2_\epsilon\right)}{s^2_\epsilon}$$

où n est le nombre d'observations.

Yatchew (2003) montre que le numérateur de V est approximativement égale à :

$$\sqrt{n} \frac{1}{n} \sum \epsilon_i \epsilon_{i-1} \xrightarrow{loi} N\left(0, \sigma^4_\epsilon\right)$$

[4]Nous avons essayé la version semi-paramétrique d'une fonction CES mais les données ne semblent pas s'y prêtées (message d'erreur du logiciel).

[5]Cette statistique diffère légèrement de celle proposée par Yatchew qui utilise des techniques de différentiation pour calculer les résidus issus des régressions semi-paramétriques. ici, nous n'avions pas eu besoin de ces techniques car nous avons en sortie R ces résidus.

Table 3: Estimations semiparamétriques de fonctions de coût

Composante linéaire									
	Modèle Translog			Modèle Translog Homothétique			Modèle Log-linéaire		
Variables	Coef.	e. t	Signif.	Coef.	e. t	Signif.	Coef.	e. t	Signif.
constante	-0,0002	0,0027		-0,0000	0,0029		-0,0000	0,0023	
we	0,1030	0,0137	***	0,1170	0,0114	***	0,1456	0,0074	***
wm	0,6783	0,0157	***	0,6474	0,0122	***	0,6976	0,0082	***
abon	0,0248	0,0139	*	0,0210	0,0138		0,0276	0,0169	*
we*we	0,0493	0,0088	***	0,0490	0,0086	***			
wm*wm	0,1467	0,0113	***	0,1319	0,0105	***			
yf*yl	0,0015	0,0011		0,0017	0,0009	*			
yf*we	-0,0137	0,0056	***	—	—				
yf*wm	0,0265	0,0063	***	—	—				
yl*we	0,0046	0,0034		—	—				
yl*wm	-0,0032	0,0043		—	—				
we*wm	-0,0391	0,0090	***	-0,0327	0,0085	***			
yf*prod	0,0184	0,0141		0,0984	0,0138				
we*stoc	-0,0017	0,0050		-0,0033	0,0044				
we*com	-0,0126	0,0063	**	-0,0122	0,0060	**			
wm*stoc	-0,0216	0,0053	***	-0,0083	0,0045	*			
Composante non linéaire									
	Modèle Translog			Modèle Translog Homothétique			Modèle Log-linéaire		
	dl	lissage	noeuds	dl	lissage	noeuds	dl	lissage	noeuds
f(yf)	3,21	1,606	35	3,299	1,521	35	1	527,4	35
SCR									
Significativité : à $1\%^{(***)}$, à $5\%^{(**)}$, à $10\%^{(*)}$ // SCR : Somme des carrés résiduels									

Puisque s_ϵ^2, le dénominateur de V, est un estimateur convergent de σ_ϵ^2, il déduit que V suit asymptotiquement une loi normale centrée réduite sous H_0.

Le second test, est une version du test de restriction linéaire classique $R\beta = r$. Il montre que la statistique[6] suivante :

$$W = n\frac{\left(s_{\epsilon_c}^2 - s_{\epsilon_{nc}}^2\right)}{s_{\epsilon_{nc}}^2} \xrightarrow{loi} \chi^2_{rang(R)}$$

où $s_{\epsilon_{nc}}^2$ désigne le carré des résidus dans le modèle non contraint et $s_{\epsilon_c}^2$ le carré des résidus dans le modèle contraint.

Forts de ces tests, nous pouvons chercher la meilleure spécification pour nos fonctions de coûts. Commençons par comparer les deux fonctions de coûts translog (complet et homothétique). Nous avons 2 restrictions et le modèle contraint est la fonction homothétique. Ainsi :

$$W = 147\frac{(0,00720801 - 0,00719978)}{0,00719978} = 0,17$$

Puisque $\chi^2_{5\%}(2) = 5,99$, on ne rejette pas H_o. Le meilleur modèle est donc le modèle contraint, c'est-à-dire la fonction de coût translog homothétique.

Comparons ensuite, cette dernière à la fonction de coût semiparamétrique Cobb-Douglas. Le modèle contraint est ici la fonction Cobb-Douglas et nous avons 3 restrictions. Ainsi :

$$W = 147\frac{(0,009665624 - 0,00720801)}{0,00720801} = 50,12$$

Puisque $\chi^2_{5\%}(3) = 7,81$, on rejette H_o. Le meilleur modèle est donc le modèle non contraint, c'est-à-dire la fonction de coût translog homothétique. Ainsi, cette fonction de coût translog homothétique est la meilleure spécification parmi nos fonctions de coûts semiparamétriques.

Reste maintenant à comparer cette fonction à la fonction de coût paramétrique quadratique. Le modèle contraint ici est la fonction de coût quadratique. La statistique du test est donnée par :

$$V = \sqrt{174}\frac{\left(0,1034^2 - 0,00720801\right)}{0,00720801} = 6,37$$

Puisque $z_{lue}(5\%) = 1.65$, on rejette H_o. Le meilleur modèle est donc le modèle non contraint, c'est-à-dire la fonction de coût translog homothétique.

Au total, il apparaît que le modèle purement paramétrique quadratique n'est pas adapté pour ces données. Il faut donc une spécification non paramétrique et le modèle

[6]Cette statistique diffère légèrement de celle proposée par Yatchew pour les raisons évoquées plus haut.

translog homothétique est le meilleur parmi ceux que nous avons estimé pour décrire nos données.

Conclusion

Ce travail est une première exploration de la régression non paramétrique appliquée. Il a permis de passer en revue les différents outils classiques disponibles pour ce type d'analyse statistique. Il permet de rappeler comment ces outils sont mis en oeuvre en pratique sous le logiciel R.

L'objectif visé à était de se familiariser avec quelques uns des ces outils en vue d'une utilisation future. En effet, dans le prolongement de ce papier, nous envisageons d'explorer plus en profondeur les tests non paramétriques notamment ceux portant sur les restrictions que sont la concavité, la monotonicité, la symétrie etc. Plus généralement, il serait intéressant d'aller au delà de l'estimation des modèles semi-paramétriques; celles-ci pouvant être vues comme la porte d'entrée vers les modèles purement non paramétriques. Par ailleurs, nous comptons exploiter la structure en panel de notre base de données pour mieux appréhender les spécificités des différents services d'eau qui la compose..

References