MESURE DE L'EFFICIENCE DES ÉTABLISSEMENTS DE SANTÉ

REVUE ET SYNTHÈSE MÉTHODOLOGIQUE

PIERRE OUELLETTE
PATRICK PETIT

Février 2010

Centre sur la productivité et la prospérité

HEC MONTRÉAL

Créé en 2009, le Centre sur la productivité et la prospérité de HEC Montréal a une double vocation.

Le Centre se veut d'abord un organisme voué à la recherche sur la productivité et la prospérité en ayant comme objets principaux d'étude le Québec et le Canada.

Le Centre se veut également un organisme de transferts, de vulgarisation et, ultimement, d'éducation en matière de productivité et de prospérité.

Pour en apprendre davantage sur le Centre ou pour obtenir des copies supplémentaires de ce document, visitez le **www.hec.ca/cpp** ou écrivez-nous à **info.cpp@hec.ca**.

Adresse de correspondance :
Centre sur la productivité et la prospérité
HEC Montréal
3000, chemin de la Côte-Sainte-Catherine
Montréal (Québec) Canada H3T 2A7

Téléphone : 514-340-6449
Télécopieur : 514-340-6469

Cette publication a bénéficié du soutien financier du ministère des Finances du Québec.

Remerciements

Au début de la recherche menant à ce rapport, nous avons bénéficié, suite à un appel à tous, de l'aide des membres du *Productivity Analysis Research Network* (PARN). Tout particulièrement, V. Valdmanis, M.D. Rosko et B. Hollingsworth nous ont transmis nombre de leurs travaux. Nous tenons à les remercier pour leur aide. Nous remercions aussi le Centre sur la productivité et la prospérité - HEC Montréal pour son support financier.

* Mise en garde: "Les analyses et opinions exprimées dans ce document sont celles de l'auteur et ne représentent pas nécessairement celles du FMI ou la politique du FMI."

MESURE DE L'EFFICIENCE DES ÉTABLISSEMENTS DE SANTÉ REVUE ET SYNTHÈSE MÉTHODOLOGIQUE

PIERRE OUELLETTE
Université du Québec à Montréal

PATRICK PETIT
Fonds monétaire international (FMI)

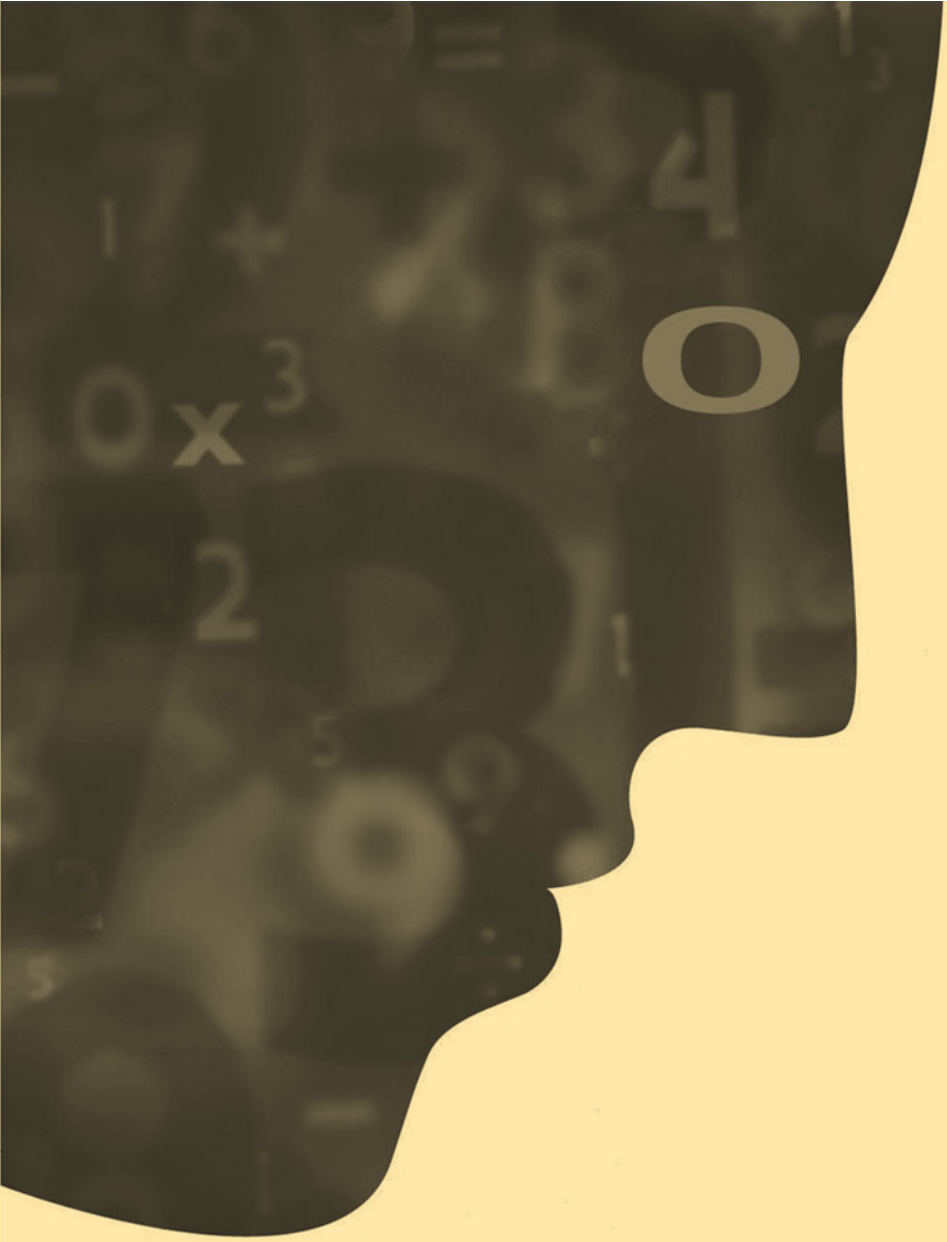
Résumé

La mesure de l'efficience a constitué un objectif de recherche majeur en économie de la santé depuis les 25 dernières années. Presque toutes les études ont souscrit à l'exigence formulée par Cowing et Stevenson en 1983 de baser leur approche sur des fondements théoriques solides tirés de l'économie de la production. À partir de ces fondements, plusieurs méthodes ont été développées et utilisées, certaines puisant dans les méthodes statistiques, d'autres en recherche opérationnelle et certaines dans des méthodes comptables. Notre objectif est de montrer que ces méthodes ont souvent le défaut de ne pas avoir pris en compte *l'ensemble* des prescriptions de la théorie. Par exemple, les auteurs ayant basé leur approche sur la théorie économique ont superposé une structure stochastique de termes d'erreur qui est parfois incompatible avec certaines propriétés de la théorie. En fait, presque tous les modèles peuvent être vus comme des cas particuliers d'un modèle général. Nous montrerons qu'à chacun de ces modèles correspond un ensemble d'hypothèses sur la nature des données et que dans certains cas, les modèles sont incohérents.

Abstract

Efficiency measurement has been a major item on the health economics agenda over the past quarter century. A thorough review of the literature shows that almost all studies met the basic requirements proposed by Cowing and Stevenson in 1983, and relied on the solid theoretical foundations of production economics. Many methods were nevertheless developed and used, with some grounded in statistics, others in operations research, or accounting. The objective of this paper is to show how these methods often fail to include all relevant theoretical considerations. For example, authors relying on economic theory have used empirical methods with stochastic error terms that are sometimes at odds with certain properties of their models. In fact, almost all models can be approached as specific cases of a general model. We will show that each model implies specific assumptions on the data, and that in some cases, the models are incoherent.

Table des matières



Introduction

Introduction

Les pressions sur les budgets gouvernementaux causées par la croissance des dépenses en santé expliquent l'intérêt que les chercheurs ont porté à l'efficience du système de santé. Si on se concentre sur les contributions des articles ayant mesuré l'efficience des institutions de santé on se rend compte que tous ces articles essayaient de répondre à l'une des questions suivantes ou les deux en même temps :

1. Quel doit être le budget d'un hôpital (ou autre institution de santé) ?
2. Quel doit être le tarif des actes médicaux ?

En fait, ces deux questions sont complémentaires. Toutes deux reposent sur la notion de coût optimal pour ce qui est produit étant donné l'environnement de cet hôpital. Ce qui les distingue, c'est le niveau d'agrégation. Dans un cas, on est au niveau de l'institution; dans l'autre, on est au niveau du service direct.

La réponse à la première question est donnée par la fonction de coût $C(w, y)$,[1] alors que la deuxième est déterminée par les propriétés de la fonction de coût. En optimum de premier rang, la tarif de l'output y_i, noté p_i, est égal au coût marginal de cet output $p_i = \frac{\partial C}{\partial y_i}$ ou encore, pour l'optimum de second rang (par exemple dans le cas où les firmes sont astreintes à l'équilibre budgétaire - aussi appelé un optimum de Boîteux-Ramsay), par le coût marginal corrigé par les élasticités des demandes.[2] Quoi qu'il en soit, au minimum, il faut connaître la fonction de coût. La grande difficulté est que cette fonction de coût est inobservable directement et qu'il faut l'inférer à partir des coûts observés. Le but de ce qui suit est justement de montrer comment on peut procéder pour récupérer la fonction de coût à partir des données disponibles sur les prix et les quantités.

[1] La notation retenue est la suivante : *C* est le coût total, w est le vecteur des prix des inputs (dont la quantité est notée x) et y est le vecteur des outputs produits. Notre but étant de procéder à un survol méthodologique, nous adopterons volontairement la forme la plus simple de fonction de coût, soit la fonction de coût total non réglementée, dans le but de simplifier la présentation. Il est possible de généraliser considérablement l'environnement de la firme en introduisant des inputs quasi-fixes ou fixes (aussi appelés inputs non discrétionnaires), le cadre règlementaire, le niveau technologique, etc. La question de la qualité des inputs et surtout des outputs sera traitée plus loin.

2 Bien que ne découlant pas de la théorie de l'optimum, la tarification au coût moyen peut aussi être envisagée par les organismes gouvernementaux ne serait-ce que pour sa simplicité.

Lien entre la fonction du coût théorique et les coûts observés

I. Lien entre la fonction du coût théorique et les coûts observés

Dans cette section, notre but est de montrer clairement les liens entre les coûts observés et les coûts minimaux. Ce n'est que dans la suite que nous ferons état des méthodes permettant de faire le lien empiriquement.

Le problème de minimisation des coûts s'écrit :

$$C(w,y) \triangleq \min_{x} \{w'x : f(y,x) \leq 0\}.$$

La solution de ce problème, si elle existe (ce que nous supposerons dans la suite), est donnée par le vecteur des demandes conditionnelles de facteurs :

$$x = x(w,y).$$

La valeur accordée par la fonction de coût est en fait le budget minimum qu'il faudrait donner à une institution de santé la plus efficace possible pour qu'elle produise le vecteur de services de santé y, étant donné les prix w des inputs x, au moyen d'une technologie représentée par la fonction de production f.

Pour sa part, le coût observé est donné par :[3]

$$C^{obs} \triangleq \sum_{i=1}^{n} w_i^{obs} x_i^{obs} = w^{obs}{}' x^{obs}.$$

Nous voulons établir le lien entre le coût observé et le coût théorique. D'une certaine façon, la question est de savoir si le coût observé correspond au minimum nécessaire pour que l'institution rende les services attendus. Autrement dit, il faut mesurer l'écart entre le coût théorique minimum tel que déterminé par la fonction de coût et le coût observé.

Les écarts entre les diverses variables en question peuvent provenir de divers types d'erreurs :

1. Erreurs d'observation;
2. Erreurs d'optimisation.

[3] De façon générale, l'indice supérieur *obs* fait référence aux variables observées.

Le terme « erreur d'optimisation » est en soi un sujet de réflexion. Qu'entend-on au juste par cette expression ? S'agit-il de simples mauvaises décisions aléatoires et qui ne se reproduisent pas ? Ou alors de décisions qui sont systématiquement sous-optimales parce que le système d'incitations est déficient ? Dans le premier cas, on ne voit pas trop comment corriger la situation et il faut espérer que ces mauvaises décisions aléatoires ne sont pas trop fréquentes, alors que dans le deuxième on peut penser que les gestionnaires réagissent optimalement à des incitations inadéquates et que c'est ce qui explique les écarts de coût. Ma (1994) a montré qu'il peut être optimal pour les hôpitaux (mais sous-optimal pour la société) de ne pas gérer efficacement leurs ressources selon le mode budgétaire auquel ils sont soumis et que le mode budgétaire leur permet d'extraire une rente.[4] Le lien entre l'inefficience et, d'autre part, l'environnement économique et réglementaire (le type de budgétisation) est tellement intime qu'on ne peut penser à l'un sans penser à l'autre. La littérature économique a retenu deux approches principales pour modéliser les erreurs, soit les modèles additif et multiplicatif.

I.1. Modèle additif

Il y a trois catégories de variables : les quantités d'inputs x, leur prix w, et les outputs y. Les erreurs d'observation sur les prix ε_w expliquent à elles seules la différence entre les prix observés et les prix utilisés par les organisations :

$$w^{obs} = w + \varepsilon_w.$$

Cependant, les écarts dans les quantités de facteurs peuvent découler à la fois des erreurs d'observation ε_x et des erreurs d'optimisation ν_x :

$$x^{obs} = x + \varepsilon_x + \nu_x.$$

Nous reviendrons plus tard sur la question des distributions statistiques des termes d'erreur. À ce stade, il est évident que les erreurs d'observation et d'optimisation expliquent les écarts entre le coût observé C^{obs} et le coût minimal C La relation s'obtient comme suit :

[4] Dans la terminologie de Ma, l'inefficience est causée par un effort sous-optimal en terme de réduction des coûts. Cet effort dépend de la pénibilité de l'effort de l'équipe de gestionnaires.

$$
\begin{aligned}
C &= C(w,y) \\
&= w'x(w,y) \\
&= \left(w+\left(w^{obs}-w^{obs}\right)\right)'\left(x(w,y)+\left(x^{obs}-x^{obs}\right)\right) \\
&= \left(w^{obs}-\left(w^{obs}-w\right)\right)'\left(x^{obs}-\left(x^{obs}-x(w,y)\right)\right) \\
&= w^{obs}{}'x^{obs}-w^{obs}{}'\left(x^{obs}-x(w,y)\right)-x^{obs}{}'\left(w^{obs}-w\right)+\left(w^{obs}-w\right)'\left(x^{obs}-x(w,y)\right) \\
&= C^{obs}-w^{obs}{}'\left(\varepsilon_x+\nu_x\right)-x^{obs}{}'\varepsilon_w+\varepsilon_w{}'\left(\varepsilon_x+\nu_x\right),
\end{aligned}
$$

d'où :

$$
\begin{aligned}
C^{obs} &= C(w,y)+w^{obs}{}'\left(\varepsilon_x+\nu_x\right)+x^{obs}{}'\varepsilon_w-\varepsilon_w{}'\left(\varepsilon_x+\nu_x\right) \\
&= C(w,y)+\underbrace{\left[w^{obs}{}'\varepsilon_x+x^{obs}{}'\varepsilon_w-\varepsilon_w{}'\varepsilon_x\right]}_{\varepsilon_C}+\underbrace{\left(w^{obs}-\varepsilon_w\right)'\nu_x}_{\nu_C}
\end{aligned}
$$

ce qui implique que l'écart entre le coût observé et le coût minimum, noté μ , est :

$$
\begin{aligned}
\mu &= \underbrace{\left[w^{obs}{}'\varepsilon_x+x^{obs}{}'\varepsilon_w-\varepsilon_w{}'\varepsilon_x\right]}_{\varepsilon_C}+\underbrace{\left(w^{obs}-\varepsilon_w\right)'\nu_x}_{\nu_C} \\
&= \varepsilon_C+\nu_C.
\end{aligned}
$$

Il reste à introduire les erreurs sur l'output. Pour l'instant, nous ne considérerons que les erreurs de mesure sur l'output :

$$
y^{obs} = y+\varepsilon_y \leftrightarrow y^{obs}-\varepsilon_y = y
$$

et ainsi :

$$
C^{obs} = C\left(w^{obs}-\varepsilon_x, y^{obs}-\varepsilon_y\right)+\underbrace{\left[w^{obs}{}'\varepsilon_x+x^{obs}{}'\varepsilon_w-\varepsilon_w{}'\varepsilon_x\right]}_{\varepsilon_C}+\underbrace{\left(w^{obs}-\varepsilon_w\right)'\nu_x}_{\nu_C}
$$

et

$$
x_i^{obs} = x_i\left(w^{obs}-\varepsilon_x, y^{obs}-\varepsilon_y\right)+\left(\varepsilon_{x_i}+\nu_{x_i}\right), \forall i = 1,\ldots,n.
$$

On remarque que la relation de Shephard tient à la fois pour le coût minimum (une simple application du lemme de l'enveloppe) et le coût observé :

$$\frac{\partial C(w,y)}{\partial w} = x(w,y) \text{ et } \frac{\partial C^{obs}}{\partial w^{obs}} = x^{obs}.$$

Plus important, le terme d'inefficience dépend obligatoirement des prix des inputs. Cette caractéristique a quasi-systématiquement été ignorée dans les travaux économétriques.

I.2. Modèle multiplicatif

Dans ce modèle, les erreurs, au lieu d'être additives, sont multiplicatives. On a :

$$\begin{aligned} w^{obs} &= we^{\varepsilon_w} \leftrightarrow w^{obs}e^{-\varepsilon_w} = w \\ x^{obs} &= xe^{\varepsilon_x+\nu_x} \leftrightarrow x^{obs}e^{-(\varepsilon_x+\nu_x)} = x \\ y^{obs} &= ye^{\varepsilon_y} \leftrightarrow y^{obs}e^{-\varepsilon_y} = y. \end{aligned}$$

On procède comme précédemment pour obtenir la relation entre le coût minimum et le coût observé :

$$\begin{aligned} C &= C(w,y) \\ &= w^{obs\prime}x^{obs} - w^{obs\prime}\left(x^{obs} - x(w,y)\right) - x^{obs\prime}\left(w^{obs} - w\right) + \left(w^{obs} - w\right)'\left(x^{obs} - x(w,y)\right) \\ &= w^{obs\prime}x^{obs} - w^{obs\prime}\left(x^{obs} - x^{obs}e^{-(\varepsilon_x+\nu_x)}\right) - x^{obs\prime}\left(w^{obs} - w^{obs}e^{-\varepsilon_w}\right) + \left(w^{obs} - w^{obs}e^{-\varepsilon_w}\right)'\left(x^{obs} - x^{obs}e^{-(\varepsilon_x+\nu_x)}\right) \\ &= w^{obs\prime}x^{obs} - w^{obs\prime}x^{obs}\left(1 - e^{-(\varepsilon_x+\nu_x)}\right) - x^{obs\prime}w^{obs}\left(1 - e^{-\varepsilon_w}\right) + \left(w^{obs} - w^{obs}e^{-\varepsilon_w}\right)'\left(x^{obs} - x^{obs}e^{-(\varepsilon_x+\nu_x)}\right) \\ &= w^{obs}e^{-\varepsilon_w\prime}x^{obs}e^{-(\varepsilon_x+\nu_x)} = \sum_i^n w_i^{obs}x_i^{obs}e^{-\left(\varepsilon_{w_i}+\varepsilon_{x_i}+\nu_{x_i}\right)}. \end{aligned}$$

Ce dernier résultat peut être réécrit comme suit :

$$C^{obs} = \frac{C\left(w^{obs}e^{-\varepsilon_w}, y^{obs}e^{-\varepsilon_y}\right)}{\sum_i^n \frac{w_i^{obs}x_i^{obs}}{C^{obs}}e^{-\left(\varepsilon_{w_i}+\varepsilon_{x_i}+\nu_{x_i}\right)}},$$

ou encore, après transformation logarithmique :

$$\ln C^{obs} = \ln C\left(w^{obs}e^{-\varepsilon_w}, y^{obs}e^{-\varepsilon_y}\right) - \ln \sum_i \frac{w_i^{obs}x_i^{obs}}{C^{obs}} e^{-\left(\varepsilon_{w_i}+\varepsilon_{x_i}+v_{x_i}\right)}.$$

S'il n'y avait qu'un input $(i = 1)$, on retrouverait un modèle où le terme d'erreur est additif après transformation logarithmique :

$$\ln C^{obs} = \ln C\left(w^{obs}e^{-\varepsilon_w}, y^{obs}e^{-\varepsilon_y}\right) + \underbrace{\left(\varepsilon_w + \varepsilon_x\right)}_{\varepsilon_C^m} + \underbrace{v_x}_{v_C^m}.$$

En présence de plusieurs inputs, le modèle double-log avec terme d'erreur additif n'est plus valable. Cette remarque met à mal la plupart des modèles de frontières stochastiques de coût où les termes d'erreur sont additifs sans tenir compte de la composition exprimée précédemment. En fait, ce qui est perdu, c'est la cohérence du modèle qui établit une relation entre le coût et les demandes/parts de facteur.[5]

Le système de parts, notées S_i, est obtenu comme suit :

$$S_i = \frac{w_i x_i}{C} = \frac{w_i^{obs}x_i^{obs}e^{-\left(\varepsilon_{w_i}+\varepsilon_{x_i}+v_{x_i}\right)}}{C} = \frac{w_i^{obs}x_i^{obs}}{C^{obs}} \times e^{-\left(\varepsilon_{w_i}+\varepsilon_{x_i}+v_{x_i}\right)} \times \frac{C^{obs}}{C} = S_i^{obs} \times e^{-\left(\varepsilon_{w_i}+\varepsilon_{x_i}+v_{x_i}\right)} \times \frac{C^{obs}}{C}$$

$$\leftrightarrow$$

$$\begin{aligned}
S_i^{obs} &= S_i\left(w^{obs}e^{-\varepsilon_w}, y^{obs}e^{-\varepsilon_y}\right) \times e^{\left(\varepsilon_{w_i}+\varepsilon_{x_i}+v_{x_i}\right)} \times \frac{C}{C^{obs}} \\
&= S_i\left(w^{obs}e^{-\varepsilon_w}, y^{obs}e^{-\varepsilon_y}\right) \times e^{\left(\varepsilon_{w_i}+\varepsilon_{x_i}+v_{x_i}\right)} \times \sum_j \frac{w_j^{obs}x_j^{obs}}{C^{obs}} e^{-\left(\varepsilon_{w_j}+\varepsilon_{x_j}+v_{x_j}\right)} \\
&= S_i\left(w^{obs}e^{-\varepsilon_w}, y^{obs}e^{-\varepsilon_y}\right) \times e^{\left(\varepsilon_{w_i}+\varepsilon_{x_i}+v_{x_i}\right)} \times \sum_j S_j^{obs} e^{-\left(\varepsilon_{w_j}+\varepsilon_{x_j}+v_{x_j}\right)} \\
&= S_i\left(w^{obs}e^{-\varepsilon_w}, y^{obs}e^{-\varepsilon_y}\right) \times \sum_j \left(S_j^{obs} \times e^{\left(\varepsilon_{w_i}+\varepsilon_{x_i}+v_{x_i}\right)-\left(\varepsilon_{w_j}+\varepsilon_{x_j}+v_{x_j}\right)} \right).
\end{aligned}$$

[5] Ce manque de cohérence a déjà été mentionné par McElroy (1987) mais sans référence aux termes d'inefficience ni aux termes d'erreur de mesure sur w et y.

Puisque $S_i = \frac{\partial \ln C\left(w^{obs}e^{-\varepsilon_w}, y^{obs}e^{-\varepsilon_y}\right)}{\partial w_i} w_i = \frac{\partial \ln C\left(w^{obs}e^{-\varepsilon_w}, y^{obs}e^{-\varepsilon_y}\right)}{\partial w_i^{obs}} w_i^{obs}$, alors :

$$S_i^{obs} = \frac{\partial \ln C\left(w^{obs}e^{-\varepsilon_w}, y^{obs}e^{-\varepsilon_y}\right)}{\partial \ln w_i^{obs}} \times \sum_j \left(S_j^{obs} \times e^{\left(\varepsilon_{w_i}+\varepsilon_{x_i}+\nu_{x_i}\right)-\left(\varepsilon_{w_j}+\varepsilon_{x_j}+\nu_{x_j}\right)} \right).$$

En fait, le système de parts est un système d'équations linéaires en S_i^{obs} que l'on peut solutionner afin d'isoler les parts observées. Ce système est de rang $n-1$ à cause de l'additivité des parts qui somment à un par définition. Dans le cas le plus simple où $n = 2$, on obtient :

$$S_1^{obs} = \frac{S_1 e^{\left(\varepsilon_{w_1}+\varepsilon_{x_1}+\nu_{x_1}\right)-\left(\varepsilon_{w_2}+\varepsilon_{x_2}+\nu_{x_2}\right)}}{1-S_1+S_1 e^{\left(\varepsilon_{w_1}+\varepsilon_{x_1}+\nu_{x_1}\right)-\left(\varepsilon_{w_2}+\varepsilon_{x_2}+\nu_{x_2}\right)}}.$$

Même dans ce cas simple, la formulation est d'une grande complexité.

Finalement, on a le système coût/parts suivant :

$$\ln C^{obs} = \ln C\left(w^{obs}e^{-\varepsilon_w}, y^{obs}e^{-\varepsilon_y}\right) - \ln \sum_j S_j^{obs} e^{-\left(\varepsilon_{w_j}+\varepsilon_{x_j}+\nu_{x_j}\right)}$$

et

$$S_i^{obs} = \frac{\partial \ln C\left(w^{obs}e^{-\varepsilon_w}, y^{obs}e^{-\varepsilon_y}\right)}{\partial \ln w_i^{obs}} \times \sum_j \left(S_j^{obs} \times e^{\left(\varepsilon_{w_i}+\varepsilon_{x_i}+\nu_{x_i}\right)-\left(\varepsilon_{w_j}+\varepsilon_{x_j}+\nu_{x_j}\right)} \right)$$

$$\leftrightarrow \ln S_i^{obs} = \ln S_i\left(w^{obs}e^{-\varepsilon_w}, y^{obs}e^{-\varepsilon_y}\right) + \left(\varepsilon_{w_i}+\varepsilon_{x_i}+\nu_{x_i}\right) + \ln \sum_j S_j^{obs} e^{-\left(\varepsilon_{w_j}+\varepsilon_{x_j}+\nu_{x_j}\right)}.$$

Cette formulation est extrêmement complexe car le coût observé et les parts (les variables dépendantes) se retrouvent des deux côtés des relations. On peut cependant simplifier le système de parts en prenant les ratios de parts :

$$\frac{S_i^{obs}}{S_n^{obs}} = \frac{S_i\left(w^{obs}e^{-\varepsilon_w}, y^{obs}e^{-\varepsilon_y}\right)}{S_n\left(w^{obs}e^{-\varepsilon_w}, y^{obs}e^{-\varepsilon_y}\right)} e^{\left(\varepsilon_{w_i}+\varepsilon_{x_i}+\nu_{x_i}\right)-\left(\varepsilon_{w_n}+\varepsilon_{x_n}+\nu_{x_n}\right)}, \forall i = 1,\ldots,n$$

$$\leftrightarrow \ln\frac{S_i^{obs}}{S_n^{obs}} = \ln\frac{S_i\left(w^{obs}e^{-\varepsilon_w}, y^{obs}e^{-\varepsilon_y}\right)}{S_n\left(w^{obs}e^{-\varepsilon_w}, y^{obs}e^{-\varepsilon_y}\right)} + \left(\varepsilon_{w_i}+\varepsilon_{x_i}+\nu_{x_i}\right)-\left(\varepsilon_{w_n}+\varepsilon_{x_n}+\nu_{x_n}\right), \forall i = 1,\ldots,n-1.$$

Cela ne simplifie pas la tâche pour l'équation de coût et si on se contente d'estimer uniquement le système de parts, on ne peut récupérer l'information nécessaire pour mesurer l'inefficience absolue car, à moins de supposer que l'un des termes d'inefficience ne soit nul, on ne peut récupérer que $(n\text{-}1)$ termes d'inefficience $\left(\nu_{x_i}-\nu_{x_n}\right), \forall i = 1,\ldots,n-1.$

Pour régler ce problème, on peut ajouter des hypothèses.

Par exemple, si on élimine les erreurs sur les prix et quantités des inputs, $\varepsilon_{w_i} = \varepsilon_{x_i} = 0, \forall i,$ et si on suppose que tous les inputs sont également inefficients, $\nu_{x_i} = \nu_C, \forall i,$ alors :

$$\ln C^{obs} = \ln C\left(w^{obs}, y^{obs}e^{-\varepsilon_y}\right) + \nu_C$$

$$S_i^{obs} = \frac{\partial \ln C\left(w^{obs}, y^{obs}e^{-\varepsilon_y}\right)}{\partial \ln w_i}.$$

L'absence de termes d'erreur sur les parts (sauf pour la mesure de l'output) est naturellement plutôt restrictif et semble à ce stade éliminatoire. On peut amoindrir l'hypothèse sur les erreurs en supposant que $\varepsilon_{w_i} + \varepsilon_{x_i} = \varepsilon_C, \forall i,$ tout en maintenant $\nu_{x_i} = \nu_C, \forall i,$ et ainsi :

$$\ln C^{obs} = \ln C\left(w^{obs}e^{-\varepsilon_w}, y^{obs}e^{-\varepsilon_y}\right) + \varepsilon_C + \nu_C$$

$$S_i^{obs} = \frac{\partial \ln C\left(w^{obs}e^{-\varepsilon_w}, y^{obs}e^{-\varepsilon_y}\right)}{\partial \ln w_i^{obs}}.$$

À titre d'illustration, prenons le cas Cobb-Douglas. On aura :

$$\begin{aligned}\ln C^{obs} &= a + b_1 \ln\left(w_1^{obs} e^{-\varepsilon_{w_1}}\right) + b_2 \ln\left(w_2^{obs} e^{-\varepsilon_{w_2}}\right) + b_y \ln\left(y^{obs} e^{-\varepsilon_y}\right) + \varepsilon_C + \nu_C \\ &= a + b_1 \ln w_1^{obs} + b_2 \ln w_2^{obs} + b_y \ln y^{obs} - \left(b_1\varepsilon_{w_1} + b_2\varepsilon_{w_2} + b_y\varepsilon_y\right) + \varepsilon_C + \nu_C \\ S_i^{obs} &= b_i.\end{aligned}$$

On remarque que le système de parts (qui n'admet pas d'erreurs et les parts doivent être constantes) donne immédiatement les coefficients $\left(b_1, b_2\right)$ et l'équation de coût devient :

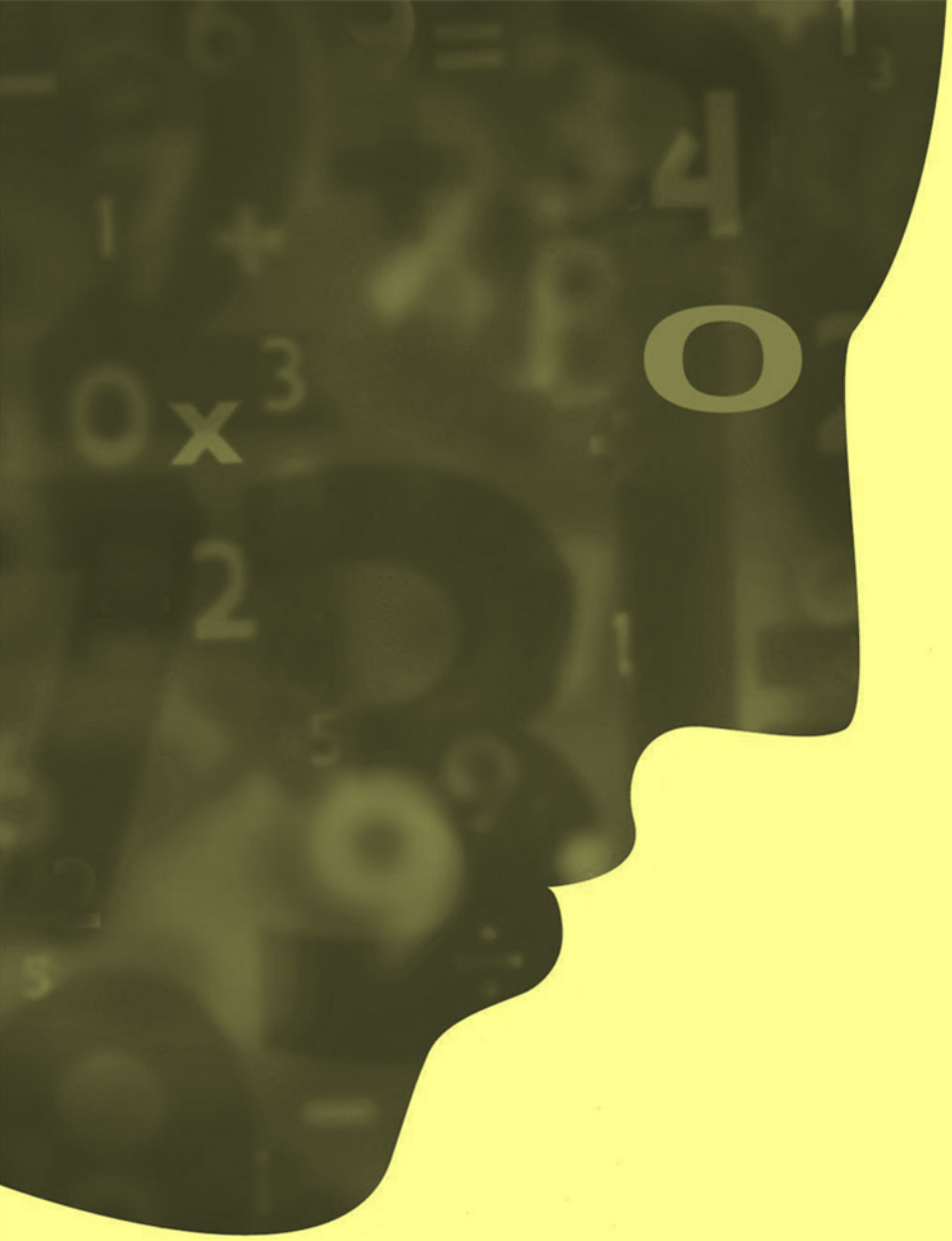
$$\ln C^{obs} - S_1^{obs} \ln w_1^{obs} - S_2^{obs} \ln w_2^{obs} = a + b_y \ln y^{obs} - \left(S_1^{obs}\varepsilon_{w_1} + S_2^{obs}\varepsilon_{w_2} + b_y\varepsilon_y\right) + \varepsilon_C + \nu_C.$$

Il reste à estimer $\left(a, b_y\right)$ en tenant compte de la dépendance du terme d'erreur par rapport aux parts et au coefficient de l'output.

Si on remplace l'hypothèse sur les erreurs de mesure sur les prix et les quantités d'inputs $\varepsilon_{w_i} + \varepsilon_{x_i} = \varepsilon_C, \forall i,$ par $\varepsilon_{w_i} = \varepsilon_w, \varepsilon_{x_i} = \varepsilon_x, \forall i,$ on obtient :

$$\begin{aligned}\ln C^{obs} - S_1^{obs} \ln w_1^{obs} - S_2^{obs} \ln w_2^{obs} &= a + b_y \ln y^{obs} - \left(S_1^{obs}\varepsilon_w + S_2^{obs}\varepsilon_w + b_y\varepsilon_y\right) + \varepsilon_w + \varepsilon_x + \nu_C \\ &= a + b_y \ln y^{obs} + \left(\varepsilon_x - b_y\varepsilon_y\right) + \nu_C.\end{aligned}$$

Ce qui ressort des deux modèles (additif et multiplicatif) est que les hypothèses sur les termes d'erreur des variables à un impact immédiat sur la formulation des termes d'erreur des équations. La cohérence d'ensemble (incluant le lemme de Shephard) a des implications importantes pour les méthodes de mesure empirique de la fonction de coût, qui sont décrites dans les trois prochaines sections.



Les modèles statistiques

II. Les modèles statistiques

II.1. Les frontières stochastiques

Traditionnellement, la méthode des frontières stochastiques se limite à la seule fonction de coût. Le modèle additif est :

$$C^{obs} = C\left(w^{obs} - \varepsilon_w, y^{obs} - \varepsilon_y\right) + \underbrace{\left[w^{obs}{}'\varepsilon_x + x^{obs}{}'\varepsilon_w - \varepsilon_w{}'\varepsilon_x\right]}_{\varepsilon_C} + \underbrace{\left(w^{obs} - \varepsilon_w\right)'\nu_x}_{\nu_C}.$$

Les hypothèses usuelles sont :

$$\begin{aligned}
&\varepsilon_y = 0,\\
&\left[w^{obs}{}'\varepsilon_x + x^{obs}{}'\varepsilon_w - \varepsilon_w{}'\varepsilon_x\right] = \mu_C \sim \aleph\left(0, \sigma_\mu^2\right),\\
&0 \le \left(w^{obs} - \varepsilon_w\right)'\nu_x = \nu_C \sim ?\left(\overline{\nu}_C, \sigma_\nu^2\right).
\end{aligned}$$

Pour l'instant, le choix de la distribution du terme d'inefficience (indiquée par « ? ») est laissée de côté. Le traitement des erreurs de mesures et d'optimisation est délicat. Il est impensable de supposer que les termes d'erreur soient tels que cette expression puisse être traitée comme indépendante des prix et quantités observés. Cette dépendance n'a pas retenu l'attention des chercheurs.

Si on ajoute l'hypothèse que les prix sont observés sans erreur, $w = w^{obs} \leftrightarrow \varepsilon_w = 0$, alors l'erreur de mesure sur le coût devient $w^{obs}{}'\varepsilon_x = \mu_C \sim \aleph\left(0, \sigma_\mu^2\right)$ et le terme d'inefficience est donné par $0 \le w^{obs}{}'\nu_x = \nu_C \sim ?\left(\overline{\nu}_C, \sigma_\nu^2\right)$. La fonction de coût se simplifie :

$$C^{obs} = C\left(w^{obs}, y^{obs}\right) + \underbrace{w^{obs}{}'\varepsilon_x}_{\varepsilon_C} + \underbrace{w^{obs}{}'\nu_x}_{\nu_C}.$$

Cette hypothèse n'élimine pas la dépendance des termes d'erreur et d'inefficience par rapport aux prix des inputs. Néanmoins, si on est disposé à accepter cette hypothèse, alors il faut encore imposer une forme fonctionnelle à la fonction de coût, $C\left(w^{obs}, y^{obs}; \beta\right)$, où β est le vecteur de paramètres à estimer.

Les résultats obtenus sont conditionnels aux hypothèses retenues. Le choix de la forme fonctionnelle n'est pas anodin (voir Gagné et Ouellette, 1998 et 2002). De même, le choix de la distribution pour le terme d'inefficience (représenté par le « ? » ci-haut) semble être problématique et influencer les résultats.

Cette méthode requiert la validation de la fonction de coût estimée qui doit respecter des propriétés de monotonicité en (w, y) (incluant le lemme de Shephard), de courbure (concavité en w) et d'homogénéité de degré un en w. Certaines de ses propriétés peuvent être imposées (par exemple, les contraintes sous forme d'égalité comme l'homogénéité) alors que d'autres ne seront que testées (par exemple, les propriétés sous forme d'inégalités comme la concavité dans les prix des inputs).

Un développement fréquemment utilisé dans le domaine de l'efficience des institutions de santé consiste à endogénéiser les termes d'inefficience. Cela revient à invoquer des arguments expliquant la présence d'un terme d'inefficience. En fait, quel que soit le modèle invoqué, on se ramène à :

$$\nu_C = \nu_C(Z),$$

où Z est le vecteur des déterminants de l'inefficience. Par exemple, on peut invoquer la force de la concurrence (indice de Gini, indice de Herfindhal, proximité des concurrents), la réglementation, le type de budgétisation, les idiosyncrasies des institutions en question (hôpital universitaire ou non), le type de clientèle desservie (pourcentage de patients Medicare ou Medicaid, les soins de charité), etc.

On peut procéder en deux étapes ou en une seule. Dans la démarche à une étape, on commence par estimer l'inefficience, puis on régresse cette mesure sur des déterminants. Dans la démarche à une étape, l'estimation de l'impact des déterminants se fait simultanément avec celle des coefficients technologiques. Dans tous les cas, on obtient une mesure du terme d'inefficience ν_C, défini comme l'espérance de la distance entre le point observé et la frontière de coût conditionnellement au terme d'erreur sur la mesure μ_C.

L'endogénéisation du terme d'inefficience peut être perçue comme une avancée. Cependant, nous soumettons que cette endogénéisation peut en fait constituer une erreur de spécification. Par exemple, invoquer la règlementation comme source d'inefficience est en fait une indication

que l'appareil au complet doit être modifié pour en tenir compte. La fonction de coût doit tenir compte de la règlementation qui limite les choix des institutions et il faut spécifier directement une fonction de coût réglementée et non utiliser une fonction de coût non réglementée puis corriger le terme d'inefficience en endogénéisant ce terme en fonction de la réglementation. La réglementation modifie l'ensemble des relations entre les inputs et les outputs et pas seulement la décomposition de la mesure de l'inefficience.[6]

Naturellement, cette discussion sur l'estimation des seules fonctions de coût ne doit pas faire oublier que l'estimation conjointe des fonctions de coût et des demandes de facteurs (ou les parts correspondantes) est préférable et plus *efficiente* en termes économétriques.

II.2. Méthode des moindres carrés corrigés

Une méthode relativement simple de mesurer et de comparer les niveaux d'efficience consiste à supposer que les différentes firmes ont une même technologie à un terme additif près et que ce terme représente les différences d'efficience. La technologie de la firme i est :

$$C_i^{obs} = \eta_i + C\left(w_i^{obs}, y_i^{obs}\right) + \varepsilon_{C_i}.$$

Cela revient à supposer que :

$$\varepsilon_{C_i} \equiv \left[w_i^{obs}{}' \varepsilon_{x_i} + x_i^{obs}{}' \varepsilon_{w_i} - \varepsilon_{w_i}{}' \varepsilon_{x_i} \right]$$
$$\nu_{C_i} \equiv \left(w_i^{obs} - \varepsilon_{w_i} \right)' \nu_{x_i} \triangleq \eta_i.$$

La dépendance des termes ε_{C_i} et η_i par rapport aux prix et quantités des inputs n'est pas prise en compte. Si la forme fonctionnelle comporte une constante, disons η_0, il faut normaliser le paramètre d'efficience en fixant l'un d'eux à 0 :[7]

$$C_i^{obs} = \eta_i + \left[\eta_0 + C^* \left(w_i^{obs}, y_i^{obs} \right) \right] + \varepsilon_{C_i}.$$

[6] Sur la fonction de coût réglementée, on peut consulter Färe et Logan (1983), Ouellette et Vigeant (2001a et b, et 2010).

[7] Naturellement, ce modèle, tout comme les frontières stochastiques, peut être estimé avec des méthodes de panel.

On estime en incorporant une variable binaire pour chaque firme (sauf une) pour récupérer les termes d'efficience. Si la plus petite valeur est positive, alors la firme qui n'a pas de binaire est la plus efficiente (toutes les autres ont une structure de coût plus élevée). Sinon, on procède comme suit :

$$\eta_i^{corr} = \eta_i - \min\left\{\eta_1, \ldots, \eta_n\right\}.$$

La firme qui a le plus petit η_i se verra accorder une valeur nulle et les autres auront toutes une valeur positive. Les paramètres η_i représentent les écarts d'efficience (mesurés en termes de coût) entre firmes.

On peut aussi utiliser un modèle multiplicatif (ou additif en log) :

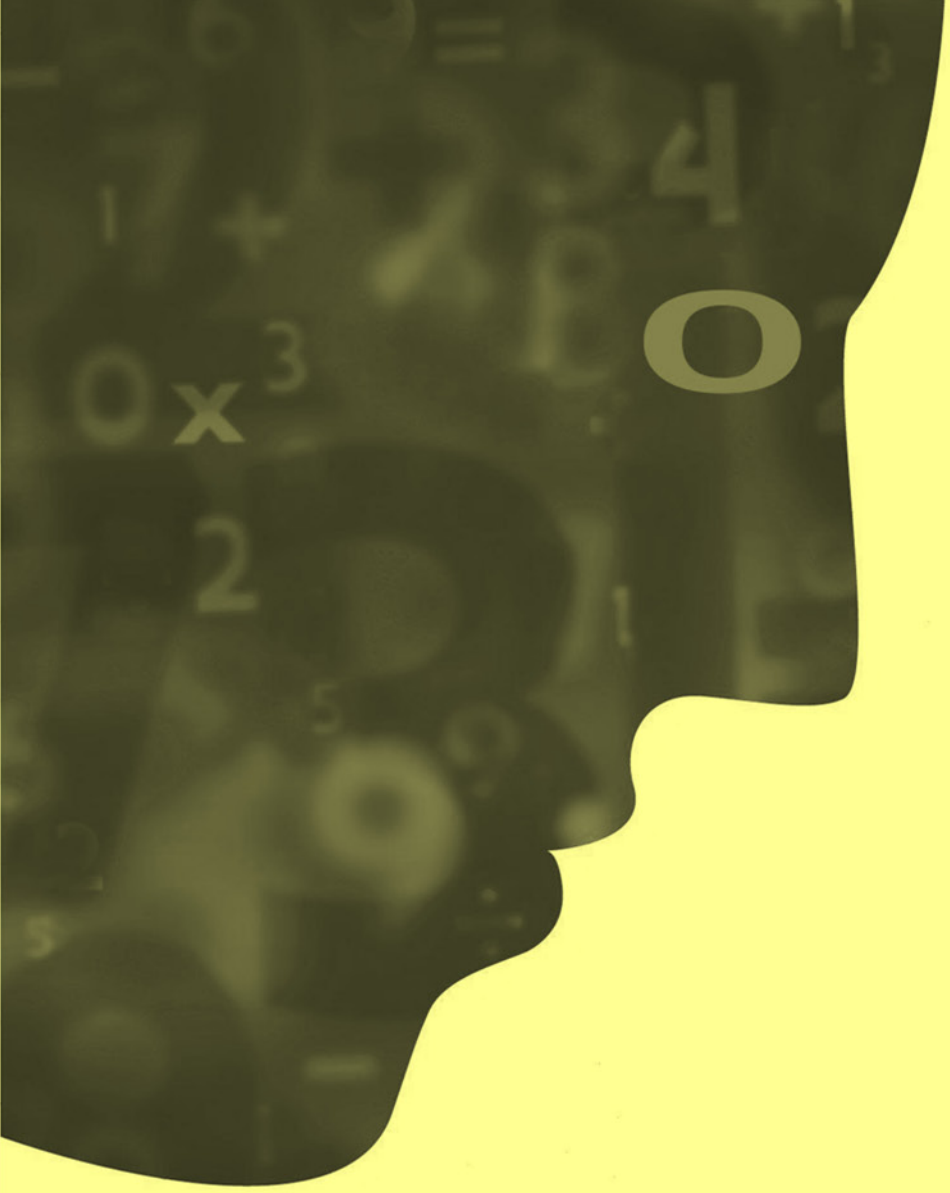
$$C_i^{obs} = e^{\eta_i} \times C\left(w_i^{obs}, y_i^{obs}\right) \times e^{\varepsilon_{C_i}} \leftrightarrow \ln C_i^{obs} = \eta_i + \ln C\left(w_i^{obs}, y_i^{obs}\right) + \varepsilon_{C_i},$$

et les écarts d'efficience en termes de coût sont :

$$C\left(w, y\right) \times \left(e^{\eta_i} - 1\right).$$

Naturellement, encore une fois, on corrige pour que les écarts de coût soient positifs ou nul dans le cas de la firme la plus efficiente.

À la lumière de ce qui a été dit au début, ce modèle comporte plusieurs défauts comme la perte de cohérence entre le coût et les demandes de facteurs et l'imposition d'une structure stochastique qui ne tient pas compte de la relation entre les termes d'erreurs et les prix et les quantités de facteur.



Les modèles de recherche opérationnelle

III. Les modèles de recherche opérationnelle

III.1. *Data Envelopment Analysis* (Farrell, 1957)

Comme toutes les méthodes économétriques, la méthode des frontières stochastiques trace une courbe dans un nuage d'observations en fonction de critères statistiques pour répartir les observations de chaque côté de la courbe. Pour sa part, la méthode de Farrell est en fait une façon de recouvrir l'ensemble de production[8] au moyen de plusieurs sous-ensembles définis par l'adoption de quelques hypothèses économiques. Parmi celles le plus souvent retenues, on retrouve :

- libre disposition des inputs;
- libre disposition des outputs;
- convexité de l'ensemble.

Si on ne retient que les deux premières, on obtient le modèle FDH (*Free disposal hull)*. L'ajout de la troisième nous donne le DEA (*Data envelopment analysis*).

Avant de poursuivre sur cette veine, revenons aux concepts de base et aux relations entre les observations et les valeurs optimales.

III.1.1. Modèle additif [9]

Par définition, nous avons $F\left(y, x\left(w, y\right)\right) \equiv 0.$ Remplaçons ces variables par les valeurs observées : $F\left(y^{obs} - \varepsilon_y, x^{obs} - \varepsilon_x - \nu_x\right) \equiv 0.$ Toujours par définition, nous aurons $F\left(y^{obs} - \varepsilon_y, x^{obs} - \varepsilon_x\right) \leq 0.$ Il y aura égalité en absence d'inefficience $\left(\nu_x = 0\right)$ et

[8] Ou tout autre représentation de la technologie comme la fonction de coût, les isoquantes, la fonction de profit, de distance, etc.

[9] Les sections III.1.1 et III.1.2 découlent naturellement de notre choix concernant les variables exogènes (c'est-à-dire, le prix des inputs et le vecteur des outputs) et endogènes (le vecteur des inputs). On peut modifier la nature (l'orientation) de la mesure, ici orientée inputs, en obtenir une orientation outputs si on utilise la maximisation des revenus ou encore une orientation mixte, inputs et outputs, dans le cas de la maximisation des profits. Dans le cas de la maximisation des revenus, ce sont les outputs qui sont endogènes alors que le vecteur des prix d'output est exogène de même que les quantités d'inputs. Dans le cas de la maximisation des profits, les quantités d'inputs et d'outputs sont endogènes alors que leurs prix sont exogènes.

l'inégalité sera stricte en présence d'inefficience $(\nu_x > 0)$. Si le terme d'inefficience est le même pour tout input (ν_x est un scalaire dans ce cas), une façon de récupérer la technologie serait de résoudre le problème suivant :

$$\max_{\nu_x} \left\{ \nu_x : F\left(y^{obs} - \varepsilon_y, x^{obs} - \varepsilon_x - \nu_x \right) \right\}.$$

Si on ne fait pas l'hypothèse d'égalité des termes d'inefficience, il faut définir un agrégat de ces termes. Par exemple, on peut choisir de minimiser la perte monétaire réelle associée à l'inefficience :

$$\max_{\nu_x} \left\{ \left(w^{obs} - \varepsilon_w \right)' \nu_x : F\left(y^{obs} - \varepsilon_y, x^{obs} - \varepsilon_x - \nu_x \right) \right\}.$$

On peut aussi prendre comme critère de minimiser la perte observée :

$$\max_{\nu_x} \left\{ w^{obs} \,' \nu_x : F\left(y^{obs} - \varepsilon_y, x^{obs} - \varepsilon_x - \nu_x \right) \right\}.$$

Les deux convergent quand il n'y a pas d'erreurs d'observation sur les prix.

III.1.2. Modèle multiplicatif

Encore une fois, par définition, nous avons $F\left(y, x\left(w, y\right)\right) \equiv 0$. Remplaçons ces variables par les valeurs observées : $F\left(y^{obs} e^{-\varepsilon_y}, x^{obs} e^{-\varepsilon_x - \nu_x} \right) \equiv 0$. Toujours par définition, nous aurons $F\left(y^{obs} e^{-\varepsilon_y}, x^{obs} e^{-\varepsilon_x} \right) \leq 0$. Il y aura égalité en absence d'inefficience $(\nu_x = 0)$ et l'inégalité sera stricte en présence d'inefficience $(\nu_x > 0)$.

Si on suppose que les termes d'inefficience sont égaux, une façon de récupérer la technologie serait de résoudre le problème suivant :

$$\max_{\nu_x} \left\{ \nu_x : F\left(y^{obs} e^{-\varepsilon_y}, x^{obs} e^{-\varepsilon_x} e^{-\nu_x} \right) \right\} \leftrightarrow \min_{\nu_x} \left\{ e^{-\nu_x} : F\left(y^{obs} e^{-\varepsilon_y}, x^{obs} e^{-\varepsilon_x} e^{-\nu_x} \right) \right\}.$$

Après avoir défini $\theta_x = e^{-\nu_x}$, on peut écrire le problème :

$$\min_{\theta_x} \left\{ \theta_x : F\left(y^{obs} e^{-\varepsilon_y}, \theta_x x^{obs} e^{-\varepsilon_x} \right) \leq 0 \right\}.$$

Si on ne fait pas l'hypothèse d'égalité des termes d'inefficience, alors il faut définir un agrégat d'inefficience. On peut à nouveau choisir la valeur monétaire réelle de l'inefficience :

$$\min_{\theta_x} \left\{ \left(w^{obs} e^{-\varepsilon_w} \right)' \theta_x : F\left(y^{obs} e^{-\varepsilon_y}, \theta_x x^{obs} e^{-\varepsilon_x} \right) \leq 0 \right\}$$

ou encore la valeur monétaire observée :

$$\min_{\theta_x} \left\{ w^{obs} \, ' \theta_x : F\left(y^{obs} e^{-\varepsilon_y}, \theta_x x^{obs} e^{-\varepsilon_x} \right) \leq 0 \right\}.$$

Si les prix sont observés sans erreur, les deux mesures convergent.

Avec cette formulation, nous sommes très près du modèle de Farrell et de la fonction de distance de Shephard (notée D et définie ci-après). En fait, si on suppose que les ν_{x_i} sont égaux, $\nu_{x_i} = \nu \leftrightarrow \theta_{x_i} = \theta$, et qu'il n'y a pas d'erreurs d'observation sur les variables, nous obtenons la fonction de distance de Shephard :

$$D\left(y^{obs}, x^{obs} \right)^{-1} \triangleq \min_{\theta} \left\{ \theta : F\left(y^{obs}, \theta x^{obs} \right) \leq 0 \right\} \leftrightarrow D\left(y^{obs}, x^{obs} \right) \triangleq \max_{\phi} \left\{ \phi : F\left(y^{obs}, \frac{x^{obs}}{\phi} \right) \leq 0 \right\}.$$

III.1.3. L'enveloppement des données

Les hypothèses de libre-disposition permettent d'engendrer des espaces réalisables pour chaque observation et l'union des espaces constitue une approximation intérieure de l'ensemble de possibilités correspondant à la représentation de la technologie retenue comme, par exemple, l'ensemble de possibilités de production ou de coût. Typiquement, grâce à cette méthode, nous obtenons des technologies en forme de marches d'escalier. L'hypothèse de convexité permet l'obtention d'un polyhèdre convexe, plus près des représentations conventionnelles de la technologie que l'on retrouve dans les livres de microéconomie.

Le choix de l'hypothèse de convexité est avant tout une affaire de préférences personnelles. Il est souvent dicté par des considérations de substitutions entre les inputs. On sera plus facilement convaincu de l'à-propos de cette hypothèse au niveau des institutions ou des systèmes de santé où le niveau d'agrégation implique la possibilité de substitutions. Cependant, au niveau le plus micro, ces substitutions pourraient être absentes à cause d'effets *putty-clay*. Nous entendons par là, que la technologie est sans doute à coefficients fixes au niveau le plus bas : on ne peut remplacer le chirurgien dans le bloc opératoire par plus de scalpels ou de fils de suture. À un niveau plus agrégé, l'hôpital par exemple, les substitutions apparaissent : on peut remplacer une opération (et ainsi le chirurgien) par des médicaments et un suivi médical. Ces substitutions peuvent intervenir rapidement ou prendre un temps considérable selon les contraintes techniques et organisationnelles.

La méthode DEA a connu beaucoup de popularité depuis les 20 dernières années. Cela s'explique par le peu d'hypothèses que requiert son utilisation. Contrairement aux méthodes économétriques, il n'est pas nécessaire d'imposer une forme fonctionnelle ou une forme particulière aux distributions des termes d'erreur. La question des tests de la théorie ne se pose pas non plus. Cependant, cette méthode a comme défaut d'être sensible aux données extrêmes. Mais c'est son incapacité à calculer des intervalles de confiance qui a constitué sa plus grande lacune pendant longtemps. En fait cela revient à supposer que les termes $\left(\varepsilon_x, \varepsilon_w, \varepsilon_y\right)$ sont nuls (ou négligeables).[10] Dans ce cas, la fonction de coût devient :

Modèle additif : $\max_{\nu_x} \left\{ w^{obs}{}'\nu_x : F\left(y^{obs}, x^{obs} - \nu_x\right)\right\}$;[11]

Modèle multiplicatif : $\min_{\theta_x} \left\{ w^{obs}{}'\theta_x : F\left(y^{obs}, \theta_x x^{obs}\right) \le 0\right\}$;

[10] Cela implique $w^{obs} = w, x^{obs} = x + \nu_x, y^{obs} = y$ et $\nu_C = w'\nu_x$.

[11] Puisque $x^{obs} - \nu_x = x$, on peut réécrire, dans le cas du modèle additif :

$$\max_{\nu_x} \left\{ w^{obs}{}'\left(x^{obs} - x\right) : F\left(y^{obs}, x^{obs} - \nu_x\right)\right\} = C^{obs} - \min_x \left\{ w^{obs}{}'x : F\left(y^{obs}, x\right)\right\}$$

et résoudre :

$$\min_x \left\{ w^{obs}{}'x : F\left(y^{obs}, x\right)\right\}$$

qui est le cas standard de minimisation des coûts par choix des inputs. Dans le cas multiplicatif, la relation n'est pas aussi simple.

et le but du DEA est de calculer à l'aide de la programmation linéaire la valeur du terme $w'\nu_x$ ou de $w'\theta_x$ pour chaque unité de décision (firmes, services, départements, systèmes de santé, etc.).

Si on fait l'hypothèse supplémentaire que les termes d'inefficience sont égaux, on obtient :

Modèle additif : $\max_{\nu_x} \left\{ \nu_x : F\left(y^{obs}, x^{obs} - \nu_x \right) \right\}$;

Modèle multiplicatif : $\min_{\theta_x} \left\{ \theta_x : F\left(y^{obs}, \theta_x x^{obs} \right) \leq 0 \right\}$.

Ce dernier modèle est le modèle standard utilisé dans la plupart des applications. Inutile de mentionner que rien ne permet de justifier l'égalité des termes d'inefficience.

Le fait que le DEA *calcule* l'inefficience et ne l'estime pas fait en sorte que la dépendance par rapport aux prix n'a plus d'importance et cela constitue un grand avantage de cette méthode.

Si on ne fait pas l'hypothèse que les termes $\left(\varepsilon_x, \varepsilon_w, \varepsilon_y \right)$ sont nuls ou négligeables, alors la mesure de l'inefficience dépend des erreurs de mesure et il devient nécessaire de calculer des intervalles de confiance. Le recours à des méthodes de bootstrap (Simar et Wilson, 1998) a permis ce calcul mais la validité de ces intervalles n'est toujours pas établie. En fait, on peut actuellement se demander si on peut avoir confiance dans les mesures de confiance.

Finalement, on peut aussi se demander si le respect assuré de la théorie est un avantage ou un inconvénient. En fait, tout dépend si on voit la théorie comme un outil de travail ou comme une construction qui exige des tests avant d'être utilisée. Autrement dit, le DEA repose sur l'hypothèse que la théorie économique est valide (à tout le moins, les hypothèses de libre disposition et de convexité)[12]. Les méthodes économétriques permettent de tester la théorie mais au prix d'autres hypothèses sur la forme fonctionnelle et sur les distributions des termes d'erreur.

La méthode DEA ne permet pas de calculer facilement les changements technologiques (voir cependant Diewert et Parkan, 1983). Pour cette raison, la jonction entre les fonctions de

[12] Il faut ajouter une hypothèse comportementale dans le cas du DEA avec minimisation des coûts.

distance utilisées en DEA et les indices de Malmquist ont permis diverses décompositions de l'efficience qui ont connu beaucoup de popularité depuis les travaux de Färe *et al*. (1992) et Caves *et al.* (1982).

III.2. Malmquist

Les indices de Malmquist sont en fait des ratios de fonctions de distance. Caves *et al.* (1982) ont montré qu'il était possible de définir des indices de productivité à partir des indices de Malmquist en faisant l'hypothèse d'efficience. Il reviendra à Färe *et al.* (1992) de généraliser leur contribution sans cette hypothèse et de montrer comment calculer cet indice à l'aide de techniques non-paramétriques issues du DEA.

À partir de cet indice de productivité, il est possible de procéder à diverses décompositions comme l'a montré Färe *et al*. Par exemple, il est maintenant classique de décomposer le changement de productivité en changement de l'efficience et en changement technologique. Depuis, de multiples décompositions ont été proposées pour tenir compte des économies d'échelle, de la présence d'inputs quasi-fixes (ou non–discrétionnaires), de la réglementation, des effets de composition de l'output ou des inputs, etc. Cela rejoint les travaux des années 70 et 80 sur la décomposition des indices de Solow.

Comme ces indices de Malmquist reposent sur les méthodes non-paramétriques utilisées pour le DEA, on retrouve les mêmes avantages (pas de formes fonctionnelles, pas de termes d'erreur) mais aussi les mêmes défauts (intervalles de confiance absents ou problématiques, sensibilité aux données extrêmes, théorie imposée).

III.3. Modèle de Aigner et Chu (1968)

Le modèle de Aigner et Chu (1968) n'a pas suscité beaucoup de travaux. Ces auteurs utilisent la programmation linéaire pour calibrer des technologies sous contrainte de respect de la théorie. Dans ce qui suit, nous nous contenterons de présenter un des modèles de Aigner et Chu. Bien que différents dans le détail, leurs divers modèles sont assez similaires. Tout repose sur la fonction de production et la définition de l'efficience.

Dans ce modèle, on peut écrire la fonction de production et la fonction de coût :[13]

$$F\left(x^{obs}-\varepsilon_x, y^{obs}-\varepsilon_y\right)\geq 0$$

et

$$C^{obs}-C\left(w^{obs}-\varepsilon_w, y^{obs}-\varepsilon_y\right)-\underbrace{\left[w^{obs}{}'\varepsilon_x+x^{obs}{}'\varepsilon_w-\varepsilon_w{}'\varepsilon_x\right]}_{\varepsilon_C}-\underbrace{\left(w^{obs}-\varepsilon_w\right)'\nu_x}_{\nu_C}\geq 0$$

.

On commence par choisir une forme fonctionnelle pour la technologie :

$$F\left(x^{obs}-\varepsilon_x, y^{obs}-\varepsilon_y;\alpha\right)\geq 0$$

et

$$C^{obs}-C\left(w^{obs}-\varepsilon_w, y^{obs}-\varepsilon_y;\beta\right)-\underbrace{\left[w^{obs}{}'\varepsilon_x+x^{obs}{}'\varepsilon_w-\varepsilon_w{}'\varepsilon_x\right]}_{\varepsilon_C}\geq 0.$$

Si on suppose que la technologie est linéaire dans les paramètres (possiblement après transformation logarithmique comme dans le cas Cobb-Douglas retenu par Aigner et Chu) :

$$\left(x^{obs}-\varepsilon_x\right)'\alpha_x-\left(y^{obs}-\varepsilon_y\right)'\alpha_y\geq 0$$

et

$$C^{obs}-\left(w^{obs}-\varepsilon_w\right)'\beta_w-\left(y^{obs}-\varepsilon_y\right)'\beta_y-\underbrace{\left[w^{obs}{}'\varepsilon_x+x^{obs}{}'\varepsilon_w-\varepsilon_w{}'\varepsilon_x\right]}_{\varepsilon_C}\geq 0.$$

L'inefficience ν est introduite comme suit :

$$\left(x^{obs}-\varepsilon_x-\nu_x\right)'\alpha_x-\left(y^{obs}-\varepsilon_y\right)'\alpha_y=0$$

et

$$C^{obs}-\left(w^{obs}-\varepsilon_w\right)'\beta_w-\left(y^{obs}-\varepsilon_y\right)'\beta_y-\underbrace{\left[w^{obs}{}'\varepsilon_x+x^{obs}{}'\varepsilon_w-\varepsilon_w{}'\varepsilon_x\right]}_{\varepsilon_C}-\underbrace{\left(w^{obs}-\varepsilon_w\right)'\nu_x}_{\nu_C}=0.$$

On isole les termes d'erreur et d'inefficience :

[13] Nous ne distinguerons pas entre le modèle additif et le modèle multiplicatif pour des raisons qui deviendront évidentes sous peu. Comme il faudra linéariser la technologie (fonction de production ou de coût), les variables du modèle multiplicatif seront sous forme logarithmique et le modèle résultant sera en tout point équivalent en prenant en compte que dans le cas additif, les variables seront en niveau et dans le cas multiplicatif elles seront en log.

$$x^{obs}{}'\alpha_x - y^{obs}{}'\alpha_y = \underbrace{\left[\varepsilon_x{}'\alpha_x + \varepsilon_y{}'\alpha_y\right]}_{\varepsilon_\alpha} + \underbrace{\nu_x{}'\alpha_x}_{\nu_F}$$

et

$$C^{obs} - w^{obs}{}'\beta_w - y^{obs}{}'\beta_y = \underbrace{\left[\varepsilon_w{}'\beta_w + \varepsilon_y{}'\beta_y\right]}_{\varepsilon_\beta} + \underbrace{\left[w^{obs}{}'\varepsilon_x + x^{obs}{}'\varepsilon_w - \varepsilon_w{}'\varepsilon_x\right]}_{\varepsilon_C} + \underbrace{\left(w^{obs} - \varepsilon_w\right)'\nu_x}_{\nu_C}.$$

Pour l'entreprise i ($i = 1$ à n), on peut écrire :

$$x_i^{obs}{}'\alpha_x - y_i^{obs}{}'\alpha_y = \underbrace{\left[\varepsilon_{x_i}{}'\alpha_x + \varepsilon_{y_i}{}'\alpha_y\right]}_{\varepsilon_{\alpha_i}} + \underbrace{\nu_{x_i}{}'\alpha_x}_{\nu_{F_i}}, \forall i = 1,\ldots,n$$

et

$$C_i^{obs} - w_i^{obs}{}'\beta_w - y_i^{obs}{}'\beta_y = \underbrace{\left[\varepsilon_{w_i}{}'\beta_w + \varepsilon_{y_i}{}'\beta_y\right]}_{\varepsilon_{\beta_i}} + \underbrace{\left[w_i^{obs}{}'\varepsilon_{x_i} + x_i^{obs}{}'\varepsilon_{w_i} - \varepsilon_{w_i}{}'\varepsilon_{x_i}\right]}_{\varepsilon_{C_i}} + \underbrace{\left(w_i^{obs} - \varepsilon_{w_i}\right)'\nu_{x_i}}_{\nu_{C_i}}, \forall i = 1,\ldots,n.$$

De façon à imposer la positivité des termes d'erreur, Aigner et Chu supposent que les erreurs de mesure sont nulles ou négligeables et ainsi :

$$x_i^{obs}{}'\alpha_x - y_i^{obs}{}'\alpha_y = \nu_{x_i}{}'\alpha_x = \nu_{F_i} \geq 0, \forall i = 1,\ldots,n$$

et

$$C_i^{obs} - w_i^{obs}{}'\beta_w - y_i^{obs}{}'\beta_y = w_i^{obs}{}'\nu_{x_i} = \nu_{C_i} \geq 0, \forall i = 1,\ldots,n.$$

Ils traitent les termes ν_{F_i} et ν_{C_i} comme n'importe quel terme d'erreur à ceci près qu'ils sont nécessairement non négatifs, $\nu_{F_i} \geq 0$ et $\nu_{C_i} \geq 0$.

La résolution du problème requiert un critère de décision. On peut prendre la minimisation des termes d'erreur au carré comme avec les MCO :

$$\min_{\alpha \geq 0} \sum_{i=1}^{n} \nu_{F_i}{}^2 = \sum_{i=1}^{n} \left(x_i^{obs}{}'\alpha_x - y_i^{obs}{}'\alpha_y\right)^2$$

et

$$\min_{\beta_w \geq 0, \beta_y \geq 0} \sum_{i=1}^{n} \nu_{C_i}{}^2 = \sum_{i=1}^{n} \left(C_i^{obs} - w_i^{obs}{}'\beta_w - y_i^{obs}{}'\beta_y\right)^2.$$

sous contrainte que :

$$\nu_{F_i} = x_i^{obs}{}'\alpha_x - y_i^{obs}{}'\alpha_y \geq 0, \forall i = 1,...,n$$

et

$$\nu_{C_i} = C_i^{obs} - w_i^{obs}{}'\beta_w - y_i^{obs}{}'\beta_y \geq 0, \forall i = 1,...,n.$$

Cela exige le recours à des méthodes programmation quadratique. Cependant, cette méthode est sensible aux valeurs extrêmes. Pour cette raison, Aigner et Chu propose de minimiser la somme des erreurs :[14]

$$\min_{\alpha \geq 0} \sum_{i=1}^{n} \nu_{F_i} = \sum_{i=1}^{n} \left(x_i^{obs}{}'\alpha_x - y_i^{obs}{}'\alpha_y \right)$$

et

$$\min_{\beta_w \geq 0, \beta_y \geq 0} \sum_{i=1}^{n} \nu_{C_i} = \sum_{i=1}^{n} \left(C_i^{obs} - w_i^{obs}{}'\beta_w - y_i^{obs}{}'\beta_y \right)$$

sous contrainte que :

$$\nu_{F_i} = x_i^{obs}{}'\alpha_x - y_i^{obs}{}'\alpha_y \geq 0, \forall i = 1,...,n$$

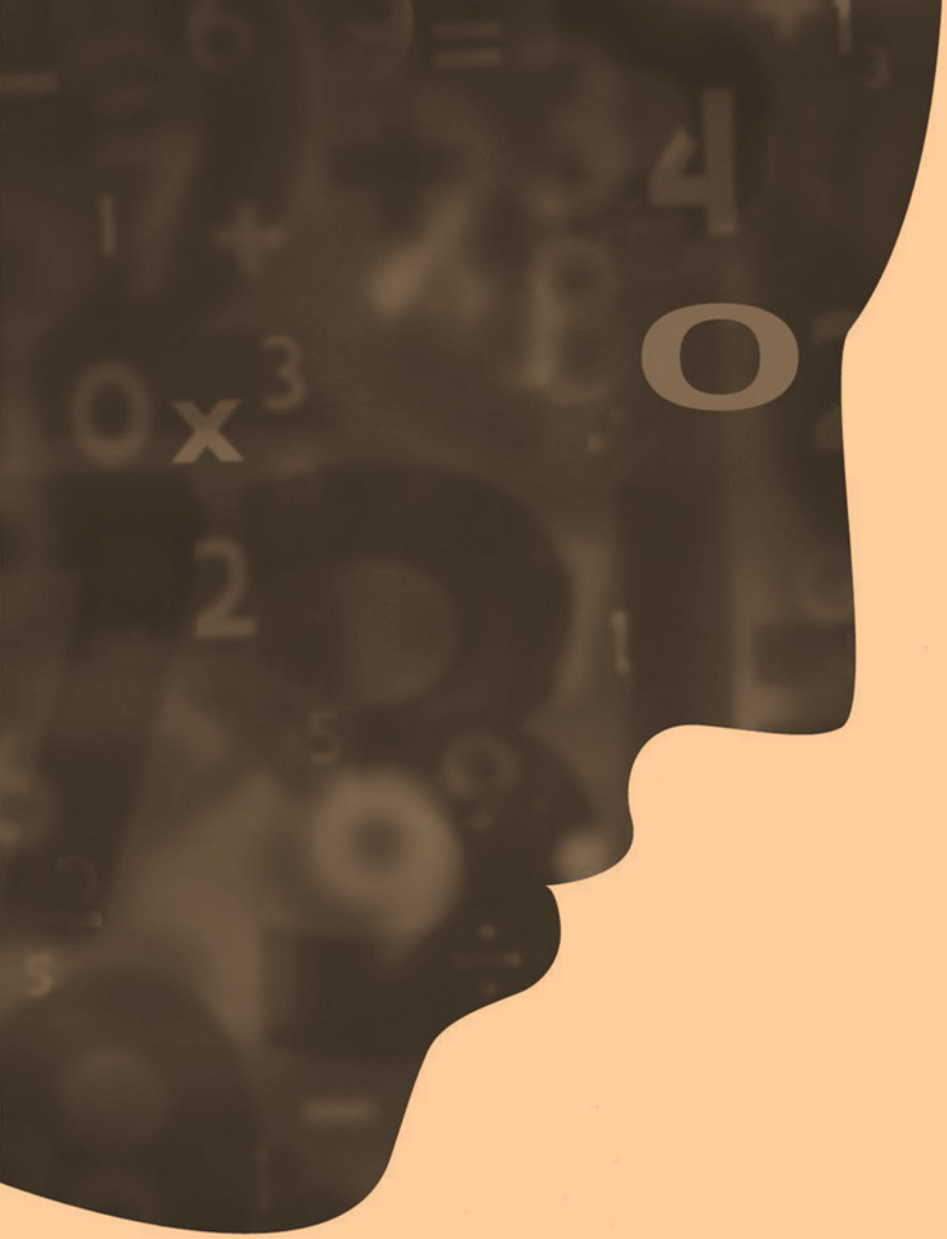
et

$$\nu_{C_i} = C_i^{obs} - w_i^{obs}{}'\beta_w - y_i^{obs}{}'\beta_y \geq 0, \forall i = 1,...,n.$$

Des méthodes simples de programmation linéaire suffisent pour solutionner ce problème. On peut ajouter des contraintes à ce programme linéaire. Par exemple, si les variables sont sous forme logarithmique et si on impose que les rendements d'échelle sont constants, alors $\alpha_x{}'\overline{1}_x = \alpha_y{}'\overline{1}_y$ et $\beta_y{}'\overline{1}_y = 1$, où $\overline{1}_x$ (resp. $\overline{1}_y$) est un vecteur de 1 de même dimension que le vecteur x (resp. y). On peut imposer l'homogénéité de degré un dans les prix soit en prenant les prix et le coût en prix relatifs ou encore sous forme de restriction sur les coefficients des prix ($1 = \beta_w{}'\overline{1}_w$).

[14] On voit ici l'importance de l'hypothèse qu'il n'y ait pas d'erreur de mesure. Cette hypothèse implique $\nu_{F_i} \geq 0$ et $\nu_{C_i} \geq 0$ et ainsi la somme des erreurs est une mesure d'inefficience adéquate, ce qui ne serait pas le cas si ce terme pouvait prendre des valeurs négatives qui compenseraient les termes positifs.

Cette méthode est intermédiaire entre le DEA et les méthodes économétriques. Les méthodes utilisées sont celles du DEA (programmation linéaire) et n'exige pas le recours à un choix pour la distribution des termes d'erreur, mais la façon de définir les paramètres est celle de l'économétrie (minimisation de la somme des erreurs, au carré ou non, et choix de la forme fonctionnelle). Bien entendu, un grand désavantage de cette méthode est le recours à des fonctions de production linéaire dans les paramètres sans compter l'absence d'intervalles de confiance.



Méthode *Sur le dos d'une enveloppe*

IV. Méthode *Sur le dos d'une enveloppe*

IV.1. Méthode comptable

La détermination des coûts unitaires de référence aux fins de budgétisation trouve sa plus simple expression dans la méthode comptable. Revenons à la relation entre le coût observé et le coût minimum. Pour un hôpital h au temps t, nous avons :

$$C_{ht}^{obs} = C\left(\boldsymbol{w}_{ht}^{obs} - \varepsilon_{w_{ht}}, \boldsymbol{y}_{ht}^{obs} - \varepsilon_{q_{ht}}, t\right) + \underbrace{\left[w_{ht}^{obs}{}'\varepsilon_{x_{ht}} + x_{ht}^{obs}{}'\varepsilon_{w_{ht}} - \varepsilon_{w_{ht}}{}'\varepsilon_{x_{ht}}\right]}_{\varepsilon_{C_{ht}}} + \underbrace{\left(w_{ht}^{obs} - \varepsilon_{w_{ht}}\right)'\nu_{x_{ht}}}_{\nu_{C_{ht}}}.$$

Deux hypothèses sont nécessaires : l'hôpital (ou le département) ne produit qu'un seul output, i.e., y est un scalaire, et l'établissement produit à rendements constants. Dans ce cas, le coût observé devient :

$$C_{ht}^{obs} = c\left(w_{ht}^{obs} - \varepsilon_{w_{ht}}, t\right) \times \left(y_{ht}^{obs} - \varepsilon_{y_{ht}}\right) + \varepsilon_{C_{ht}} + \nu_{C_{ht}}.$$

On peut aussi travailler directement à partir du coût unitaire. Il suffit de diviser par l'output. Le coût unitaire, noté c^{obs}, s'écrit :

$$c_{ht}^{obs} \equiv \frac{C_{ht}^{obs}}{y_{ht}^{obs}} = c\left(w_{ht}^{obs} - \varepsilon_{w_{ht}}, t\right) \times \frac{\left(y_{ht}^{obs} - \varepsilon_{y_{ht}}\right)}{y_{ht}^{obs}} + \frac{\varepsilon_{C_{ht}} + \nu_{C_{ht}}}{y_{ht}^{obs}}.$$

Si on est prêt à faire l'hypothèse qu'il n'y a pas d'erreur sur les variables, le coût unitaire devient :

$$c_{ht}^{obs} = c\left(w_{ht}^{obs}\right) + \frac{\nu_{C_{ht}}}{y_{ht}^{obs}}.$$

Et si, en plus, on suppose que les prix sont les mêmes dans tous les établissements, il devient possible de comparer les coûts unitaires des différentes institutions de santé entre elles et à travers le temps. Par exemple, on peut fixer comme règle que le coût unitaire servant aux fins de budgétisation en t est la moyenne des coûts unitaires de la période en $t-1$:

$$c^{référence} = \text{moyenne}\left\{c\left(w^{obs}, t-1\right) + \frac{v_{C_{ht-1}}}{y^{obs}_{ht-1}}\right\}$$
$$= \text{moyenne}\left\{c\left(w^{obs}, t-1\right)\right\} + \text{moyenne}\left\{\frac{v_{C_{ht-1}}}{y^{obs}_{ht-1}}\right\}.$$

Bien entendu, d'autres règles peuvent être retenues. Par exemple, il est possible de prendre la médiane plutôt que la moyenne sous prétexte que la médiane est moins sensible que la moyenne aux valeurs extrêmes :

$$c^{référence} = \text{médiane}\left\{c\left(w^{obs}, t-1\right) + \frac{v_{C_{ht-1}}}{y^{obs}_{ht-1}}\right\}.$$

Et, finalement, il est possible de prendre le minimum observé. Dans ce cas, on retrouve une formulation très proche de celle de moindres carrés corrigés. Tout écart de coût unitaire devient une mesure d'inefficience. En résumé, la méthode comptable est simple, mais repose sur des hypothèses très fortes et certainement inacceptables d'un point de vue théorique. Parmi les hypothèses les moins crédibles, on retrouve les rendements d'échelle constants et un output unique, sans parler de l'absence d'erreurs observationnelles.

IV.2. Modification de la méthode comptable

Reprenons la formule du coût observé, égal au coût efficient plus le coût de l'inefficience (on conserve l'hypothèse d'un seul output et on omet les erreurs d'observation) :

$$C^{obs}_{ht} = C\left(\boldsymbol{w}^{obs}_{ht}, y^{obs}_{ht}, t\right) + v^{obs}_{ht}.$$

Invariablement, il y aura des écarts entre les coûts unitaires des différents établissements h. La question est de savoir si ces écarts sont le fait d'inefficiences ou le résultat de facteurs qui défavorisent un établissement particulier. Cette question est au centre de la problématique, car elle remet en question le désir d'en arriver à définir un coût de référence unique pour tous les établissements. Si certains facteurs autres que l'inefficience et hors du contrôle de

l'établissement impliquent une hausse de ses coûts, il est impératif d'en tenir compte de façon à ne pas pénaliser cet établissement. Cet aspect du problème est explicitement pris en compte par les frontières stochastiques et le DEA par l'ajout d'autres variables comme les prix et la qualité (voir section V) mais pas par la méthode comptable présentée dans la section précédente.

En fait, il est possible d'exprimer les coûts d'un établissement en prenant le contexte d'un autre établissement comme référence. Prenons l'indice r pour l'institution de référence. Les variables explicatives du coût de l'établissement de référence sont $\left(w_{rt}^{obs}, y_{rt}^{obs}, v_{C_{rt}}\right)$. Le coût de l'institution de référence est $C_{rt}^{obs} = C\left(w_{rt}^{obs}, y_{rt}^{obs}, t\right) + v_{C_{rt}}$. En prenant une expansion de Taylor au premier ordre de la fonction de coût efficient de l'institution de référence $C\left(w_{rt}^{obs}, y_{rt}^{obs}, t\right)$ autour du point de l'institution h, i.e., $\left(w_{ht}^{obs}, y_{ht}^{obs}\right)$, nous obtenons (pour une année t donnée) :[15]

$$C\left(\boldsymbol{w}_{rt}^{obs}, y_{rt}^{obs}, t\right) - C\left(\boldsymbol{w}_{ht}^{obs}, y_{ht}^{obs}, t\right) = \sum_{i=1}^{I} x_{ht}^{i} \times \left(w_{rt}^{i,obs} - w_{ht}^{i,obs}\right) + \frac{\partial C}{\partial y} \times \left(y_{rt}^{obs} - y_{ht}^{obs}\right).$$

Chacun des termes de droite constitue un facteur expliquant partiellement l'écart de coût entre l'institution h et l'institution de référence r.

Le premier terme indique que les coûts seront différents, toutes choses étant égales par ailleurs, si l'institution h fait face à des prix différents des prix de référence. Il lui en coûtera plus cher si ses prix sont plus élevés. L'autre terme fait référence à l'échelle de production (le terme faisant appel à la dérivée partielle par rapport à y). L'écart des coûts devient :

$$C\left(\boldsymbol{w}_{rt}^{obs}, y_{rt}^{obs}, t\right) - C\left(\boldsymbol{w}_{ht}^{obs}, y_{ht}^{obs}, t\right) = \sum_{i=1}^{I} x_{ht}^{i} \times \left(w_{rt}^{i,obs} - w_{ht}^{i,obs}\right) + \frac{\partial C}{\partial y} \times \left(y_{rt}^{obs} - y_{ht}^{obs}\right).$$

On peut aussi écrire :

$$C\left(\boldsymbol{w}_{rt}^{obs}, y_{rt}^{obs}, t\right) = C\left(\boldsymbol{w}_{ht}^{obs}, y_{ht}^{obs}, t\right) + \sum_{i=1}^{I} x_{ht}^{i} \times \left(w_{rt}^{i,obs} - w_{ht}^{i,obs}\right) + \frac{\partial C}{\partial y} \times \left(y_{rt}^{obs} - y_{ht}^{obs}\right).$$

[15] Nous utilisons le lemme de Shephard : $\partial C / \partial w^{i} = x^{i}$.

Cette expression fait ressortir que le coût minimum de l'institution h corrigé par un certain nombre de termes est égal au coût minimum de l'institution de référence r. La présence de la dérivée partielle par rapport à l'output exige une hypothèse supplémentaire. En présence de rendements d'échelle constants, on peut remplacer le coût marginal $\partial C/\partial y$ par le coût unitaire efficient $C\left(w_{ht}^{obs}, y_{ht}^{obs}, t\right)\Big/ y_{ht}^{obs}$. Après substitution, il s'ensuit :

$$C\left(\boldsymbol{w}_{ht}^{obs}, y_{ht}^{obs}, t\right) = C\left(\boldsymbol{w}_{rt}^{obs}, y_{rt}^{obs}, t\right) + \sum_{i=1}^{I} x_{ht}^{i,obs} \times \left(w_{ht}^{i,obs} - w_{rt}^{i,obs}\right) + \frac{C\left(\boldsymbol{w}_{ht}^{obs}, y_{ht}^{obs}, t\right)}{y_{ht}^{obs}} \times \left(y_{ht}^{obs} - y_{rt}^{obs}\right),$$

$$0 = C\left(\boldsymbol{w}_{rt}^{obs}, y_{rt}^{obs}, t\right) + \sum_{i=1}^{I} x_{ht}^{i,obs} \times \left(w_{ht}^{i,obs} - w_{rt}^{i,obs}\right) - \frac{C\left(\boldsymbol{w}_{ht}^{obs}, y_{ht}^{obs}, t\right)}{y_{ht}^{obs}} \times y_{rt}^{obs}.$$

En divisant par y_{rt} des deux côtés, nous obtenons :

$$\frac{C\left(\boldsymbol{w}_{ht}^{obs}, y_{ht}^{obs}, t\right)}{y_{ht}} = \frac{C\left(\boldsymbol{w}_{rt}^{obs}, y_{rt}^{obs}, t\right)}{y_{rt}} + \sum_{i=1}^{I} \frac{x_{ht}^{i,obs}}{y_{rt}} \times \left(w_{ht}^{i,obs} - w_{rt}^{i,obs}\right).$$

Il nous reste à substituer les coûts observés :

$$\frac{C_{ht}^{obs} - v_{ht}}{y_{ht}^{obs}} = \frac{C_{rt}^{obs} - v_{rt}}{y_{rt}^{obs}} + \sum_{i=1}^{I} \frac{x_{ht}^{i,obs}}{y_{rt}^{obs}} \times \left(w_{ht}^{i,obs} - w_{rt}^{i,obs}\right),$$

d'où

$$\frac{C_{ht}^{obs}}{y_{ht}^{obs}} = \frac{C_{rt}^{obs}}{y_{rt}^{obs}} + \sum_{i=1}^{I} \frac{x_{ht}^{i,obs}}{y_{rt}^{obs}} \times \left(w_{ht}^{i,obs} - w_{rt}^{i,obs} \right) + \frac{v_{ht}}{y_{ht}^{obs}} - \frac{v_{rt}}{y_{rt}^{obs}}.$$

Notons qu'en choisissant un établissement de référence (ici r), cela revient à supposer que cet établissement est efficient.[16] Autrement dit $\frac{v_{rt}}{y_{rt}^{obs}} = 0$ et ainsi le coût efficient de l'établissement de référence est égal à son coût observé, C_{rt}^{obs}. En substituant cette définition dans l'équation précédente, il s'ensuit :

$$\frac{C_{ht}^{obs}}{y_{ht}^{obs}} = \left\{ \frac{C_{rt}^{obs}}{y_{rt}^{obs}} + \sum_{i=1}^{I} \left[\frac{x_{ht}^{i,obs}}{y_{rt}^{obs}} \times \left(w_{ht}^{i,obs} - w_{rt}^{i,obs} \right) \right] \right\} + \frac{v_{ht}}{y_{ht}^{obs}}.$$

Le terme entre accolades représente le coût unitaire attendu pour l'établissement h. Le terme $\frac{v_{ht}}{y_{ht}^{obs}}$ représente l'écart dans les coûts unitaires dû à l'inefficience de h, écart qui ne doit pas être retenu dans le budget de cet établissement.

Naturellement, cette correction suppose que les prix et quantité de chacun des inputs variables sont observables. Si ce n'est pas le cas, il faudra se restreindre aux seuls inputs pour lesquels ces variables sont observées.

Note : La formule de correction précédente est basée sur une approximation de la fonction de coût efficient de l'établissement de référence r autour du point de l'établissement h. Il est bien entendu possible de procéder en sens contraire et d'approximer la fonction de coût efficient de l'établissement de référence h autour du point de l'établissement r. On obtient un facteur de correction légèrement différent :

$$\frac{C_{ht}^{obs}}{y_{ht}^{obs}} = \left\{ \frac{C_{rt}^{obs}}{y_{rt}^{obs}} + \sum_{i=1}^{I} \left[\frac{x_{rt}^{i,obs}}{y_{ht}^{obs}} \times \left(w_{ht}^{i,obs} - w_{rt}^{i,obs} \right) \right] \right\} + \frac{v_{ht}}{y_{ht}^{obs}}.$$

[16] Il serait plus juste de parler d'efficience relative.

IV.3. Recours à la méthode économétrique pour compléter la méthode comptable

À ce stade, nous avons une mesure corrigée du coût attendu qui incorpore les différences dans le prix des facteurs variables (du moins ceux à notre disposition). La section précédente montre que les rendements non constants peuvent expliquer l'écart des coûts unitaires. Nous verrons dans la section suivante que d'autres facteurs peuvent aussi se rajouter. Cela implique qu'une partie du surcoût $\frac{v_{ht}}{y_{ht}^{obs}}$ peut en fait s'expliquer par l'un ou plusieurs de ces facteurs. Pour l'instant, nous nous contenterons des seuls rendements non constants. Il est facile de modifier ce qui suit en incorporant d'autres facteurs.

À partir de la relation entre le coût efficient de h et de r :

$$C\left(\boldsymbol{w}_{rt}^{obs}, y_{rt}^{obs}, t\right) - C\left(\boldsymbol{w}_{ht}^{obs}, y_{ht}^{obs}, t\right) = \sum_{i=1}^{I} x_{ht}^{i,obs} \times \left(w_{rt}^{i,obs} - w_{ht}^{i,obs}\right) + \frac{\partial C}{\partial y} \times \left(y_{rt}^{obs} - y_{ht}^{obs}\right),$$

on peut montrer que :

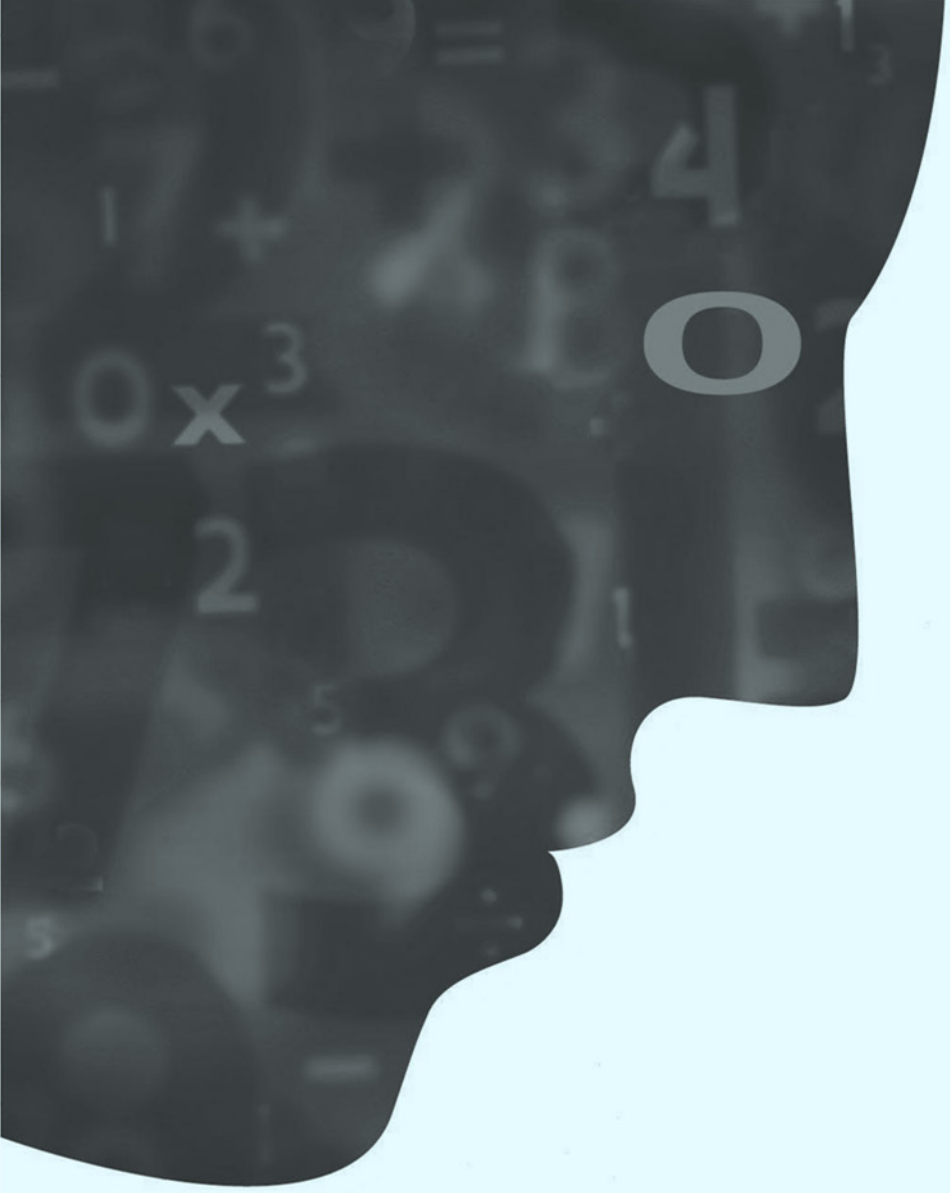
$$\frac{C_{ht}^{obs}}{y_{ht}^{obs}} = \left\{\frac{C_{rt}^{obs}}{y_{rt}^{obs}} + \sum_{i=1}^{I} x_{ht}^{i,obs} \times \left(w_{ht}^{i,obs} - w_{ht}^{i,obs}\right)\right\} + \frac{1}{y_{rt}^{obs}}\left\{\left(\frac{\partial C}{\partial y} - \frac{C\left(\boldsymbol{w}_{ht}^{obs}, y_{ht}^{obs}, t\right)}{y_{ht}^{obs}}\right) \times \left(y_{ht}^{obs} - y_{rt}^{obs}\right)\right\} + \frac{v_{ht}}{y_{ht}^{obs}}.$$

Le premier terme de droite est le coût attendu selon la méthode comptable corrigée. Le troisième terme est la mesure de l'inefficience. Le second terme a été supposé nul ou négligeable dans la méthode comptable corrigée. Il représente l'impact des rendements non constants (l'écart entre le coût marginal et le coût moyen). Nous obtenons la relation :

$$\frac{C_{ht}^{obs}}{y_{ht}^{obs}} - \left\{\frac{C_{rt}^{obs}}{y_{rt}^{obs}} + \sum_{i=1}^{I} x_{ht}^{i,obs} \times \left(w_{ht}^{i,obs} - w_{ht}^{i,obs}\right)\right\} = \frac{1}{y_{rt}^{obs}}\left\{\left(\frac{\partial C}{\partial y} - \frac{C\left(\boldsymbol{w}_{ht}^{obs}, y_{ht}^{obs}, t\right)}{y_{ht}^{obs}}\right) \times \left(y_{ht}^{obs} - y_{ht}^{obs}\right)\right\} + \frac{v_{ht}}{y_{ht}^{obs}}.$$

Le terme de droite comprend tous les termes mesurables. Cette équation implique que l'écart entre le coût unitaire de l'établissement h et le coût unitaire de la méthode comptable corrigée, i.e., $\frac{C_{ht}^{obs}}{y_{ht}^{obs}} - \left\{\frac{C_{rt}^{obs}}{y_{rt}^{obs}} + \sum_{i=1}^{I} x_{ht}^{i,obs} \times \left(w_{ht}^{i,obs} - w_{ht}^{i,obs}\right)\right\}$, doit être lui-même modifié pour incorporer l'impact de ces variables. Malheureusement, la présence des dérivées partielles ne

permet pas de correction à partir des seules observations. Il est possible de régresser ce surcoût sur ces variables explicatives $\left(y_{ht}^{obs}\right)$ à condition d'avoir une banque de données fiables. On peut généraliser ce qui précède en incorporant des erreurs de mesure sur les variables exogènes.



Le passage de la théorie à la pratique

V. Le passage de la théorie à la pratique

On peut penser qu'en présence de bonnes données, les méthodes précédentes donnent toutes à peu près les mêmes résultats. Les études qui ont comparé les diverses méthodes n'arrivent pas à cette conclusion. Pourquoi ?

V.1. Surplus de variables : agrégation et qualité

Le grand nombre d'inputs utilisés par un hôpital et la grande diversité d'outputs qu'il produit sont une source de problèmes importants. Par exemple, dans le cas d'une fonction de coût translog avec m inputs et n outputs et un terme de tendance pour le changement technologique, il y aura $(1+(m+n)+(m+n)\times(m+n+1)/2$ paramètres.[17] Dans des travaux antérieurs, nous avons constitué une banque de données pour les hôpitaux québécois. Il y avait avant agrégation plus d'une centaine d'inputs[18] et des milliers d'outputs (autant que d'actes médicaux, de types d'examens de laboratoire et de tests, de services d'hôtellerie et de buanderie, etc. À titre d'exemple prenons 100 inputs et 1000 outputs. On obtient plus de 600 000 paramètres. Le Québec ne compte qu'une centaine d'hôpitaux. Pour estimer autant de paramètres, il faudrait 6 000 ans d'observations avec des données annuelles. Il faut donc réduire le nombre de variables et le recours à l'agrégation est un passage obligé. L'omission de variables est évidemment à rejeter. Mais comment agréger ?

Cette question est délicate et le plus souvent elle est résolue en maximisant le contenu informationnel des banques de données. Autrement dit, c'est le contenu de données disponibles qui dicte la nature des données utilisées Dans ce contexte, la question de l'omission de variables revient sur le tapis avec tout ce que cela impose comme limite à la confiance dans les résultats qui deviennent immédiatement susceptibles de biais.

Si on laisse de côté l'omission de variables, comment agréger les données à notre disposition ? Le processus d'agrégation consiste essentiellement à trouver des fonctions agrégatives W, X et Y telle que :

[17] On a tenu compte de la propriété d'homogénéité.

[18] Il suffit de mentionner les seuls médecins : une trentaine de spécialités et cinq statuts. Si on ajoute les catégories d'infirmières, de préposés aux bénéficiaires, le personnel administratif, etc. on arrive facilement à des centaines d'inputs.

$$F(x_1,...,x_m;y_1,...,y_n) = F\left(X_1(x_1,...,x_a),...,X_\delta(x_{c+1},...,x_m);Y_1(y_1,...,y_d),...,Y_\gamma(y_{f+1},...,y_n)\right)$$
$$= F(X_1,X_2,...,X_\delta;Y_1,Y_2,...,Y_\gamma)$$

et

$$C(x_1,...,x_m;y_1,...,y_n) = C\left(W_1(w_1,...,w_a),...,W_\delta(w_{c+1},...,w_m);Y_1(y_1,...,y_d),...,Y_\gamma(y_{f+1},...,y_n)\right)$$
$$= C(W_1,W_2,...,W_\delta;Y_1,Y_2,...,Y_\gamma).$$

Naturellement, tout cela n'a de sens que si $\delta < m$ et $\gamma < n$.

La question de l'agrégation, exacte ou non, est certes fondamentale mais elle ne nous retiendra que pour mieux aborder la question de la qualité. On peut se référer à Blackorby *et al.* (1978) pour une revue extensive sur l'agrégation. Certaines propriétés ne sont pas nécessairement évidentes et il est parfois difficile de les invoquer pour justifier les agrégats retenus. Dire que pour deux outputs d'un agrégat, les taux marginaux de substitution sont indépendants des outputs hors agrégat n'est pas économiquement limpide et vérifiable *a priori*. Ou alors, dire qu'il faut que le taux de croissance de cet agrégat soit une somme pondérée des taux de croissance des divers outputs sur la base des coûts marginaux ne mène nulle part quand justement on ne connaît pas les coûts marginaux et qu'aucun marché n'existe pour les remplacer par des prix de marché en concurrence. Ces difficultés expliquent le côté *ad hoc* des agrégats effectivement retenus par les chercheurs en santé. Cependant, même en ne tenant pas compte de ces difficultés, il n'en demeure pas moins que l'agrégation doit être minimalement cohérente avec les caractéristiques des institutions étudiées. En fait, la question est la suivante : dans le processus d'agrégation, que perd-on ? Et est-il possible de compenser cette perte ?

Si l'agrégation n'est pas exacte, alors cela revient à introduire un nouveau terme d'erreur. Si l'agrégation est exacte, nous aurons (nous prenons le cas de l'agrégation des outputs, mais le cas des autres variables est identique) :

$$Y_i = Y_i(y_1,\ldots,y_d).$$

Si l'agrégation est inexacte, nous aurons :

$$\tilde{Y}_i = \tilde{Y}_i(y_1,\ldots,y_d) + \tilde{\varepsilon}_y.$$

Le terme $\tilde{\varepsilon}_y$ constitue un nouveau terme d'erreur qui se rajoute aux autres termes d'erreur de mesure et d'optimisation. Rien ne permet de conclure que ce terme soit indépendant des autres variables ce qui complique encore plus le traitement des frontières stochastiques.

Dans le cas de l'agrégation inexacte, on peut utiliser des fonctions à valeurs vectorielles et remplacer le vecteur des outputs non observés par un ensemble de variables de dimension réduite. Pour illustrer cette question, prenons le cas des patients ayant subi une appendicectomie. Supposons que, pour des raisons médicales, on considère que les appendicites ne sont pas traitées de la même façon selon que le patient soit jeune, adulte ou agé, et que seuls deux résultats sont possibles : réussite et échec suivi de la mort. Cela nous laisse six outputs :

- nombre d'appendicectomies d'un jeune avec succès;
- nombre d'échecs de l'appendicectomie d'un jeune ;
- nombre d'appendicectomies d'un adulte avec succès;
- nombre d'échecs de l'appendicectomie d'un adulte ;
- nombre d'appendicectomies d'une personne âgée avec succès;
- nombre d'échecs de l'appendicectomie d'une personne âgée.

Ces six outputs peuvent être trop nombreux en présence de peu de données. On peut ramener le nombre d'outputs à trois :

- nombre d'appendicectomies (somme des six catégories d'appendicectomies);
- pourcentage d'échecs;
- âge moyen des patients.

Ceci n'est qu'un exemple d'agrégation possible. Selon le nombre de degrés de liberté, on peut être amené à agréger encore plus en laissant tomber l'âge moyen par exemple. Bien entendu, l'agrégation n'est retenue qu'en cas de manque de données, mais cela sera toujours le cas. Ce type d'agrégation est souvent utilisé, en santé et ailleurs comme en transport, et on l'appelle hédonique.

Ici on peut s'attarder sur la notation. Dans l'exemple précédent, on passe de six mesures d'outputs à trois. Dans la terminologie retenue par les chercheurs en économie de la santé, ces variables ne sont pas traitées de façon symétrique. En fait, certaines sont appelées des outputs alors que d'autres sont des variables de qualité. Dans notre exemple, le nombre

d'appendicectomies serait un output alors que le pourcentage d'échec et l'âge moyen seraient des variables de qualité. Cette question est purement rhétorique et n'apporte rien au débat. En fait, les trois variables sont des variables *qualitatives* qui n'ont de sens que prises simultanément afin de quantifier un agrégat de six outputs.

Si nous sacrifions à l'usage, nous regrouperons les mesures quantitatives de services (comme le nombre d'appendicectomies) dans le vecteur des agrégats d'outputs (noté Y) et les autres variables dans le vecteur de qualité (noté q_o). La fonction de coût s'écrira :

$$C\left(w, Y, q_o\right) \triangleq \min_x \left\{ w'x : f\left(x, Y, q_o\right) \leq 0 \right\}.$$

Le même raisonnement peut être utilisé du côté des inputs. En présence de diverses catégories de travail d'infirmières (notées $\left(x_1, ..., x_a\right)$ ou l'indice $i = 1, ..., a$ représente le nombre d'années d'expérience), on a diverses options :

- on peut simplement additionner les heures travaillées en supposant que les infirmières sont de parfaits substituts sans égard à leur expérience, i.e. $X\left(x_1, ..., x_a\right) = \sum_{i=1}^{a} x_i$, hypothèse souvent retenue; ou encore,
- on peut créer pour chacun des hôpitaux un indice d'ancienneté pour tenir compte des gains de productivité liés à l'expérience : $X\left(x_1, ..., x_a\right) = \left(\sum_{i=1}^{a} x_i, \sum_{i=1}^{a} i \frac{x_i}{\sum_{j=1}^{a} x_j} \right)$.[19]

Ici on passe de a catégories à seulement deux catégories.

En procédant comme précédemment, nous pouvons définir un vecteur de qualité des inputs (noté q_i) et un vecteur de quantité des inputs (noté X) avec le vecteur des prix d'inputs correspondant (noté W) et écrire la fonction de coût avec agrégats d'inputs :

$$C\left(W, q_i, Y, q_o\right) \triangleq \min_x \left\{ W'X : f\left(X, q_i, Y, q_o\right) \leq 0 \right\}.$$

[19] Cet indice d'ancienneté, et tout autre indice de qualité des inputs, est introduit dans le vecteur q_i.

Quand on utilise des données d'institutions, le besoin d'agrégation est encore plus important car le nombre d'outputs et d'inputs s'accroît. Plus il y a d'inputs et d'outputs, plus il faudra agréger. Dans ce contexte, le choix des indices de qualité (comme le taux de succès ou de mortalité, ou le degré d'ancienneté) devient de plus en plus important. Malheureusement, les indices de qualité n'abondent pas et il devient difficile de trouver un indicateur global satisfaisant. Cela explique pourquoi tant d'études se penchent sur cette question sans l'avoir résolue à la satisfaction générale.

Plusieurs études ont essayé de résoudre ce problème en ayant recours aux DRG. Les DRG sont en fait des agrégats basés sur des *a priori* quand à l'intensité relative des ressources utilisées pour traiter des patients. Si l'utilisation des DRG permet de réduire le nombre de variables, rien ne permet de penser que les biais d'agrégation ne soient pas importants.

V.2. Endogénéité de la production

Jusqu'à présent, nous avons supposé que l'output était exogène. Cela suppose une grande rigidité de la demande aux actions des hôpitaux, même dans le cas où les patients n'ont rien à payer. Les patients peuvent être attirés par la réputation de l'hôpital, la longueur des files d'attente, la qualité des soins telle que perçue par les patients, la présence d'autres institutions à proximité, la rigidité des contraintes imposées par les assureurs (droit de fréquenter un seul hôpital ou plusieurs au choix), etc.

Une façon de traiter ce problème est d'introduire la demande en tant que condition supplémentaire via la condition d'égalité entre le revenu marginal et le coût marginal. Le revenu marginal de l'institution i est donné par la demande qui s'adresse à l'institution en question. Naturellement, l'expression de cette demande dépend aussi des conditions de marché d'où l'inclusion de variables de concurrence (par exemple, indice d'Herfindhal, etc.).

Dans ce contexte, il faut estimer conjointement la fonction de coût (et les parts/demandes de facteur) et la condition d'optimalité sur l'output :

$$C^{obs} = C(w, y) + \text{ terme d'erreur de la fonction de coût}$$
$$S_i^{obs} = S_i(w, y) + \text{ terme d'erreur de la fonction de part}$$
$$Rm_i(y, \phi_i) = \frac{\partial C}{\partial y_i} + \text{ terme d'erreur de la condition d'optimalité sur l'output,}$$

où ϕ_i représente les caractéristiques de la demande et du marché pour l'output i.

Une alternative consiste à imposer la contrainte que l'offre de la firme est égale à la demande :

$$C^{obs} = C(w, y) + \sigma_C$$
$$S_i^{obs} = S_i(w, y) + \sigma_i$$
$$y_i = D_i(p, \phi_i) + \sigma_y.$$

où p est le vecteur des prix servant de signaux aux consommateurs et ϕ_i est le vecteur des autres informations et caractéristiques de la demande pour le services de santé i. On peut substituer cette condition dans la fonction de coût :

$$C^{obs} = C(w, D(p, \phi) + \sigma_y) + \sigma_C$$
$$S_i^{obs} = S_i(w, D(p, \phi) + \sigma_y) + \sigma_i.$$

Cette formulation est très proche de celle présentée plus haut. Les mêmes remarques s'imposent. Naturellement, on peut ajouter la question de l'erreur d'observation sur les prix et les caractéristiques de la demande.

Conclusion

Conclusion

Dans ce texte, notre but était de présenter un survol des méthodes visant à mesurer l'efficience des institutions de santé. Dans un document compagnon (Ouellette et Petit, 2010), nous présentons une revue de littérature de 600 articles publiés depuis les années 50 qui ont utilisé l'une ou l'autre des méthodes présentées ici et appliquées au domaine de la santé. Le lecteur intéressé s'y référera pour obtenir une longue liste d'exemples et de contributions souvent novatrices, parfois incohérentes. En fait, on comprend que ce survol s'applique à tout type d'industrie et que seuls les exemples portaient sur les institutions de santé.

En un sens, notre démarche a été l'inverse de celle empruntée par la plupart des chercheurs. Plutôt que de formuler un modèle économique *puis* d'ajouter une structure d'erreurs, nous avons commencé par définir les termes d'erreur *puis* nous avons incorporé les variables dans un modèle économique. Cette inversion en apparence banale a permis de formuler des modèles intrinsèquement cohérents tant du point de vue de la théorie économique que de celui des méthodes utilisées pour mesurer l'efficience. À la lumière de nos résultats, il ressort que les méthodes les plus courantes ne respectent pas l'ensemble des prescriptions de la théorie et que les modèles sont soit incohérents, soit incomplets.

Notre approche nous a aussi permis de préciser quelles étaient les hypothèses sous-jacentes des modèles présentement utilisés.

Nous ne nous sommes pas bornés à critiquer les travaux passés. Nous avons aussi précisé sous quelle forme les modèles devaient être spécifiés. Selon les outils de prédilection (statistique, recherche opérationnelle ou comptabilité), nous avons présenté divers modèles intrinsèquement cohérents. Ces modèles se caractérisent par la même facilité d'utilisation et ne requièrent aucune innovation en termes de méthodes économétriques ou de programmation mathématique.

Références

Blackorby, C., D. Primont et R. Russell (1978), « Duality, Separability, and Functional Structure: Theory and Economic Applications », North-Holland, Amsterdam.

Caves, D.W., L.R. Christensen et W.E. Diewert (1982), « The Economic Theory of Index Numbers and the Measurement of Input, Output, and Productivity », Econometrica, vol. 50, n° 6, p. 1393-1414.

Diewert, W.E., et C. Parkan (1983), « Linear Programming Tests of Regularity Conditions for Production Functions », p. 131-158, in W. Eichhorn, R. Henn, K. Neumann and R.W. Shephard (Eds.), Quantitative Studies on Production and Prices, , Wien: Physica Verlag.

Färe, R., S. Grosskopf, B. Lindgren et P. Roos (1992), « Productivity Changes in Swedish Pharamacies 1980–1989: A Non-parametric Malmquist Approach », Journal of Productivity Analysis, vol. 3, n° 1-2, p. 85-101.

Färe, R. et J. Logan (1983), «The Rate-of-Return Regulated Firm: Cost end Production Duality », The Bell Journal of Economics, vol. 14, no. 3, p. 405-414.

Gagné, R., et P. Ouellette (1998), « On the Choice of Functional Forms: Summary of a Monte Carlo Experiment », Journal of Business and Economic Statistics, vol. 16, n° 1, p. 118-124.

Gagné, R., et P. Ouellette (2002), « The Effect of Technological Change and Technical Inefficiencies on the Performance of Functional Forms », Journal of Productivity Analysis, vol. 17, n° 3, p. 233-247.

Ma, C.T.A. (1994), « Health Care Payment Systems: Cost and Quality Incentives », Journal of Economics and Management Strategy, vol. 3, p. 93-112.

McElroy, M.B. (1987), « Additive General Error Models for Production, Cost, and Derived Demand or Factor Share Systems », Journal of Political Economy, vol. 95, n° 4, p. 737-755.

Ouellette, P., P. Petit (2010), « Efficience budgétaire des institutions de santé : une revue de littérature », Centre sur la productivité et la prospérité, HEC Montréal.

Ouellette, P., P. Petit, L.-P. Tessier-Parent et S. Vigeant (2010), « Introducing Regulation in the Measurement of Efficiency, with an Application to the Canadian Air Carriers Industry », European Journal of Operational Research, vol. 200, p. 216-226.

Ouellette, P., et S. Vigeant (2001a), « Cost and Production Duality: The Case of the Regulated Firm », Journal of Productivity Analysis, vol. 16, n° 3, p. 203-224.

Ouellette, P., et S. Vigeant (2001b), « On the Existence of a Regulated Production Function », Journal of Economics, vol. 73, n° 2, p. 193–200.

Simar, L., et P.W. Wilson (1998) « Sensitivity Analysis of Efficiency Scores: How to Bootstrap in Nonparametric Frontier Models », Management Science, vol. 44, n° 1, p. 49-61.