Université de Montréal

Mesurer l'efficacité technique des établissements de soins de santé

-Portée et limites de la méthode DEA-

par
Catherine Gendron-Saulnier

sous la direction de
M. Abraham Hollander
et
M.Yves Richelle

Département de Sciences Économiques
Faculté des Arts et Sciences

Rapport de recherche présenté
en vue de l'obtention du grade de Maîtrise
en Sciences Économiques

Août, 2009

Résumé

Ce rapport de recherche tente d'analyser le potentiel de la méthode DEA afin d'évaluer l'efficacité technique des établissements d'un système de soins de santé. Un accent particulier est mis sur la signification d'une analyse effectuée par DEA pour les établissements du système de santé québécois. Les modèles CCR et BCC sont présentés de même que diverses extensions de ceux-ci afin de mener une discussion sur la façon la plus adéquate de construire une analyse DEA dans le domaine de la santé. Suite à une revue de plusieurs études empiriques, nous mettons également en perspective la manière dont nous pouvons interpréter les résultats de la méthode. À notre avis, les modèles CCR et BCC ne produisent qu'une évaluation partielle de l'efficacité technique. Ainsi, leurs résultats ne peuvent pas être interprétés comme une mesure spécifique de l'efficacité technique. Finalement, nous avançons que la méthode DEA, et ce, en utilisant un modèle bien particulier, le modèle ERM, ne produit que des indicateurs d'efficacité. Nous appuyons notre argument en présentant une série de propriétés que devrait posséder un indicateur d'efficacité et en montrant que seul le modèle ERM respecte l'ensemble de ces propriétés.

Mots-clés: Data Envelopment Analysis (DEA); Enhanced Russell Graph Efficiency Measure (ERM); Efficacité technique; Établissements de santé; Système de santé; Indicateur d'efficacité.

Table des matières

Table des figures

Mesurer l'efficacité technique des établissements de soins de santé: Portée et limites de la méthode DEA

1 Introduction

Les manchettes des quotidiens font état, presque tous les jours, des déficiences du système de santé québécois. Engorgement des salles d'urgence, délais d'attente excessifs, pénurie de main-d'oeuvre (médecins de famille, infirmières, etc.), surcharge de travail pour le personnel et erreurs médicales sont parmi les nombreux exemples qui ébranlent la confiance de la population dans le système de soins de santé. Plusieurs en viennent à douter de la capacité du système à répondre convenablement aux besoins de la population.

Pour certains, cette situation est attribuable à une crise de financement du système. Pourtant, au Québec, en 2009-2010 c'est près de 45 % du budget qui sera consacré à la santé et aux services sociaux (Secrétariat du Conseil du trésor, 2009), il en est de même dans les autres provinces canadiennes, et ce, malgré les nombreuses contraintes financières et économiques auxquelles font face les gouvernements. En fait, les dépenses publiques en matière de soins de santé en constituant une part si importante des budgets, ont atteint un seuil qu'il apparaît difficile d'excéder. De plus, certaines études ont démontré qu'il était impossible de déterminer le niveau optimal des dépenses en matière de santé et qu'il demeurait difficile de lier l'augmentation des dépenses en santé à une amélioration de l'état de santé de la population (Bardey et Pichetti, 2004). Supposer que la résolution des problèmes actuels puisse se résumer à une question d'investissements supplémentaires ou à une réorganisation du mode de financement nous apparaît réducteur. Nous croyons qu'une analyse d'un système de soins de santé doit s'attarder d'abord à évaluer l'efficacité du système, pour ensuite pouvoir aborder les questions de son financement.

En effet, l'importance budgétaire d'un système de soins de santé laisse entrevoir que si certaines composantes du système sont inefficaces, ces inefficacités seront probablement d'ampleur proportionnelle et donc passablement importante en termes monétaires. De plus, dans le cadre d'un système centralisé, comme c'est

le cas au Québec, l'absence des pressions concurrentielles laisse présager que des inefficacités substantielles et persistantes peuvent exister. Il est donc pertinent de s'attarder au développement de méthodes qui soient aptes à évaluer et à révéler l'existence d'inefficacités, et ce, dans la perspective de mettre en oeuvre des réformes et des politiques visant à diminuer, voire éliminer, le gaspillage des ressources.

À cet égard, bon nombre de groupes de travail et de publications gouvernementales posent la réflexion sur les améliorations à apporter au système sur les bases de l'efficacité. Cependant, s'il semble exister une volonté d'améliorer l'efficacité du système, ce qui permettrait d'offrir des services à la fois accessibles et de qualité et d'utiliser adéquatement les ressources financières, il faut néanmoins disposer d'outils de mesure qui sont adaptés aux particularités du domaine d'étude.

Voici donc l'objet de ce rapport de recherche, c'est-à-dire de démontrer dans quels cadres, ce qui inclut les méthodes et les hypothèses qu'elles posent, il est possible d'effectuer une analyse d'efficacité des établissements d'un système de soins de santé en accordant une attention particulière au contexte québécois. Il sera question de s'intéresser à la mesure empirique de l'efficacité, c'est-à-dire aux méthodes qui ont été développées pour mesurer le niveau d'efficacité d'une organisation quelconque ou d'un groupe d'organisations et de mener une discussion approfondie sur les particularités d'une analyse d'efficacité en santé et de la mesure dans laquelle les méthodes permettent de rendre compte de ces caractéristiques.

Il demeure légitime de se questionner sur la validité des méthodes, car le fait est qu'il n'est pas facile d'obtenir une évaluation empirique de l'efficacité. Bien que la notion d'efficacité occupe une place centrale dans les cours de théorie microéconomique qui exposent les diverses conditions nécessaires afin de l'établir et sur les nombreux mécanismes à mettre en place afin de la restaurer, la réalité est que bien peu de méthodes satisfaisantes de la mesure de l'efficacité ont été développées à ce jour, notamment parce que nous ne savons pas tout à fait ce que nous devons mesurer. Si le concept d'efficacité revêt pour le sens commun, un sens générique de l'atteinte de la plus grande productivité possible, pour le scientifique, l'opérationnalisation en une mesure quantifiable, exige cependant une définition plus restrictive.

L'analyse économique nous indique que le niveau de profits et d'autres signaux donnés par le marché peuvent représenter une mesure d'efficacité des firmes, mais

dans le cas où nous ne disposons pas d'un tel indicateur, comment pouvons-nous définir et mesurer ce qu'est l'efficacité?

Dans le secteur de la santé, tout comme dans plusieurs secteurs publics et dans certains secteurs des services, la difficulté est que l'efficacité est un concept propre au domaine étudié, c'est-à-dire qu'une multitude de facteurs extérieurs à la technologie de production doivent intervenir. Non seulement nous ne détenons pas de mesure de profitabilité, mais des défis conceptuels provenant de l'existence de plusieurs objectifs se posent de manière critique. Par exemple, devons-nous définir l'efficacité en fonction de la production d'un système de santé, soit en comptabilisant le nombre de soins qui sont prodigués? Devons-nous plutôt considérer l'efficacité en fonction des résultats obtenus, c'est-à-dire en analysant les indicateurs de mortalité et de morbidité des populations desservies par le système? Pour paraphraser Jacobs et al. (2006), les soins de santé ne sont pas demandés pour leur valeur intrinsèque, à la place, ils sont demandés parce qu'ils sont supposés contribuer de façon positive à l'état de santé d'un individu. Nous verrons au cours de ce rapport de recherche, qu'une foule d'autres questions du genre se posent et que celles-ci influenceront de manière significative l'analyse de l'efficacité dans le domaine de la santé.

À la question de la définition de l'efficacité d'un système de soins de santé est associée le problème de la mesure. Comment pouvons-nous évaluer le niveau d'efficacité, à partir de quelle référence? Pouvons-nous définir une fonction de production théorique et mesurer les déviations des organisations à partir de celle-ci? Parallèlement, connaissons-nous le meilleur niveau d'efficacité possiblement atteignable? Dans le contexte d'un système de soins de santé, de telles questions sont véritablement répondues par la négative et c'est là que se pose tout le problème de la mesure. Conséquemment, nous montrerons que les méthodes économétriques traditionnelles dont fait usage la théorie économique pour estimer des fonctions de production et de coûts sont insuffisantes et que des méthodes alternatives sont nécessaires. Parmi ces méthodes, nous identifierons la méthode de la frontière stochastique (SFA) et la méthode du *Data Envelopment Analysis* (DEA), mais c'est à cette dernière que nous nous attarderons spécifiquement. C'est elle qui constituera le cadre d'analyse de ce rapport de recherche.

Précisément, notre objectif consistera à faire une revue générale, mais formelle, de la méthode DEA en discutant des hypothèses posées et de leur importance dans

le cadre d'une analyse d'efficacité dans un système de santé comme le système québécois. Nous aborderons également la manière dont cette méthode a permis de mener des analyses empiriques en faisant la revue de plusieurs études. Une fois que ceci sera accompli, nous pourrons nous interroger sur les conclusions de la méthode DEA, à savoir si la méthode produit des résultats qui sont inhérents à la façon dont elle est construite et les hypothèses qu'elle sous-tend, ou si elle est réellement apte à évaluer l'efficacité. Ensuite, nous discuterons de la portée de la méthode DEA en regard de deux aspects: le calcul du niveau d'efficacité et son utilisation en tant qu'indicateur d'efficacité.

Plus qu'une simple revue de la littérature, ce travail a l'ambition de mener une discussion la plus exhaustive possible sur les aspects fondamentaux de l'efficacité dans un système de soins de santé et de leurs conséquences pour l'outil de mesure sélectionné, la méthode DEA. Bien qu'une multitude de papiers s'attardent soit à appliquer la méthode à des secteurs précis en santé, soit à discuter théoriquement de certaines de ses hypothèses et de leur influence sur la mesure DEA, aucun n'a cependant suggéré une analyse comme nous proposons de le faire dans ce rapport de recherche.

Pour finir, ce rapport de recherche est organisé de la manière suivante: la section 2 présente les différentes méthodes de mesure d'efficacité, la section 3 s'attarde à la présentation de la méthode DEA de manière formelle, la section 4 discute de plusieurs extensions techniques de la méthode DEA et de la contrepartie pratique de ces ajouts dans le domaine de la santé, la section 5 fait la revue d'une multitude d'études empiriques, finalement, la section 6 traite de la construction d'un d'indicateur d'efficacité à partir des résultats de la méthode DEA.

2 Définir et mesurer l'efficacité

2.1 Farrell et la mesure de l'efficacité empirique

Parmi les articles fondateurs qui concernent l'analyse empirique de l'efficacité au niveau microéconomique, nous comptons l'article «*The Measurement of Productive Efficiency*» de M. J. Farrell publié en 1957. Dans ce papier, ayant exercé une influence considérable dans le développement de la mesure de l'efficacité, Farrell

y adresse deux questions principales (Farrell, 1957): comment définir l'efficacité et la productivité? Comment mesurer et calculer celle-ci?

À cette première question, Farrell propose de décomposer l'efficacité des organisations en fonction de trois types d'efficacité qu'il définit comme suit:

- Efficacité technique: la capacité à produire un maximum d'outputs possible à partir d'une quantité d'inputs donnée;
- Efficacité allocative: la capacité à combiner les inputs en proportions optimales en fonction de leur prix et de la technologie afin de produire une quantité d'outputs maximale;
- Efficacité totale: cette dernière étant le produit des deux autres.

Selon Farrell, il est possible *a priori* que des firmes soient inefficaces en regard d'un de ces aspects. Une mesure de l'efficacité se doit alors d'être menée à partir d'une frontière d'un ensemble des possibilités de production, plutôt qu'à partir d'une analyse économétrique standard, car celle-ci en recherchant l'équation d'une droite décrivant le mieux possible les observations, traite les données extrêmes comme des données aberrantes et produit une évaluation moyenne des performances (Førsund et Sarafoglou, 2002). Plutôt, ce serait l'ensemble des données qui pourrait contribuer à l'estimation de l'efficacité, le niveau extrême de certaines performances pourrait identifier des pratiques efficaces ou encore inefficaces. Ainsi, en proposant le recours à l'estimation d'une frontière, Farrell énonce véritablement un changement de paradigme (Badillo et al., 1999).

La contribution principale de Farrell consiste donc d'abord, à obtenir l'ensemble des possibilités de production, à déterminer sa frontière et enfin, à mesurer l'efficacité comme étant la distance qui sépare une entité de cette frontière. L'ampleur de la déviation à la frontière serait donc interprétée comme une mesure de l'inefficacité des firmes (Farrell, 1957).

L'illustration de la page suivante (Farrell, 1957) permet de saisir les concepts tels qu'ils sont définis dans l'article de Farrell.

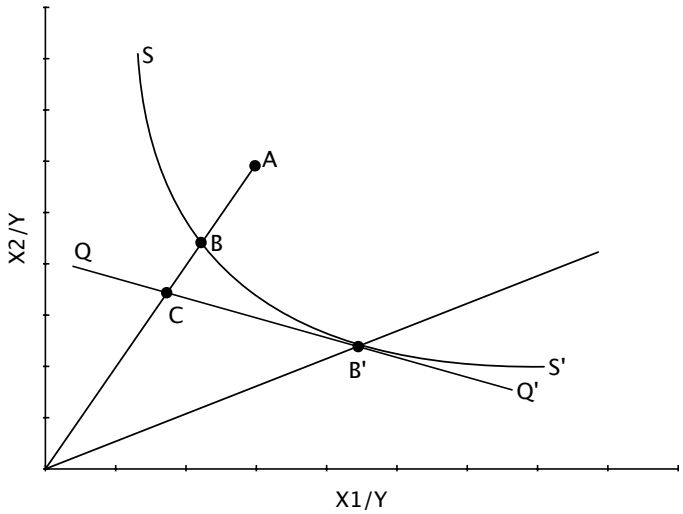

Fig. 1: Illustration de l'efficacité selon Farrell

Le segment $\overline{SS'}$ est l'isoquant efficace de production, c'est-à-dire l'ensemble des combinaisons possibles d'inputs 1 et 2 minimums qu'une firme efficace peut utiliser afin de produire une quantité donnée d'outputs. Puisque nous avons normalisé les deux inputs par la quantité d'outputs, nous pouvons interpréter le segment $\overline{SS'}$ comme l'ensemble des combinaisons possibles d'inputs 1 et 2 minimums afin de produire une unité d'output.

Le segment $\overline{QQ'}$ représente un segment dont la pente correspond au rapport des prix des deux inputs. Les firmes B et A utilisent la même proportion de chacun des inputs puisqu'elles sont situées sur le même rayon. Cependant, B produit le même niveau d'output à une fraction $^{OB}/_{OA}$ des inputs utilisés par A. La firme B est donc plus efficace que la firme A, et la mesure d'inefficacité technique de A est donnée par ce rapport $^{OB}/_{OA}$.

La théorie économique indique qu'une firme efficace du point de vue allocatif produit au point où le taux marginal de substitution de ses inputs est égal au rapport des prix de ceux-ci. Dans ces circonstances, c'est le point B' qui est efficace par rapport au prix des inputs et non B. Les coûts de production au point B' seront une fraction $^{OC}/_{OB}$ de ce qu'ils sont au point B. Il est donc sensé de croire que ce rapport représente l'efficacité allocative de la firme A.

Nous pouvons alors combiner ces deux mesures d'efficacité pour obtenir l'efficacité totale de la firme A, ${}^{OB}/_{OA} \times {}^{OC}/_{OB} = {}^{OC}/_{OA}$.

La figure 1 suppose le segment $\overline{SS'}$ connu, mais en réalité ce ne sera que rarement le cas. Il sera, en général, nécessaire d'estimer cette frontière. Farrell, sans toutefois être explicite sur la façon d'obtenir cette frontière, spécifie certaines des propriétés qu'elle doit posséder. Premièrement, il doit s'agir du *«most pessimistic piecewise linear envelopment of the data»* (Farrell, 1957, p.256), c'est-à-dire le segment qui est le plus près possible des données. Ensuite, en aucun point sa pente ne doit être positive et aucun point ne doit être compris entre la frontière et l'origine du plan cartésien.

Si le travail amorcé par Farrell reste considérablement important en regard de l'analyse qu'il propose de mener, il demeure que des méthodes pour estimer les frontières doivent toutefois encore être développées. C'est donc autour de cet objectif que les recherches seront orientées. Celles-ci pourront se distinguer en fonction de leur caractère paramétrique ou non paramétrique et de leur caractère déterministe ou stochastique. Les prochaines sections s'affairent à présenter ces différentes méthodes.

2.2 De multiples modèles

2.2.1 Modèle de la frontière stochastique[1]

Du point de vue économétrique, l'estimation de frontières de production est intéressante, car le concept de «maximalité» impose une limite à la variable dépendante (Førsund et al., 1980). L'estimation de la frontière de production à l'aide des méthodes de régressions usuelles n'est pas appropriée puisque l'idée qui consiste à trouver une fonction moyenne implique que certaines firmes pourraient se retrouver au-dessus de la frontière à l'image de la figure 2, et donc apparaîtrait comme surefficaces. Pour cette raison, l'intérêt a été porté sur la spécification de modèles qui puissent accommoder cette notion de frontière. De nombreuses propositions ont émergé de cette réflexion, toutefois, nous nous concentrerons uniquement sur le modèle de la frontière stochastique (SFA) puisqu'il nous semble être le plus achevé des modèles économétriques[2].

[1] La présentation de cette section est inspirée de Coelli et al. (2005, pp. 241-288.)

[2] Pour un survol des autres méthodes, consulter Førsund et al. (1980).

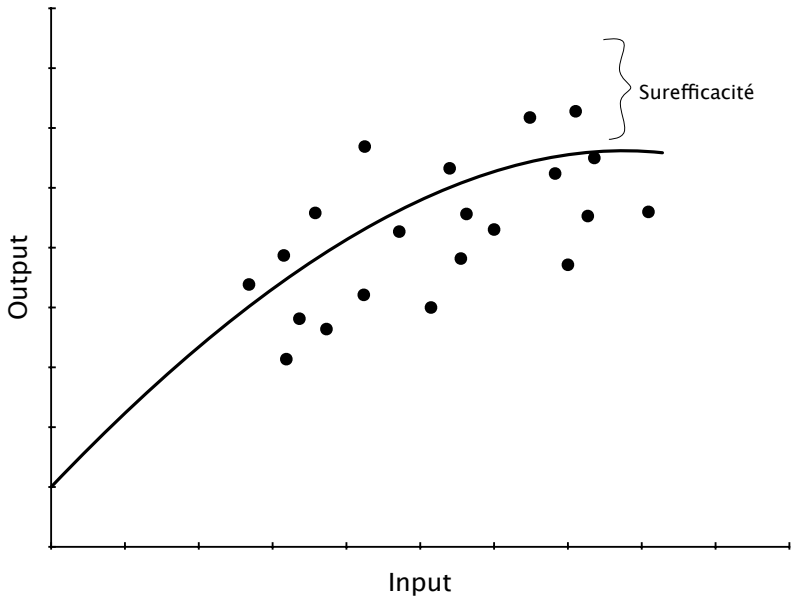

Fig. 2: Un modèle de régression standard

Le modèle de la frontière stochastique est proposé de manière indépendante à la fois par Aigner, Lovell et Schmidt (1977) et par Meeusen et van den Broeck (1977) comme une méthode permettant d'estimer une frontière de production à caractère paramétrique qui inclut une composante stochastique. En vertu de cette méthode, le niveau de production d'une firme, y_i , est le produit d'une composante déterministe des inputs x_i dont la forme est donnée sans perte de généralités par une fonction Cobb-Douglas, du bruit statistique u_i et de l'inefficacité v_i.

$$ln\ y_i = x_i'\beta + u_i - v_i$$
$$y_i = exp(x_i'\beta) \times exp(u_i) \times exp(-v_i)$$

Le terme d'erreur se compose de deux éléments: d'abord, une composante symétrique permettant des variations aléatoires qui prennent compte des erreurs de mesure, des autres bruits statistiques et des chocs exogènes aux firmes, soit la composante symétrique u_i qui est dotée des propriétés standards du modèle de régression classique (espérance nulle, identiquement et indépendamment distribué (i.i.d.)), ensuite, d'une composante asymétrique qui capture l'inefficacité

des firmes en rapport à la frontière, soit la composante v_i qui est une variable aléatoire non négative d'espérance non nulle et qui est distribuée de façon i.i.d. également.

Ce genre de modèle peut être estimé par la méthode des MCO corrigés (COLS[3]) ou préférablement par maximum de vraisemblance en supposant que la variable u_i est normalement distribuée et que la variable v_i suit une loi asymétrique comme une loi normale tronquée ou encore une loi gamma.

Une fois l'estimation des paramètres complétée, une mesure de l'efficacité technique pour une firme s'obtient en comparant la production réelle au niveau de production efficace prédite par le modèle, ce qui inclut la composante déterministe et le bruit statistique et exclut la composante d'inefficacité.

Il importe de prendre note que les hypothèses postulées en ce qui concerne à la fois la distribution de l'inefficacité et du bruit statistique et la relation entre les inputs et les outputs sont infiniment importantes en regard des résultats qui seront obtenus. L'absence d'indication ou d'intuition théorique pour justifier ce genre de choix est alors l'un des principaux obstacles à l'utilisation de ce type de méthode dans des analyses d'efficacité.

2.2.2 Modèle du Data Envelopment Analysis

Du côté non paramétrique, l'une des méthodes les plus importantes est sans aucun doute la méthode du *Data Envelopment Analysis* (DEA). Elle fait son apparition en tant que méthode unifiée pour la première fois dans un article publié en 1978 par A. Charnes, W.W. Cooper et E. Rhodes.

Ces auteurs contribuent de manière fondamentale au développement des analyses de performance en proposant le recours à une méthode de programmation linéaire pour estimer une frontière de production et le niveau d'efficacité des firmes. Leur idée de départ se résume en quelques mots (Cooper et al., 2005):

> *As a fundamental assumption behind the computation of [...] efficiency is that if a given firm A is capable of producing Y(A) using X(A) inputs, then other firms should be also be able to do the same if they were to operate efficiently*

[3] Corrected Ordinary Least Squares.

Il s'agit donc de développer une méthode afin de comparer entre elles les performances d'organisations faisant partie d'un ensemble quelconque. L'hypothèse qui est sous-jacente à une telle entreprise réfère à la difficulté de connaître le niveau théorique de production qu'il serait possible d'atteindre dans les domaines des sciences sociales et en management. Pour cette raison, la mesure de l'efficacité doit être fondée empiriquement et définie selon des termes relatifs, c'est-à-dire sur la base des meilleures performances observées des organisations d'un ensemble qui nous intéresse.

La méthode DEA tente alors d'identifier quelles sont les firmes avec ces meilleures performances, dans l'objectif de construire une frontière d'efficacité au sens de Farrell. Ensuite, la méthode procède à l'évaluation de l'efficacité de chacune des organisations en mesurant la distance de celles-ci par rapport à la frontière d'efficacité.

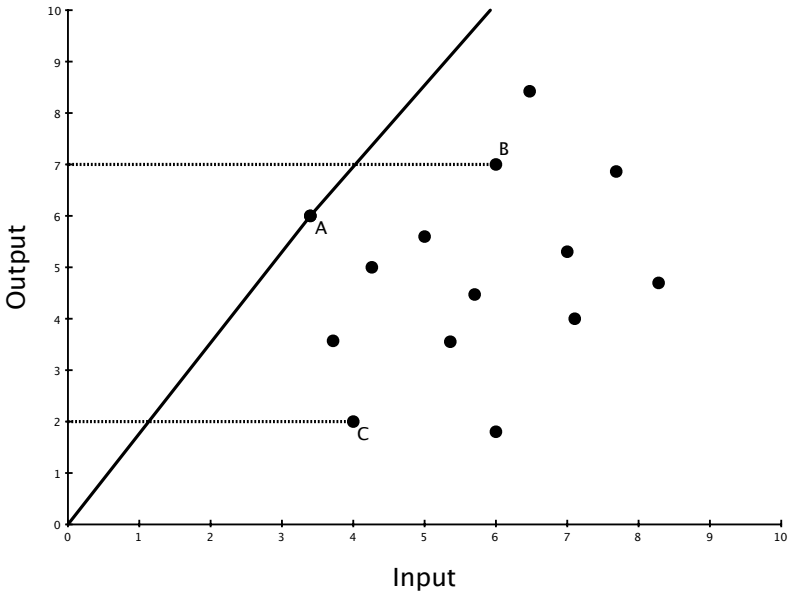

Fig. 3: Un input et un output

La figure 3 donne un exemple de l'analyse DEA. Nous y retrouvons un ensemble de points qui illustre la consommation en input et la production d'output de 14 firmes. La meilleure performance est vraisemblablement atteinte par la firme A, car c'est elle qui produit le plus d'output pour un niveau d'input donné. Le segment partant de l'origine et passant par le point A représente alors la frontière d'efficacité du modèle DEA. La firme A étant la seule à être située sur ce segment

est l'unique firme efficace de l'ensemble. Les 13 autres firmes sont donc inefficaces, ce qui signifie qu'elles auraient pu réduire leur consommation de l'input pour produire la même quantité d'output et ainsi atteindre la frontière. Le niveau d'inefficacité de ces firmes est évalué à partir de la distance de chacune d'entre elles de la frontière d'efficacité et se mesure comme le ratio de la distance de l'axe des ordonnées à la frontière sur la distance de l'axe des ordonnées à la position de la firme. Pour la firme B cette mesure est donnée par le ratio $4/6 = 0.6667$, et pour la firme C celui-ci est de $1.1/4 = 0.275$.

Si nous considérons maintenant deux inputs et toujours un seul output tel qu'illustré à la figure 4, la façon de procéder demeure assez similaire. Les firmes A et D sont les firmes efficaces de l'ensemble puisque ce sont elles qui consomment le moins de ressources pour produire une unité d'output. Elles définissent alors la frontière qui s'interprète comme un isoquant à l'image de celui de Farrell et à partir de laquelle le niveau d'inefficacité des autres firmes sera calculé. Ce calcul s'effectue encore en considérant que les firmes inefficaces auraient pu réduire proportionnellement la consommation des deux inputs et toujours produire le même niveau d'output si elles avaient été efficaces. Cette réduction s'effectue sur le rayon reliant une firme à l'origine de sorte à conserver la proportion entre les deux inputs. Par exemple, l'inefficacité au point B est donnée par le ratio $3/4 = 0.75$ et celle au point C par le ratio $4/6 = 0.6667$.

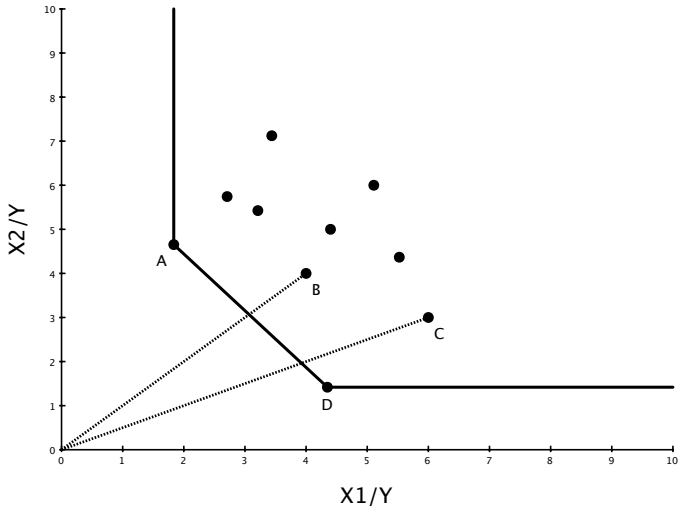

Fig. 4: Deux inputs et un output

Comme nous venons de le démontrer, la logique derrière la méthode DEA est de comparer la performance courante des organisations à la meilleure performance possible, soit celle qui se situe sur la frontière efficace, que chacune pourrait vraisemblablement atteindre.

Essentiellement, la méthode DEA fait intervenir une représentation graphique de l'efficacité et de sa mesure. Toutefois, le recours à la technique de la programmation linéaire est nécessaire afin d'obtenir une solution plus générale qui permette d'augmenter la dimensionalité du problème. En effet, nous avons simplifié la présentation de la méthode en ne l'illustrant qu'avec un ou deux inputs et un seul output. Il faut cependant noter que la méthode peut accommoder jusqu'à m inputs et s outputs et que dans ce cas, la représentation graphique devient impossible.

Bien que notre présentation de la méthode DEA demeure assez synoptique, nous pouvons déjà apprécier certaines de ses particularités. L'utilisation de la méthode DEA comme méthode pour analyser l'efficacité donne la possibilité d'ignorer les prix, d'avoir à formuler des hypothèses sur l'importance relative à accorder à certains inputs ou outputs, tout comme elle ne nécessite aucune spécification d'une fonction de production comme c'était le cas pour les méthodes paramétriques.

2.3 Point de vue adopté dans cette recherche

La mesure de l'efficacité est un vaste sujet de recherche au sein duquel il est facile de perdre le fil, en partie parce qu'il existe une multitude de méthodes disponibles. C'est pourquoi il importe de préciser les aspects sur lesquels nous nous concentrerons dans ce rapport de recherche.

D'abord, rappelons que l'objectif premier est de mener une discussion sur la façon de poser une analyse d'efficacité dans le domaine de la santé. Ce faisant, nous devons adopter une définition de l'efficacité qui soit appropriée à ce secteur.

Nous devons noter que le processus de production d'un système de soins de santé est complexe, non seulement à cause de la structure du marché, mais également parce que la production en tant que telle est difficile à définir. Comme le note Jacobs et al. (2006) : «*Not only do prices not exist, but outputs are difficult to define. Health is a complex concept for which there has been no readily available*

valuation and there is no market for health in the conventional sense». Dans le contexte québécois, mais également dans un cadre plus général, cette complexité et la prévalence d'un système public ont pour conséquence de rendre laborieuse une analyse d'efficacité allocative.

Cela dit, une analyse d'efficacité technique semble, quant à elle, plus appropriée pour certaines raisons. Premièrement, s'attarder à la mesure de l'efficacité technique est une première étape qui cadre avec une analyse plus générale qui viserait à mesurer l'efficacité totale. Ensuite, comme nous l'attesterons tout au long de ce rapport, les établissements de santé au Québec sont certainement limités dans leur choix lorsqu'il est question de l'allocation des ressources et de budgets, les décisions importantes étant prises au niveau des agences régionales et du Ministère. Par conséquent, ce sont les décisions dans la gestion des ressources, c'est-à-dire sur la façon dont elles seront organisées et interagiront au sein des établissements de santé afin de produire un système de soins, qui ultimement relèvent des diverses organisations. En ce sens, une analyse d'efficacité technique semble s'arrimer de façon plus réaliste à la nature des décisions sous le contrôle des établissements de santé québécois.

Nous poursuivons maintenant en présentant la définition de l'efficacité technique que nous utiliserons au cours de ce rapport. Cette définition se fonde sur les notions d'efficacité de Pareto et de Koopmans (Cooper et al., 2004):

> *Une organisation est efficace techniquement, si et seulement si, il est impossible pour cette organisation d'augmenter le niveau d'outputs produit pour un niveau de ressources donné, ou encore de diminuer le niveau de ressources consommées pour produire une quantité donnée d'output.*

Finalement, il importe de spécifier que le cadre d'analyse qui sera utilisé dans cette recherche est celui proposé par la méthode DEA. Nous tenterons donc d'apprécier la façon dont celle-ci propose d'effectuer une analyse d'efficacité en santé. Ce choix de la méthode DEA se justifie en deux temps. D'une part parce qu'elle est certainement la méthode qui est de loin la plus utilisée lorsqu'il est question d'analyser l'efficacité dans le domaine de la santé. Hollingsworth (2008) rapporte que plus de 67% des études dans ce domaine utilisent la méthode DEA sous une de ses formes. D'autre part, parce qu'elle présente des avantages intéressants en comparaison des méthodes stochastiques. Parmi ces avantages, notons

le fait que la méthode DEA analyse chacune des organisations séparément par rapport à l'ensemble des données en déterminant l'efficacité de chacune par rapport au groupe de pairs ayant la meilleure pratique, qu'elle ne nécessite aucune paramétrisation, qu'elle prenne en compte facilement un grand nombre d'outputs et qu'elle évalue simultanément la contribution de toutes les variables à la mesure de l'efficacité.

3 Cadre d'analyse: la méthode DEA

Cette section est destinée à présenter de manière plus précise la méthode DEA, en exposant ses postulats de base, les programmes linéaires auxquels elle a recours afin de mesurer l'efficacité de chacune des organisations d'un ensemble et la signification de ces hypothèses pour une analyse d'efficacité dans un contexte général.

3.1 Construction de l'ensemble de production

Comme nous l'avons souligné, la méthode DEA est une méthode qui est dirigée à l'approximation de la frontière de l'ensemble de production à partir d'observations d'une population d'unités de production, communément appelées les DMUs[4]. Dans le cadre de ce rapport, les DMUs pourront prendre diverses formes. En général, il s'agira d'établissements producteurs de soins de santé, il pourra s'agir de centres hospitaliers, de centres de soins pour personnes âgées, de départements de chirurgie, de salles d'urgence, etc.

L'idée de départ de la méthode est qu'en l'absence de prérogatives sur la technologie de production d'un secteur, les performances observées peuvent être utilisées comme évidences pour estimer ce qui est réalisable, c'est-à-dire un ensemble de production, et identifier du même coup les meilleures réalisations, soit un sous-ensemble de production efficace constituant une frontière d'efficacité.

[4] Le terme DMU est l'acronyme du terme anglais *decision making unit*. La définition adoptée des DMUs reste par ailleurs assez large. En général, il s'agit d'entités responsables de convertir des inputs en outputs et desquelles nous désirons évaluer l'efficacité. À défaut de trouver une traduction adéquate, nous utiliserons ce terme et son acronyme pour désigner les unités dont nous désirons évaluer l'efficacité.

La première étape à la mesure de l'efficacité est la définition d'un ensemble de production. Il faut définir ce que les DMUs ont la possibilité de réaliser dans le plan input-output, ce n'est qu'ensuite qu'il sera possible de déterminer quelles combinaisons d'inputs et d'outputs sont efficaces et enfin d'évaluer les réalisations des organisations en rapport à cet ensemble efficace.

Alors, supposons que nous détenons un ensemble d'observations (X_j, Y_j) $j = 1, ..., n$, où n est le nombre de DMUs, $X_j = (x_{j1}, x_{j2}, ..., x_{jm})$ représente un vecteur de m inputs et $Y_j = (y_{j1}, y_{j2}, ..., y_{js})$ un vecteur de s outputs.

Posons T, l'ensemble de production que tente de définir la méthode DEA. Dans sa forme la plus restrictive, cet ensemble doit posséder les propriétés suivantes (Charnes et al., 1978):

1. Convexité: Si $(X_j, Y_j) \in T, j = 1, ..., n$ et $\lambda_j \geq 0$ sont des scalaires non négatifs tels que $\sum_{j=1}^{n} \lambda_j = 1$, alors $(\sum_{j=1}^{n} \lambda_j X_j, \sum_{j=1}^{n} \lambda_j Y_j) \in T$.

2. Libre disponibilité forte ou propriété d'inefficacité: i) Si $(X, Y) \in T$ et $\bar{X} \geq X$, alors $(\bar{X}, Y) \in T$; ii) Si $(X, Y) \in T$ et $\bar{Y} \leq Y$, alors $(X, \bar{Y}) \in T$.

3. Expansion radiale: Si $(X, Y) \in T$ alors $(\kappa X, \kappa Y) \in T \quad \forall\, \kappa > 0$.

4. Extrapolation minimale: T est l'intersection de tous les $\hat{T}$ satisfaisants les propriétés $1, 2$ et 3 et pour lesquels chacune des observations $(X_j, Y_j) \in \hat{T},\ j = 1, ..., n$.

La méthode DEA cherche donc à construire le plus petit ensemble convexe qui puisse contenir les observations. En utilisant ce principe de l'extrapolation minimale, la méthode DEA identifie la frontière des meilleures pratiques, puisque nous ignorons où se situe la véritable frontière de production (Agrell et Bogetoft, 2001). En ce sens, la frontière DEA est une estimation conservatrice de l'ensemble de production, elle est le standard d'efficacité minimum étant donné les observations.

La formulation mathématique DEA nous donnera la possibilité d'illustrer comment la méthode recherche un sous-ensemble de l'ensemble analysé qui constituera une surface d'enveloppement pour les observations, à savoir cette frontière d'efficacité à laquelle nous référons depuis le début de ce travail. Nous verrons également comment la méthode procède pour disposer d'une mesure d'efficacité pour chacune des organisations analysées et identifier les sources d'inefficacités et leur ampleur.

3.2 Formulation mathématique du problème

Nous exposons ici la formulation mathématique de la méthode DEA dont la présentation est tirée de Cooper et al. (2007). Nous ferons référence au modèle présenté comme le modèle CCR, du nom de ses auteurs: W.W. Cooper, A. Charnes et E. Rhodes.

Rappelons que nous désirons évaluer n DMUs, où chacune consomme une quantité variable de m inputs différents afin de produire s différents outputs. Plus précisément, la DMU$_j$ utilise x_{ij} d'input i et produit une quantité y_{rj} d'output r. Pour des besoins notationnels, nous réfèrerons à une DMU particulière comme étant la DMU$_o$.

Pour faire l'évaluation de cette DMU$_o$, le modèle CCR procède à la construction d'un indice de productivité qui inclut tous les inputs utilisés et tous les outputs produits par celle-ci. Ce ratio prend la forme suivante :

$$\textit{Indice de productivité } CCR = \frac{\textit{Output virtuel}}{\textit{Input virtuel}}$$

Les notions d'output et d'input virtuels signifient que chacun des éléments du ratio est pondéré d'une façon particulière. De la sorte, il s'agit donc de construire cet indice de productivité en cherchant les pondérations à accorder à chacun des outputs et des inputs. Plus précisément, le problème consiste à trouver les poids optimaux, c'est-à-dire ceux qui maximisent le ratio de la DMU$_o$ sous un certain nombre de contraintes. Le problème à résoudre est alors:

$$max_{v,u} \quad h_o = \frac{\sum_{r=1}^{s} u_r y_{ro}}{\sum_{i=1}^{m} v_i x_{io}} \tag{1}$$

$$sujet\ à \quad \frac{\sum_{r=1}^{s} u_r y_{rj}}{\sum_{i=1}^{m} v_i x_{ij}} \leq 1 \quad j = 1, ..., n \tag{2}$$

$$u_r, v_i \geq 0 \quad \forall\, i, r \tag{3}$$

Il s'agit donc de trouver un vecteur de poids (u^*, v^*) qui permet de maximiser le ratio des outputs sur les inputs, soit le score h_o, pour la DMU$_o$ tout en considérant que ces poids sont également réalisables pour l'ensemble des autres DMUs, c'est-à-dire que $h_j(u_o^*, v_o^*) \leq 1, \ \forall\, j$. La recherche du ratio maximum se justifie par le besoin de faire apparaître la DMU$_o$ la plus performante possible sur la base des performances des autres DMUs de sorte à pouvoir identifier les meilleures performances de l'ensemble.

Notons que les poids optimaux (u^*, v^*) s'interprètent comme la contribution marginale d'une unité de chaque input ou de chaque output au score d'efficacité h_o.

Afin d'obtenir un vecteur de poids et un score d'efficacité h_o pour chacune des DMUs, il faut résoudre n problèmes similaires à (1). Nous disposons donc ainsi d'une évaluation $h_o \in [0, 1]$ pour chacune des unités analysées de l'ensemble. Les DMUs caractérisées comme étant efficaces seront celles dont la valeur de h_o est égale à 1, ce sont celles-ci qui définiront la frontière d'efficacité.

Toutefois, remarquons que la formulation fractionnaire en (1) comporte une infinité de solutions. En effet, si (u^*, v^*) est une solution optimale, alors $(\beta u^*, \beta v^*)$ est aussi optimale pour $\beta > 0$. Le problème est donc linéarisé pour obtenir une solution représentative en ajoutant une contrainte normalisatrice sur $\sum_{i=1}^{m} v_i x_{io}$.

Le problème devient:

$$max_{v,u} \quad z_o = \sum_{r=1}^{s} u_r y_{ro} \tag{4}$$

$$sujet\ à \quad \sum_{r=1}^{s} u_r y_{rj} - \sum_{i=1}^{m} v_i x_{ij} \leq 0 \quad j = 1, ..., n \tag{5}$$

$$\sum_{i=1}^{m} v_i x_{io} = 1 \tag{6}$$

$$u_r, v_i \geq 0 \quad \forall\, i, r \tag{7}$$

Nous pouvons recourir à la théorie de la programmation linéaire afin d'écrire ce problème sous sa forme dual, cette manipulation facilitera la résolution du problème linéaire de même que son interprétation:

$$min_{\lambda,\theta} \quad \theta \tag{8}$$

$$sujet\ à \quad \sum_{j=1}^{n} x_{ij}\lambda_j = \theta x_{io} - s_i^- \quad i = 1, ..., m \tag{9}$$

$$\sum_{j=1}^{n} y_{rj}\lambda_j = y_{ro} + s_r^+ \quad r = 1, ..., s \tag{10}$$

$$\lambda_j, s_i^-, s_r^+ \geq 0 \quad \forall i, j, r \tag{11}$$

En fonction du théorème de la dualité en programmation linéaire, nous avons l'égalité des solutions optimales des problèmes (4) et (8), c'est-à-dire que $z_o^* = \theta^*$. Ainsi, une DMU n'est efficace que si $\theta^* = 1$. Le théorème de Kuhn-Tucker des *complementary slackness conditions* ou des «conditions de relâchement supplémentaire» nous indique également que nous avons la relation suivante entre le problème primal en (4) et le problème dual en (8): $v_i^* s_i^{-*} = 0 \quad \forall\, i$ et $u_r^* s_r^{+*} = 0 \quad \forall\, r$, où les variables s_r^+ et s_i^- représentent les variables d'écart, communément appelées *slacks* dans la littérature.

Alors que le problème primal en (4) recherche les pondérations optimales à accorder à chacun des inputs et des outputs dans la construction d'un indice de productivité, le problème dual en (8), quant à lui, recherche les poids λ à accorder à chacune des DMUs de l'ensemble de façon à minimiser le coefficient θ d'utilisation des ressources de la DMU$_o$ pour un niveau d'output donné. Ceci signifie que nous recherchons une combinaison de DMUs qui produit le même niveau d'output que la DMU$_o$ et qui utilise les inputs dans la même proportion, mais à un niveau qui soit le plus inférieur possible. Ce niveau est donné par le facteur θ minimum du problème que nous désignerons comme une mesure d'efficacité radiale.

Les contraintes (9) et (10) incluent chacune des variables d'écarts puisqu'il est possible qu'une combinaison construite à partir des lambdas ne puisse pas reproduire la même proportion dans l'utilisation des inputs ou dans la production des outputs que la DMU$_o$. Nous montrerons un peu plus loin ce que cela signifie.

Pour l'instant, passons à la figure 5 qui présente bien ce dont nous parlons avec la combinaison de DMU utilisant les ressources à une proportion θ de la DMU$_o$. Nous avons 7 DMUs, A, B, C, D, E, F et G utilisant chacune deux inputs x_1, x_2 pour produire un output y. Elles sont représentées dans le plan $x_1/y, x_2/y$. Prenons le point D et recherchons la combinaison constituée des autres éléments de l'ensemble qui utilise la même proportion d'inputs que celui-ci, mais à un niveau qui soit le plus petit possible. Nous recherchons alors une combinaison de DMUs qui soit située sur le rayon partant de l'origine jusqu'au point D. Comme nous pouvons le constater sur la figure, celle-ci se situe au point D' et est une composition des points A et C. La mesure θ qui correspond à la proportion des inputs du point D qui est utilisée au point D' sera alors inférieure à 1. La DMU D est donc inefficace puisque les performances des autres DMUs démontrent qu'il aurait été possible de n'utiliser qu'une proportion θ des ressources qu'elle utilise.

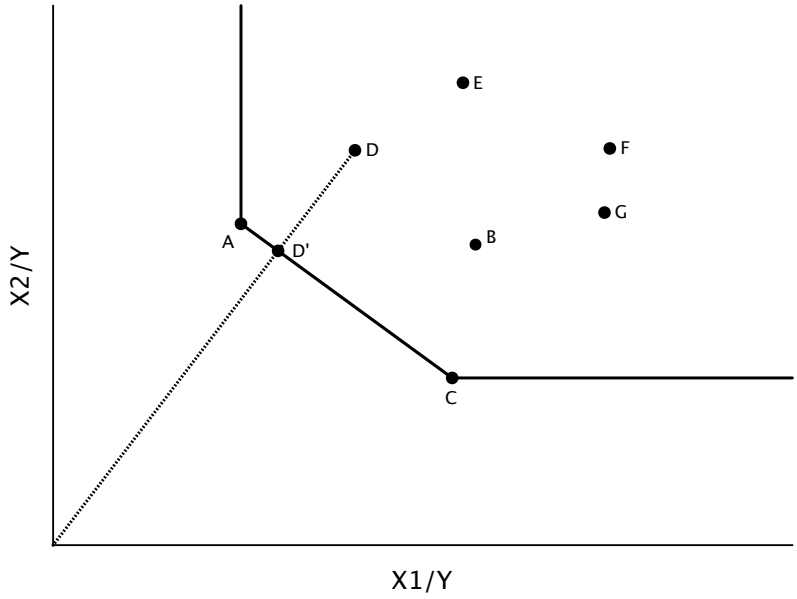

Fig. 5: À la recherche d'une combinaison de DMUs

Si nous nous attardons à l'évaluation de l'ensemble des autres points, nous remarquons que tous peuvent être reproduits comme une combinaison des points A et C comme en témoigne la figure 6 de la page suivante. Les points B, E, F et G seront donc aussi inefficaces et seuls les DMUs A et C seront efficaces. Ce sont ces deux points qui formeront la frontière. Nous retrouvons de plus deux segments parallèles aux axes, un partant du point A et l'autre partant du point C. Ces

segments constituent également une part de la frontière. Comme nous n'avons pas observé d'autres organisations ayant combiné les ressources à un niveau inférieur à A et C nous ne pouvons pas étendre l'ensemble des possibilités de production plus loin.

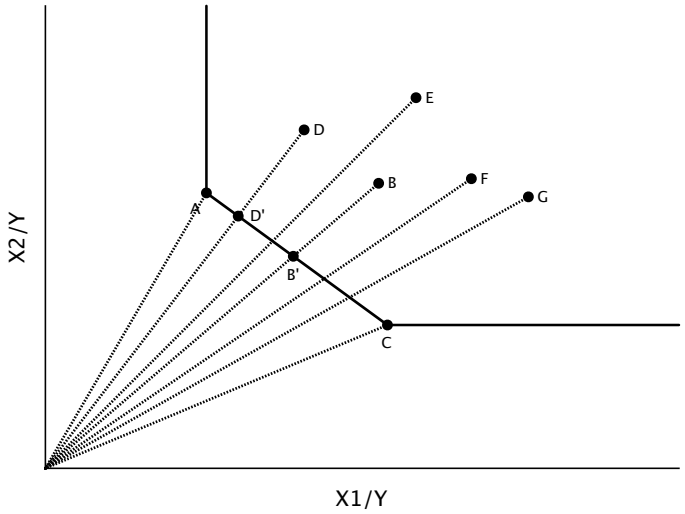

Fig. 6: Combinaisons des DMUs

Revenons maintenant aux variables d'écart, $s_r^+,\ r = 1, ..., s$ et $s_i^-,\ i = 1, ..., m$ en mentionnant qu'elles sont d'une très grande importance dans l'évaluation de l'efficacité des DMUs, dans la mesure où elles font partie intégrante de la définition de l'efficacité. En effet, il est possible qu'une DMU obtienne une mesure d'efficacité radiale égale à 1, mais qu'il soit aussi possible de diminuer la consommation de certains inputs de manière non proportionnelle. Ceci est le cas lorsque des DMUs se retrouvent sur l'un des segments parallèles aux axes de la frontière.

Par exemple, si nous reprenons l'ensemble de DMUs de la figure 5 et 6 et que nous y ajoutons une huitième DMU, le point H. La figure 7 illustre qu'il est impossible de trouver une combinaison de DMU qui utilise proportionnellement moins d'inputs que la DMU H, par conséquent celle-ci obtiendra $\theta^* = 1$. Toutefois, nous constatons que la DMU A utilise le même niveau d'input x_1 que la DMU D, mais avec un niveau d'input x_2 plus faible. La solution du problème dual évalué au point H devra alors contenir une variable d'écart positive en x_1. Il ne fait pas de sens de dire que le point H est efficace puisque le point A démontre qu'il est possible de produire une unité d'output avec moins d'inputs.

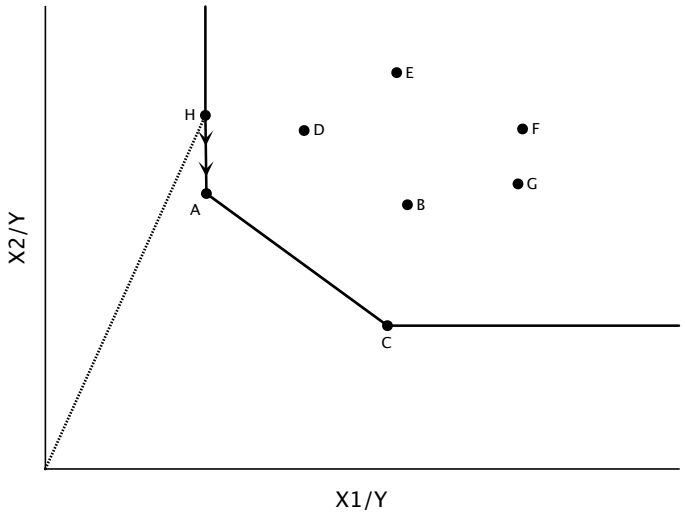

Fig. 7: Variable d'écart et réduction non proportionnelle

Cela dit, il est donc nécessaire de prendre en considération les variables d'écart lors de la résolution de (8) afin de mettre à jour l'ensemble des sources d'inefficacité possibles[5].

Pour ce faire, la méthode DEA suggère de procéder en deux phases:

1. $min\ \theta$ sous les contraintes (9), (10) et (11);
2. En utilisant $\theta^* = min\ \theta$, résoudre le problème suivant:

$$max_{\lambda,s^-,s^+} \quad \sum_{i=1}^{m} s_i^- + \sum_{r=1}^{s} s_r^+ \tag{12}$$

$$sujet\ à \quad s_i^- = \theta^* x_{io} - \sum_{j=1}^{n} x_{ij}\lambda_j \quad i = 1, ..., m \tag{13}$$

$$s_r^+ = \sum_{j=1}^{n} y_{rj}\lambda_j - y_{ro} \quad r = 1, ..., s \tag{14}$$

$$\lambda_j, s_i^-, s_r^+ \geq 0 \quad \forall i, j, r \tag{15}$$

[5] Notons que la résolution du problème en (8) ne procure par les solutions optimales pour les variables s_r^+ et s_i^-.

Si la solution optimale des deux problèmes précédents $(\theta^*, \lambda^*, s^{+*}, s^{-*})$ satisfait $\theta^* = 1$ **et** $s^{+*} = 0, s^{-*} = 0$, où $\lambda^*, s^{+*}, s^{-*}$ sont des vecteurs, alors la DMU_o est efficace au sens de la définition de Pareto-Koopmans donnée à la section 2.3.

C'est à partir des solutions obtenues pour chacune des DMUs de l'ensemble que nous pouvons construire la frontière d'efficacité. Mentionnons que nous avons présenté le modèle DEA comme une technique pour évaluer l'efficacité d'unités par rapport à une frontière d'unités efficaces. Cependant, cette frontière à laquelle nous faisons référence n'est pas explicitement formulée, elle est implicite aux scores d'efficacité θ^* qui sont calculés pour l'ensemble des DMUs. La frontière est définie comme l'ensemble des segments linéaires qui relient les DMUs pour lesquelles $\theta^* = 1$.

Enfin, le θ^* représente une mesure de contraction radiale, une valeur inférieure à 1 indique alors que tous les inputs consommés par la DMU_o peuvent être réduit proportionnellement de $(1-\theta^*)$ sur la base des évidences de l'ensemble de production défini par les données. Nous ferons référence à ce type d'inefficacité comme étant de l'*inefficacité radiale*. Une mesure de contraction radiale égale à 1 signifie que la DMU exhibe une des meilleures performances de l'ensemble et donc qu'elle se situe sur la frontière.

Par contre, le fait qu'une DMU se positionne sur la frontière n'est pas suffisant pour atteindre l'efficacité, car elle peut se situer sur une portion inefficace de celle-ci. Les variables d'écart, tel que souligné précédemment, représentent des réductions possibles propres à chacun des inputs ou des augmentations spécifiques à chacun des outputs, ce sont donc des déplacements non proportionnels. Nous référerons à ce type d'inefficacité comme étant de l'*inefficacité non proportionnelle*.

Nous pouvons aller un peu plus loin dans notre exposition du modèle CCR en notant que la formulation du problème pourrait se faire dans l'autre sens, c'est-à-dire en minimisant le ratio inverse de (1). Le programme linéaire deviendrait alors:

$$min_{v,u} \quad q_o = \sum_{i=1}^{m} v_i x_{io} \tag{16}$$

$$sujet\ à \quad \sum_{i=1}^{m} v_i x_{ij} - \sum_{r=1}^{s} u_r y_{rj} \geq 0 \quad j = 1, ..., n \tag{17}$$

$$\sum_{r=1}^{s} u_r y_{ro} = 1 \tag{18}$$

$$u_r, v_i \geq 0 \quad \forall\, i, r \tag{19}$$

et son dual:

$$max_{\eta,\mu} \quad \eta \tag{20}$$

$$sujet\ à \quad \sum_{j=1}^{n} x_{ij}\mu_j + s_i^- = x_{io} \quad i = 1, ..., m \tag{21}$$

$$\sum_{j=1}^{n} y_{rj}\mu_j - s_r^+ = \eta y_{ro} \quad r = 1, ..., s \tag{22}$$

$$\mu_j, s_i^-, s_r^+ \geq 0 \quad \forall\, i, j, r \tag{23}$$

Nous avons donc deux orientations possibles avec les modèles (8) et (20). Si celles-ci semblent faire référence à un seul et même problème, il faut néanmoins souligner que les problèmes posés diffèrent dans la mesure où (20) est orienté sur l'output puisqu'il tente de maximiser l'output en conservant la consommation d'inputs constante, tandis que (8) est orienté sur l'input en tentant de minimiser la consommation d'inputs tout en conservant le niveau d'output constant.

Il est important de distinguer ces deux formes puisque l'orientation choisie constituera une hypothèse importante dans certaines applications comme nous le verrons plus loin. Pour l'instant, nous conserverons la première orientation et son dual, tel que formulé en (4) et (8). De plus, nous travaillerons plutôt avec la forme dual pour certaines raisons que nous exposons ici (Cooper et al., 2007).

Premièrement, du côté computationel, le nombre de contraintes du problème dual est généralement inférieur au nombre de contraintes de la formulation primal ce qui simplifie la résolution des programmes. En effet, les contraintes (9) et(10) font intervenir les inputs et les outputs, tandis que (5) fait intervenir les DMUs. Le nombre de DMUs analysé sera, dans presque tous les cas, supérieur à la somme du nombre d'inputs et d'outputs. Deuxièmement, le problème primal ne permet pas de disposer des variables d'écart qui sont pourtant centrales dans l'explication et la détermination de l'efficacité des DMUs. Finalement, l'interprétation du problème dual est plus directe, puisque nous obtenons une mesure θ, qui s'interprète comme une mesure de contraction radiale de tous les inputs.

En plus de ce que nous venons de souligner, la solution au problème (8) permet de disposer pour chacune des DMUs inefficaces d'un ensemble de référence qui est constitué de DMUs efficaces lui indiquant quelles pratiques lui auraient permis d'atteindre l'efficacité. En effet, à partir des éléments de cet ensemble nous pouvons dériver les objectifs à atteindre en termes de consommation d'inputs et de production d'outputs pour permettre à ces DMUs inefficaces d'atteindre la frontière efficace. Les contraintes (9) et (10) font en sorte que les données de la DMU_o sont comparées à une combinaison de données de DMUs de la population analysée. Les $\lambda_j,\ j = 1, ...n$, obtenus lors de la phase 2 de la résolution du problème CCR, représentent les poids accordés à chacune des DMUs efficaces constituant l'ensemble de référence E_o qui est défini comme suit:

$$E_o = \left\{ j \mid \lambda_j^* > 0 \right\} \quad (j \in \{1, ..., n\}) \tag{24}$$

Il va sans dire que les DMUs efficaces seront les seuls éléments de leur propre ensemble de référence puisqu'il n'y aura pas d'autres DMUs qui pourront reproduire à un niveau inférieur leur consommation de ressources et leur production d'outputs.

Reprenons (9) et (10), à partir desquelles nous pouvons dériver ces objectifs dont nous faisons mention:

$$\hat{x}_{io} = \sum_{j \epsilon E_o} x_{ij} \lambda_j^* = \theta^* x_{io} - s_i^{-*} \quad i = 1, ..., m \tag{25}$$

$$\hat{y}_{ro} = \sum_{j \epsilon E_o} y_{rj} \lambda_j^* = s_r^{+*} + y_{ro} \quad r = 1, ..., s \tag{26}$$

Les relations (25) et (26) suggèrent, afin de ramener une DMU inefficace à l'efficacité, de diminuer les inputs par le ratio θ^* et d'en soustraire les variables d'écart s'il y a lieu, tout en augmentant les outputs des variables d'écarts ayant une valeur positive. Notons qu'en présence d'optimums multiples, plusieurs formules de projection existeront correspondant chacune à un ensemble de référence différent.

Prenons la figure 8 pour illustrer ces propos. Le point A est inefficace, sa projection sur la frontière est alors obtenue comme une combinaison des points E et D. Le point B, quant à lui, est projeté sur la frontière efficace à partir d'une combinaison des points D et C. La réduction proportionnelle de la consommation d'inputs pour les points A et B, soit le déplacement sur leur rayon, est suffisante pour les ramener à l'efficacité, il n'y aura pas de variables d'écarts positives dans leur formule de projection. Si nous regardons maintenant le point G, son ensemble de référence ne contient que le point E. Par contre G et E n'utilisent pas les deux inputs dans la même proportion, ils ne sont pas situés sur le même rayon. Ceci signifie qu'atteindre G' n'est pas suffisant pour atteindre l'efficacité, il y a une variable d'écart qui est positive en x_2. À partir de G' il est nécessaire de diminuer la consommation en x_2 afin d'atteindre le point E et du même coup, la projection efficace.

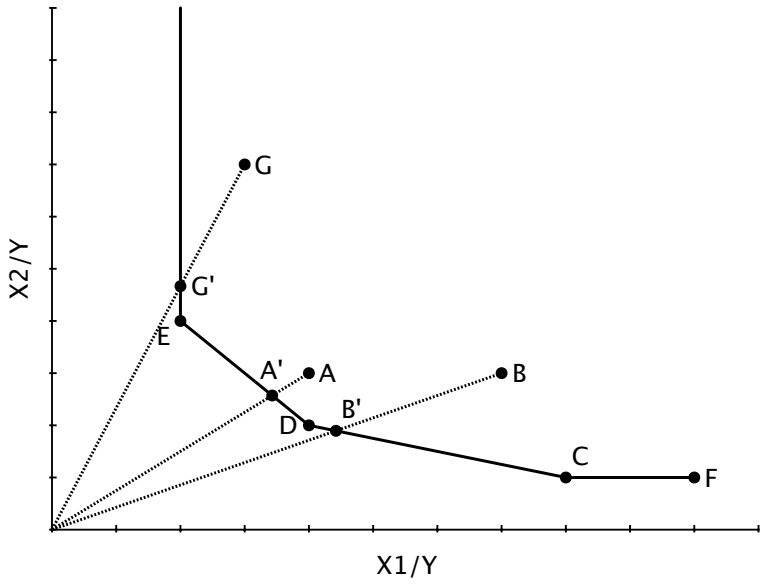

Fig. 8: Deux inputs I

Nous concluons maintenant cette section en présentant une série de théorèmes et leurs corollaires concernant la méthode DEA dont les preuves sont fournies par Cooper et al. (2007, pp.48-50):

1. La projection $(\hat{x}_o, \hat{y}_o)$ telle que définie en (25) et (26) est efficace au sens de la définition de Pareto-Koopmans;

2. La projection $(\hat{x}_o, \hat{y}_o)$ est située sur la frontière des DMUs efficaces;

3. Les DMUs composant tout ensemble de référence sont efficaces;

4. Toutes combinaisons semipositives de DMUs contenues dans un ensemble de référence sont efficaces.

3.3 Hypothèses de la méthode DEA

Pour faire suite à l'exposition technique de la méthode DEA, nous proposons dans cette section une discussion sur les hypothèses posées et sur leurs implications dans le cadre d'une analyse d'efficacité.

À ce propos, la première des choses qu'il convient de mettre en évidence est le caractère relatif de l'efficacité que propose d'évaluer la méthode DEA. Les DMUs sont évaluées les unes par rapport aux autres et ce sont les organisations qui utilisent le moins de ressources pour une production donnée qui sont caractérisées comme étant efficaces. En ce sens, la méthode DEA offre une évaluation des meilleures pratiques, plutôt qu'une évaluation de l'efficacité absolue des organisations. Néanmoins, comme nous ignorons dans la grande majorité des cas, où se situe la véritable frontière de production, son estimation sur la base d'observations ne peut produire qu'une évaluation conservatrice de celle-ci. Comme nous l'avons noté précédemment, cette frontière est le *standard d'efficacité minimum*. Les organisations étant caractérisées comme inefficaces par la méthode DEA conserveront vraisemblablement leur statut au regard de la réelle frontière de production, et ce, même si nous ne disposons d'aucune évaluation de la performance absolue des DMUs caractérisées efficaces relativement aux autres.

D'un autre côté, cette façon de faire pose comme hypothèse que le processus de transformation de chacune des DMUs est équivalent, puisqu'aucune mesure de qualité ou de réalisation n'est incorporée au modèle. Dans une analyse d'efficacité, il semble important de prendre en considération cet aspect. Cependant, si les développements récents de la technique DEA permettent maintenant de prendre en compte l'aspect qualitatif de la production, il demeure qu'une très grande proportion des applications éludent tout simplement cet aspect, nous aborderons cette question à la section 4.4.

Ensuite, la méthode DEA est une méthode non paramétrique et déterministe. Non paramétrique, dans le sens où il n'est pas nécessaire de définir une forme fonctionnelle pour la frontière de production et déterministe puisqu'aucun aléa n'est inclus dans la spécification du modèle. Ce dernier aspect a plusieurs implications dont nous devons rendre compte. En effet, en ne supposant aucun bruit, la méthode DEA présume l'absence d'erreurs de mesure sur les données. Ceci peut représenter une source de biais importante dans une technique qui utilise l'information des données extrêmes (Jacobs et al., 2006). Par exemple, nous pouvons constater dans la figure 9 l'ampleur des conséquences pour l'ensemble des possibilités de production (une nouvelle frontière en pointillé) et pour l'évaluation de l'efficacité d'un bon nombre de DMUs (celles dans la zone ombragée) si nous supposons que les données d'une seule DMU comportent certaines erreurs de mesures, le véritable point étant Z' plutôt que Z (Badillo et al., 1999).

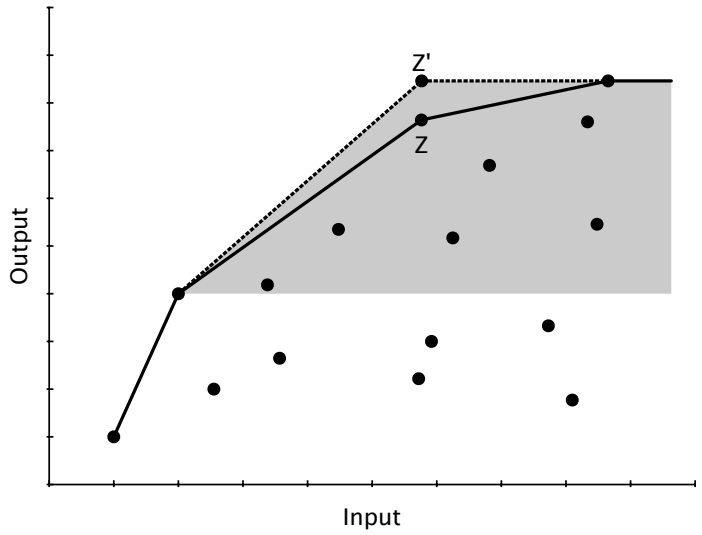

Fig. 9: Effet d'une erreur de mesure

De plus, l'absence d'erreur systématique sous-entend une spécification adéquate du modèle dans le sens où tous les inputs et les outputs pertinents ont été inclus. Cette caractéristique de la méthode entraîne comme conséquence que toutes les déviations de la frontière sont attribuables à l'inefficacité supposée inhérente aux processus des organisations (Jacobs et al., 2006).

En spécifiant le modèle sur la base des inputs et des outputs uniquement, la méthode DEA suppose que l'hétérogénéité des organisations pourra être prise en compte en accordant un vecteur de poids spécifique à chacune dans la résolution du problème primal. Cette flexibilité quant à la libre détermination des poids pour chacune des DMUs est précisément destinée à accommoder les différences qui pourraient exister entre les organisations (Ramanathan, 2003). Toujours est-il que ceci repose sur la bonne spécification du modèle, ce qui est embêtant dans la mesure où le caractère déterministe de la technique limite les tests que nous pouvons effectuer pour déterminer l'inclusion et l'exclusion de variables dans les modèles DEA.

Sur des considérations plus techniques, la méthode DEA procède à l'évaluation de l'efficacité des DMUs en comparant chacune d'entre elles à une combinaison d'observations utilisant une composition similaire d'inputs et d'outputs. Cette combinaison est construite à partir des poids λ optimaux déterminés dans la résolution de la deuxième phase du problème dual. Dans le modèle CCR, ces λ sont restreints par une contrainte de semipositivité ce qui implique que les interpolations et les extrapolations linéaires d'observations efficaces sont possibles selon la définition de l'ensemble de production. De plus, en fonction de la propriété d'expansion radiale que nous avons citée comme troisième propriété de l'ensemble de production, il est tout aussi possible d'inférer une diminution ou une augmentation de taille à partir de n'importe quel point appartenant à l'ensemble; ce nouveau point appartiendra également à l'ensemble de production. Ceci signifie alors que le modèle CCR pose l'hypothèse de l'existence de rendements d'échelle constants.

Géométriquement, l'ensemble de production T défini par la méthode CCR est un cône convexe, car pour deux vecteurs $(X_j, Y_j), (X'_j, Y'_j) \in T$, et pour des scalaires $\alpha \geq 0, \beta \geq 0$, nous avons, $\alpha(X_j, Y_j) + \beta(X'_j, Y'_j) \in T$. Ceci découle des propriétés de convexité et de rendements constants de l'ensemble (Mas-Colell et al., 1995, p.134).

À propos des résultats de la méthode DEA, nous poursuivons en relevant la double utilité du score d'efficacité θ estimé pour chacune des DMUs. D'une part, il permet de distinguer les organisations se situant sur la frontière ($\theta^* = 1$), des organisations qui n'y sont pas ($\theta^* < 1$). D'autre part, il permet de définir une projection $(\hat{x}_o, \hat{y}_o)$ qui ramènerait les organisations inefficaces sur la portion efficace de la frontière. Cette projection se retrouve bien sur la frontière efficace puisqu'elle est définie comme une combinaison linéaire d'organisations qui se retrouvent elles-mêmes sur cette portion de la frontière comme le montre (25) et (26). En supposant ce déplacement possible et en le rendant nécessaire pour retrouver l'efficacité, la méthode admet que tous les points entre deux organisations se situant sur la frontière sont réalisables. Ceci correspond à concevoir une parfaite substitution entre les façons de combiner les divers inputs.

La figure 10 montre pour certains points inefficaces, les projections qui les ramèneraient sur la frontière. Ces projections sont obtenues à partir de combinaisons de points efficaces. Toutefois, il est possible qu'aucune réalisation ne soit observée entre deux technologies frontières. Par exemple, le point A' est dérivé des points G et H. En définissant ce point A', nous supposons qu'il soit possible d'utiliser une technologie mitoyenne entre G et H, et ce, même si cela n'a jamais été observé. Une telle hypothèse est forte en implications, car elle entraîne la définition de points efficaces sur la frontière qui n'ont pas été observés. Soulignons qu'ultimement cela résulte de la façon dont l'ensemble de production a été défini, c'est-à-dire en supposant sa convexité.

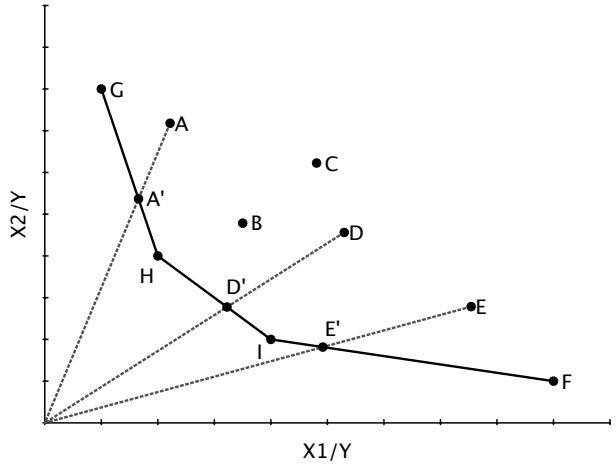

Fig. 10: Deux inputs II

La nécessité du déplacement indiqué par les projections suppose également que les gestionnaires ont le plein contrôle sur les différentes variables dans le cadre du processus de production, c'est-à-dire qu'ils ont la possibilité de diminuer leur consommation en ressources ou d'augmenter la production d'outputs (un choix dicté par l'orientation du modèle). Il faut préciser que cette présomption va plus loin que la simple propriété de libre disposition donnée à l'ensemble de production. Si cette dernière entend qu'il est possible de réduire la consommation en ressources tout en conservant le niveau d'output et que ce nouveau point appartient à l'ensemble de production, invoquer le contrôle des gestionnaires fait plutôt intervenir la notion de déplacement. L'un réfère à l'existence d'un point, tandis que l'autre réfère à la possibilité de s'y rendre, deux concepts bien différents. Il en découle alors que l'inefficacité observée au sein des organisations est totalement attribuable aux pratiques de gestion, donc éliminables, plutôt que de rigidités structurelles propres au secteur des organisations. Cette supposition est certainement l'une des plus fortes de la méthode DEA, dans le sens où elle risque fort bien de ne plus être respectée à l'extérieur du cadre théorique.

Certes, l'attrait d'une méthode comme la méthode DEA et son modèle CCR réside certainement dans la grande flexibilité offerte par la technique. Le positionnement de la frontière de production sur la base de l'évaluation relative des organisations ne nécessite aucune spécification du lien input-output.

Toutefois, le positionnement de la frontière reste sensible aux valeurs extrêmes et aux observations présentant une combinaison ou un niveau singulier d'inputs et d'outputs (Jacobs et al., 2006). En effet, Cooper, Seiford et Tone (2007) montrent bien que si une DMU détient le niveau le plus bas d'inputs, elle sera caractérisée comme efficace au même titre que la DMU qui présente le niveau d'outputs le plus élevé.

En bref, il va sans dire que la multitude d'hypothèses que nous avons mises en évidence peuvent être restrictives, quoique dans des degrés qui varient, surtout en regard des différents secteurs auxquels la technique DEA peut être utilisée. Dans les sections qui suivent, nous accorderons une attention particulière à ces hypothèses en présentant les modifications au modèle CCR qui ont été proposées.

4 Extensions théoriques et pertinence pour le secteur de la santé

Nous nous attardons maintenant à la présentation des multiples variantes et extensions du modèle CCR qui ont été développées au fil des ans. Cette présentation reste pertinente dans la mesure où elle permettra de mener une discussion sur la façon dont ces concepts influencent une analyse d'efficacité spécifique dans le cadre d'un système de soins de santé. Ainsi, nous aborderons les modèles à rendements d'échelle variables, la façon d'inclure des composantes stochastiques et des variables non discrétionnaires, les modèles qui tiennent compte de la notion de qualité, des modèles à ensemble non convexes et pour finir les modèles d'analyse intertemporelle de l'efficacité.

4.1 Modèle à rendements d'échelle variables

L'un des premiers constats qui est posé sur le modèle CCR est la restriction sur les rendements d'échelle qu'il autorise. La restriction aux rendements d'échelle constants semble être une limitation importante dans la plupart des applications pratiques. Afin d'élargir les possibilités de recherche associées au modèle DEA, certains auteurs ont adapté le modèle DEA standard pour tenir compte de rendements d'échelle variables. Nous offrons maintenant une présentation générale de ces modifications.

4.1.1 Théorie

Le modèle le plus utilisé dans ce contexte est le modèle BCC développé par Banker, Charnes et Cooper et publié dans un article de 1984. Essentiellement, le modèle reprend la formulation du modèle CCR en ajoutant une contrainte de convexité aux multiplicateurs λ.

Le programme linéaire BCC est le suivant:

$$min_{\lambda,\theta} \quad \theta \tag{27}$$

$$sujet\ à \quad \theta X_o - X\lambda \geq 0 \tag{28}$$

$$Y\lambda \geq Y_o \tag{29}$$

$$e\lambda = 1 \tag{30}$$

$$\lambda \geq 0 \tag{31}$$

où X est le vecteur d'inputs, Y le vecteur d'outputs, e un vecteur unitaire, λ un vecteur de poids et θ un scalaire. Nous pouvons écrire la forme dual de ce problème de minimisation:

$$max_{v,u,u_o} \quad z = uY_o - u_o \tag{32}$$

$$sujet\ à \quad vX_o = 1 \tag{33}$$

$$-vX + uY - u_o e \leq 0 \tag{34}$$

$$v \geq 0, u \geq 0, u_o\ libre \tag{35}$$

où v et u sont des vecteurs, z et u_o sont des scalaires.

La différence entre les modèles (4) et (27) consiste donc en la présence du scalaire u_o qui est associé à la contrainte (30) dans le modèle BCC. L'ajout de cette contrainte a pour conséquence de limiter les combinaisons d'observations possibles à des combinaisons linéaires convexes lors de la comparaison la performance de la DMU_o. L'évaluation de la DMU_o est donc faite par rapport à une autre DMU ou une combinaison de DMUs de «même taille», puisque nous limitons les extrapolations et les interpolations dans la définition de l'ensemble de production.

La résolution du problème BCC en (27) passe par la résolution des mêmes deux phases que nous avons présentées pour le modèle CCR.

Une DMU est caractérisée efficace de la même façon que dans le modèle CCR, c'est-à-dire si son score d'efficacité $\theta^{*BCC} = 1$ et que les variables d'écart $s_i^{-*} = 0 \forall i$ et $s_r^{+*} = 0\ \forall r$.

Notons, par ailleurs, que le score d'efficacité θ^{*BCC} ne peut pas être inférieur au score d'efficacité θ^{*CCR}, ceci étant dû à l'ajout de la contrainte sur les λ qui limite

l'ensemble de production et donc réduit les mesures de distances par rapport à la frontière. La figure 11 met en évidence cette relation.

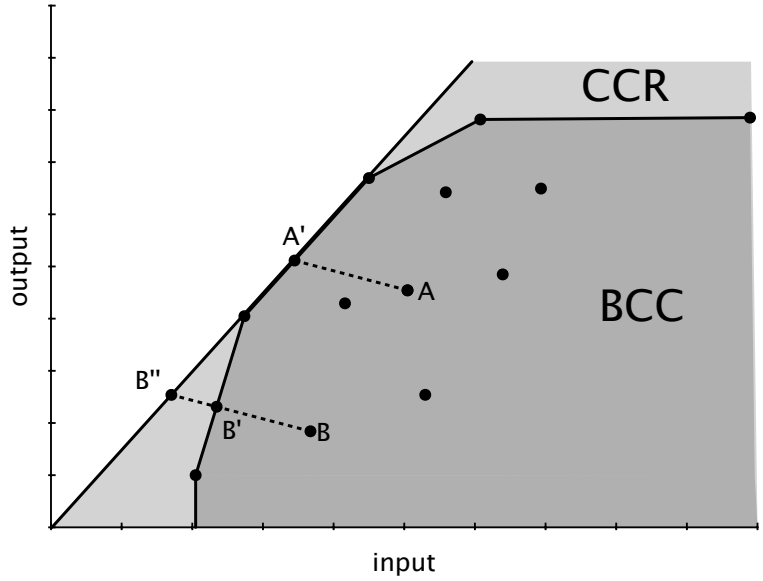

Fig. 11: Ensembles de production BCC et CCR

De plus, comme dans le modèle CCR, le modèle BCC définit les projections des points inefficaces sur la frontière:

$$\hat{x}_{io} = \theta^{*BCC} x_{io} - s_i^{-*},\ i = 1, ..., m$$
$$\hat{y}_{ro} = y_{ro} + s_r^{+*},\ r = 1, ..., s$$

Maintenant, une nouvelle composante du modèle BCC consiste en sa capacité à caractériser le niveau des rendements d'échelle pour les DMUs situées sur la frontière. Le problème tel que formulé en (32) permet de disposer de cet indicateur. Nous énonçons la façon de faire sous la forme du théorème de la page suivante.

Si (X_o, Y_o) est un point sur la frontière BCC, alors le signe de la variable u_o permet d'identifier les rendements d'échelle à ce point. Si $u_o^* < 0$ pour toute solution optimale, les rendements d'échelle sont croissants à (X_o, Y_o); si $u_o^* > 0$ pour toute solution optimale, les rendements d'échelle sont décroissants à (X_o, Y_o); si $u_o^* = 0$ pour n'importe quelle solution optimale, les rendements d'échelle sont constants à (X_o, Y_o) [6] .

La preuve de ce théorème en fonction d'une analyse graphique apparaît assez simple. Cooper et al. (2007, p.135) montre que u_o peut définir l'intercept d'une droite, ou encore le niveau d'un hyperplan supportant un point (X_o, Y_o). La figure 12 illustre l'utilisation que nous pouvons faire de ce u_o. La frontière est définie par le segment $\overline{ABCD}$. L'intercept de la droite supportant le segment $\overline{AB}$ est u_1, celui-ci étant négatif montre que les rendements d'échelle sont croissants sur ce segment $\overline{AB}$. Pour le segment $\overline{BC}$, l'intercept u_2 est nul, il identifie alors des rendements constants pour les points situés sur ce segment. Finalement, le segment $\overline{CD}$ est supporté par une droite ayant un intercept positif u_3, les points sur ce segment sont donc caractérisés par des rendements d'échelle décroissants.

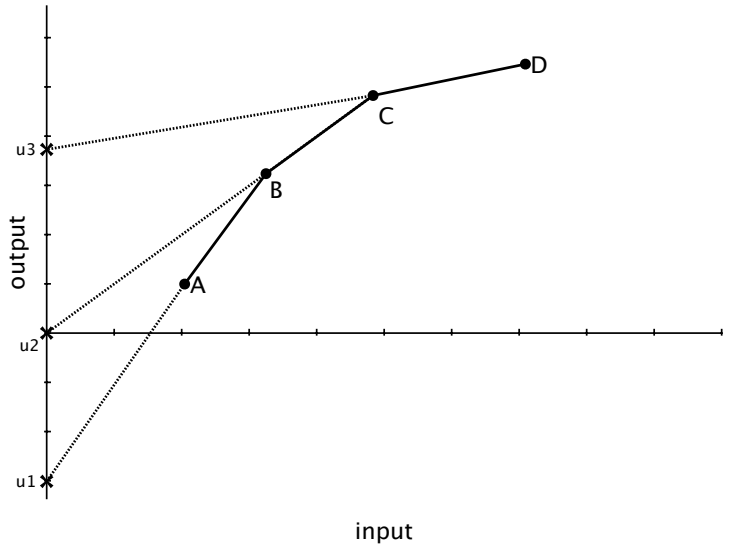

Fig. 12: Utilisation de u_o pour identifier les rendements d'échelle

[6] Ce théorême précise bien que nous devons regarder l'ensemble des solutions optimales pour les rendements d'échelle croissants et décroissants. Ceci est nécessaire pour tenir compte de points comme le point B de la figure 12, pour lesquels plusieurs hyperplans supports peuvent exister. Cooper et al. (2007, p.137) proposent une seconde étape qui consiste à maximiser ou à minimiser u_o afin de déterminer s'il est possible d'obtenir $u_o^* = 0$ et donc de conclure que le point en question serait plutôt caractérisé par des rendements d'échelle constants.

Autrement, certains auteurs ont proposé d'utiliser le modèle CCR pour caractériser la nature des rendements d'échelle (Banker et Thrall, 1992). Il s'agit d'évaluer les DMUs en regard de la frontière CCR en supposant celle-ci valide et de regarder la somme des λ^* obtenus lors de la seconde phase de résolution comme indicateur des rendements d'échelle. Les λ^* indiquent la performance d'une combinaison de DMUs efficaces que la DMU_o serait vraisemblablement capable d'atteindre si elle était efficace. De cette façon, la somme de ces multiplicateurs à l'optimum nous permet de connaître l'amplitude du déplacement qui devrait être fait par rapport aux unités efficaces sous CCR, à savoir si nous devons extrapoler à partir des points efficaces, donc augmenter la taille de la production par rapport aux unités efficaces, ou encore interpoler, ce qui signifie diminuer la taille de la production en regard des DMUs efficaces sous CCR. Nous présentons le théorème pour ensuite compléter l'explication par un exemple.

> *Si (X_o, Y_o) est un point sur la frontière efficace BCC, alors $\sum_{j \in E_o^{CCR}} \lambda_j^*$, où E_o^{CCR} est l'ensemble de référence CCR de la DMU_o, permet d'identifier les rendements d'échelle à ce point. Si $\sum \lambda_j^* < 1$ pour toute solution optimale, les rendements d'échelle sont croissants à (X_o, Y_o); si $\sum \lambda_j^* > 1$ pour toute solution optimale, les rendements d'échelle sont décroissants à (X_o, Y_o); si $\sum \lambda_{j=1}^*$ pour n'importe quelle solution optimale, les rendements d'échelle sont constants à (X_o, Y_o).*

La figure 13 illustre la façon de faire. Regardons le point A, efficace sous BCC et inefficace sous CCR. En évaluant A sous CCR nous obtenons sa projection A' sur la frontière CCR. Celle-ci est obtenue comme une combinaison de DMUs de son ensemble de référence qui ne contient ici que le point F ou que le point G, car deux optimums sont possibles. Nous obtenons donc $A' = \lambda_f^* F = \lambda_g^* G \, , 0 < \lambda_j^* < 1 \; j = \{F, G\}$. La valeur des λ_j^* est donc une indication sur le déplacement à faire. Il serait nécessaire de diminuer la taille de production, mais comme cela est impossible au point A en fonction de l'ensemble de production BCC, nous pouvons conclure que des rendements d'échelle croissants prévalent en A.

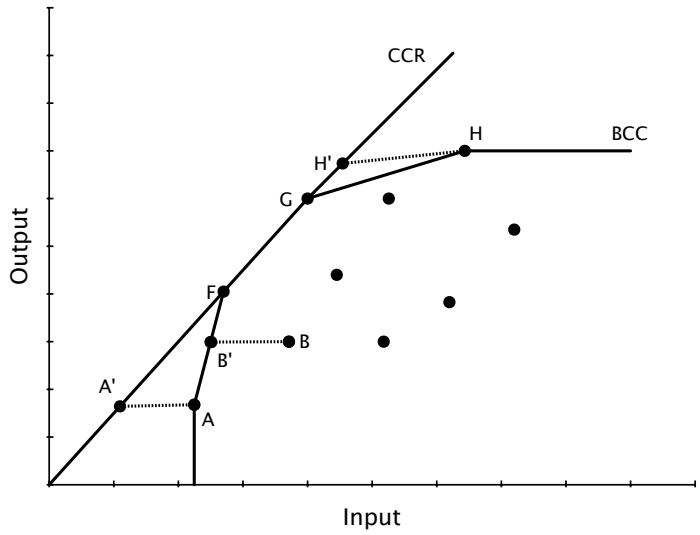

Fig. 13: Projections CCR pour identifier les rendements d'échelle

Ensuite, nous pouvons procéder de la même manière pour caractériser la nature des rendements d'échelle au point H. La projection H' est obtenue à partir de F ou G. Dans les deux cas, la valeur du multiplicateur λ à l'optimum doit être supérieure à 1. Ainsi, il serait nécessaire d'augmenter la taille de production. L'impossibilité de le faire selon l'ensemble défini par BCC nous permet de conclure qu'au point H les rendements d'échelle sont décroissants.

Il faut insister sur le fait que cette méthode pour déterminer les rendements d'échelle n'est valable que pour des DMUs situés sur la frontière BCC. En effet, dans le cas de DMUs inefficaces sous BCC, l'inefficacité est confondue avec les rendements d'échelle (Banker et Thrall, 1992).

Diverses méthodes ont été développées pour tenir compte de cette limite du théorème énoncé plus haut (Banker et al., 1996; Tone, 1996; Butler et Li, 2005). Nous nous contentons ici de citer l'idée générale de ces méthodes. Celles-ci consistent principalement à effectuer l'analyse des rendements d'échelle pour un point (X_o, Y_o) inefficace à partir de sa projection $(\hat{X}_o, \hat{Y}_o)$ qui se situe sur la frontière. Par exemple, à la figure 13, cela consisterait à évaluer les rendements d'échelle de la DMU_B à partir du point B' sur la frontière BCC.

En général, nous pouvons dire que le modèle BCC permet de mener deux analyses. D'un côté, il permet d'évaluer l'efficacité technique sous sa forme en (27) et d'un

autre, il permet de caractériser la situation des rendements d'échelle à partir de modèle comme (32).

En mettant en relation le modèle CCR et le modèle BCC, nous remarquons qu'une DMU est caractérisée efficace sous les deux mesures d'efficacité si et seulement si elle se situe dans une région de rendements d'échelle constants. Bien que cela découle de l'hypothèse de rendements d'échelle constants du modèle CCR, nous devons préciser qu'une hypothèse est sous-jacente à ce résultat. Cette hypothèse provient de la définition de l'efficacité qui est adoptée par la méthode DEA. En effet, dans la perspective DEA, l'atteinte de rendements constants est considérée comme une situation désirable que l'on assimile à l'atteinte du maximum de la productivité moyenne (Cooper et al, 2007, p.142).

À ce titre, certains modèles, appelés les modèles du *most productive scale size*, permettent d'associer l'atteinte de l'efficacité des organisations et l'atteinte de la région de rendements constants. Ces modèles proposent même une méthode pour dériver des projections qui ramènent les organisations vers cette région (Banker, 1984). Ce genre de modèle considère alors l'échelle de production comme une décision de gestion, en évacuant le fait que certains motifs comme les contraintes technologiques puissent être à la source de rendements variables.

Poursuivons encore la comparaison entre le modèle CCR et le modèle BCC. Nous avons formulé les problèmes en fonction d'une orientation sur l'input, nous aurions toutefois pu les formuler en fonction d'une orientation sur l'output. Pour le modèle CCR, cela ne modifie en rien les résultats qui sont obtenus puisque la mesure θ^{CCR} est invariable à l'orientation du modèle. Cependant pour le modèle BCC, la conséquence est tout autre, car généralement les mesures d'efficacité vont différer selon l'orientation choisie, puisque les projections sur la frontière seront différentes selon l'orientation. À cet égard, comme la caractérisation des rendements d'échelle se fait à partir des projections pour les points inefficaces, il est possible d'obtenir une situation avec des rendements d'échelles différents en fonction de l'orientation.

Dans un autre ordre d'idées, le modèle BCC et ses dérivés rendent possible une décomposition de l'efficacité en fonction de l'efficacité technique «pure» et de l'efficacité d'échelle. Comme nous l'avons montré avec la figure 13, il est possible d'obtenir une observation qui soit efficace en fonction de la frontière BCC, mais que le score d'efficacité sous le modèle CCR soit en deçà de 1. Le relâchement de la contrainte de convexité contribue donc à rendre la DMU inefficace. Le ratio

donné par les deux scores peut alors servir de mesure d'efficacité d'échelle (SE).

$$SE = \frac{\theta^{*CCR}}{\theta^{*BCC}} \leq 1 \tag{36}$$

Remarquons que l'efficacité d'échelle est atteinte lorsque les scores CCR et BCC sont équivalents, ce qui revient donc à dire que la DMU se situe dans la région des rendements constants de la frontière. Le score d'efficacité CCR caractérise l'efficacité générale (TE), alors que le score d'efficacité BCC constitue celui de l'efficacité technique pure (PTE). Nous pouvons donc réécrire (36) comme:

$$\begin{aligned} \theta^{*CCR} &= \theta^{*BCC} \times SE \\ TE &= PTE \times SE \end{aligned} \tag{37}$$

Nous constatons maintenant que la modification du modèle de base pour prendre en compte des rendements d'échelle variables respecte l'esprit général de la mesure DEA obtenue par CCR, dans le sens où les hypothèses restent sensiblement les mêmes. L'analyse d'efficacité repose toujours sur l'obtention d'une mesure de contraction radiale définissant une projection sur la frontière et les mêmes constats quant aux raisons de la déviation de la frontière et sur la possibilité pour les gestionnaires d'ajuster leurs niveaux de variables, continuent de s'appliquer.

La principale modification par rapport aux hypothèses de base du modèle CCR réside dans la définition de l'ensemble des possibilités de production. Le modèle BCC postule un ensemble de production qui soit intérieur à l'ensemble du modèle CCR, ceci étant attribuable à l'interdiction d'extrapoler l'ensemble de production au-delà des performances observées des DMUs. Le cône convexe qui définissait l'ensemble CCR est maintenant remplacé par un ensemble convexe à segments linéaires. Conséquemment, la propriété d'additivité de l'ensemble n'est plus respectée et ce faisant, le modèle BCC impose une contrainte sur la taille des DMUs dans l'ensemble de production. D'ailleurs, c'est pour cette raison que nous mentionnions que le modèle BCC comparait les unités à partir d'unités de même taille.

4.1.2 Pertinence pour une analyse en santé

Nous avons présenté les modèles CCR et BCC qui sont considérés comme les modèles principaux de la méthode DEA, il importe désormais de s'attarder à la pertinence de ces derniers lorsqu'il est question d'analyser l'efficacité de composantes d'un système de soins de santé.

Débutons en soulignant que le recours à ces modèles découle avant toute chose de la nature de la production dans un système de santé. Nous l'avons déjà noté, il est difficile de faire cadrer cette production selon les schémas d'analyse économique traditionnels, en partie parce qu'il est difficile de cibler des niveaux absolus d'efficacité. À cet égard, l'évaluation relative que propose la technique DEA est toute indiquée pour une analyse en santé. La possibilité de cibler les meilleures pratiques semble intéressante dans la mesure où les organisations inefficaces peuvent orienter leurs propres pratiques vers celles-ci.

De façon plus précise, discutons de la pertinence du modèle BCC en comparaison du modèle CCR dans une analyse d'efficacité de certaines composantes d'un système de santé. D'abord, nous l'avons mentionné, l'imposition de rendements d'échelle constants par le modèle CCR est possiblement très réductrice. Comme dans n'importe quel secteur de l'économie, il est possible que la production des établissements de soins de santé se situe à un niveau de rendements d'échelle variables. Supposer *a priori* que les établissements puissent prendre n'importe quelle taille, au sens de la définition de l'ensemble des possibilités de production, élude donc cette possibilité.

Ensuite, et de façon plus importante, c'est que la notion de taille revêt un caractère précis pour les organisations en santé au Québec, mais aussi dans toutes les régions où le système de santé est sous l'égide d'un contrôle gouvernemental serré. Au Québec, le ministère de la Santé et des Services sociaux (MSSS) est le principal acteur et régulateur des services en matière de santé. Il décide du type, du lieu et de la taille des établissements qui pourront opérer, ceci en fonction des besoins régionaux, mais également en fonction de considérations politiques (Québec. Loi sur les services de santé et les services sociaux, L.R.Q. ch.S-4.2, art. 463).

De nombreuses réformes au sein du système de santé québécois ont façonné l'organisation du réseau actuel. Au début des années 1990, plus d'un millier d'établissements constituaient le réseau de la santé et des services sociaux, en 2001,

nous n'en comptions plus que 482 (Turgeon et al., 2003), alors qu'aujourd'hui en 2009, le MSSS reporte sur son site internet que 293 établissements composent le système[7]. La diminution du nombre d'établissements s'explique par la fusion de plusieurs d'entre eux et s'accompagne ainsi de la prise en charge de plusieurs missions[8] pour un bon nombre d'établissements (Turgeon et al., 2003).

De cette manière, la taille d'opération des établissements et les services qu'ils peuvent offrir deviennent hors du contrôle des organisations et de leurs gestionnaires. Le fait que le modèle BCC permette de séparer l'apport de l'efficacité d'échelle (SE) et de l'efficacité technique pure (PTE) à l'efficacité totale est alors plus qu'utile. Nous pouvons ainsi disposer d'une mesure d'efficacité (PTE) qui soit indépendante de la taille de l'établissement et qui soit simplement liée aux pratiques de gestion et d'organisation des établissements.

La possibilité de caractériser les rendements d'échelle dans le cadre des modèles BCC et autrement par CCR, trouve sa pertinence selon le point de vue qui est adopté dans les analyses de performance. Du point de vue des gestionnaires, pouvoir disposer d'une mesure locale de rendement d'échelle est plus ou moins à propos étant donné leur contrôle limité sur la taille de leur organisation. Toutefois, du point de vue ministériel, il pourrait être plus intéressant de détenir ce genre d'information afin de l'intégrer à l'élaboration des politiques, voire à l'orientation des réformes, pour se questionner par exemple, sur la pertinence de continuer à fusionner les établissements pour créer des superétablissements de soins. Cependant, un tel usage des résultats DEA doit encore être soumis à certains tests de validité comme nous en discuterons à la section 5.4 et 6.2.

[7] Québec. Ministère de la Santé et des Services sociaux. En ligne. http://wpp01.msss.gouv.qc.ca/appl/M02/M02ListeEtab.asp?Etab=Region. (page consultée le 31 juillet 2009).

[8] Les établisements peuvent prendre en charge l'une ou les missions suivantes: centre hospitalier (CH), centre local de services communautaires (CLSC), centre d'hébergement et de soins longue durée (CHSLD), centre de protection de l'enfance et de la jeunesse (CPEJ) et centre de réadaptation (CR). (Québec. Loi sur les services de santé et les services sociaux, L.R.Q. ch.S-4.2, art. 79.)

4.2 Modèles stochastiques

Nous poursuivons notre revue de la méthode DEA en abordant les limites imposées par son contexte déterministe. Comme nous l'avons mentionné, la méthode DEA suppose que les différentes variables (inputs, outputs) sont mesurées sans erreur et par conséquent que la déviation de la frontière d'efficacité est le résultat des mauvaises pratiques de gestion de certaines DMUs. Toute distance par rapport à la frontière est assimilée à l'inefficacité.

Cependant, il semble que la performance des organisations puisse être affectée par trois facteurs: l'efficacité dans le processus de transformation d'inputs en outputs, l'environnement dans lequel les organisations opèrent et finalement les chocs aléatoires comme la chance, l'omission de variables et les problèmes associés (Fried et al., 2002). Tandis que le premier facteur est endogène aux organisations, les deux autres y sont complètement exogènes.

Idéalement, nous souhaiterions qu'une mesure d'efficacité puisse distinguer ces trois types de facteurs et leur impact sur la performance des organisations. Pourtant, ce n'est pas le cas de la méthode DEA. La notion de choc aléatoire est évidemment omise de par le caractère déterministe de la démarche DEA, alors que le contexte environnemental est difficilement pris en compte, et ce, même si des modèles incluant des variables non discrétionnaires et des catégories existent comme nous l'exposerons dans une prochaine section. C'est que l'une des hypothèses importantes de la méthode DEA est que les variables incluses dans l'analyse sont «isotoniques» dans le sens où plus d'outputs et moins d'inputs sont toujours désirables au niveau de l'efficacité. En ce qui concerne les variables environnementales, le sens dans lequel elles peuvent influencer la performance n'est pas toujours connu, ce qui contribue à complexifier leur inclusion dans les modèles généraux que nous avons présentés jusqu'à ce point.

4.2.1 Théorie

Différentes méthodes ont été proposées afin de permettre à la mesure d'efficacité DEA de considérer ces aspects. Rapportons d'ailleurs que ce champ est parmi le plus fertile de la recherche en matière de modèles DEA. Pour cette raison, nous nous contenterons d'exposer les grands principes et l'intuition qui est sous-jacente aux développements de ces multiples méthodes.

Le premier groupe de méthodes qui composent avec les notions d'environnement et de choc aléatoire est construit à partir de méthodes à étapes multiples. Dans un premier temps, il s'agit d'obtenir les scores d'efficacité de façon habituelle et ensuite, lors d'étapes subséquentes, d'intégrer des éléments stochastiques à l'aide d'une autre méthode afin de séparer l'efficacité technique des effets des variables contextuelles et aléatoires. Nous retrouvons des méthodes simples à deux étapes et des méthodes plus complexes à trois étapes ou plus.

Généralement les méthodes en deux étapes se concentrent sur l'influence qu'exerce l'environnement sur les scores θ^* d'efficacité radiale. À l'aide d'une régression où la distribution de la variable dépendante est censurée (la plupart du temps à l'aide d'un modèle Tobit), on tente d'expliquer les variations de la mesure θ dans l'ensemble de DMUs analysées par une série de variables indépendantes (Simar et Wilson 2007). Cependant, ce genre d'analyse ne permet pas d'intégrer les différences d'environnement à l'évaluation de l'efficacité comme telle puisque ce n'est qu'après l'estimation de l'efficacité que la discussion sur le contexte environnemental est menée.

C'est en fonction de cette lacune que les méthodes à trois étapes voient le jour (Fried et al., 2002). La première de ces étapes consiste toujours à obtenir le score d'efficacité radiale par DEA, mais une attention est également portée aux variables d'écart. Lors d'une seconde étape, les variables d'écarts sont régressées sur un certain nombre de variables environnementales où le terme d'erreur du modèle est de forme composée comme dans le modèle de la frontière stochastique SFA, c'est-à-dire qu'il inclut un terme asymétrique caractérisant l'inefficacité et un second terme de structure usuelle qui caractérise les chocs aléatoires. Cette étape permet alors de distinguer, parmi l'ensemble des facteurs identifiés, celui ou ceux pouvant affecter la performance des organisations. Il s'agit ensuite d'ajuster les observations. Plusieurs formules d'ajustement peuvent être proposées selon l'orientation du modèle DEA choisie, mais aussi en fonction des régressions effectuées. Ces formules d'ajustement visent à tenir compte des chocs aléatoires, à compenser les organisations opérant dans des environnements défavorables et à pénaliser les organisations avec des environnements plus favorables. Enfin, la troisième et dernière étape consiste à effectuer une seconde analyse DEA, cette fois avec les données ajustées. Les mesures d'efficacité en résultant sont alors épurées des effets associés à l'environnement et au bruit statistique.

Le second groupe de méthodes qui introduisent la stochasticité dans l'approche DEA, est reconnu sous l'appellation des *chance-constrained DEA models.* D'après Land, Lovell et Thore (1993), il est réaliste de croire qu'il existe une faible probabilité qu'une ou plusieurs contraintes du modèle DEA ne soient pas respectées à l'optimum du problème, puisque le processus de production est aléatoire. La frontière efficace doit ainsi être également aléatoire. De cette manière, il faut alors considérer un modèle comme celui-ci:

$$min_{\lambda,\theta} \quad \theta \tag{38}$$

$$sujet\ à: \quad prob\left[\sum_{j=1}^{n} x_{ij}\lambda_j \leq \theta x_{io}\right] \geq (1-\alpha) \quad i = 1, ..., m \tag{39}$$

$$\sum_{j=1}^{n} y_{rj}\lambda_j \geq y_{ro} \quad r = 1, ..., s \tag{40}$$

$$\lambda_j \geq 0 \quad \forall j \tag{41}$$

où α est le niveau de confiance, en général 5%. La résolution de ce problème non linéaire est complexe. Sans en exposer les détails, notons simplement que les solutions dépendent des hypothèses qui sont postulées sur les variances et les covariances des inputs[9].

Finalement, il existe une littérature importante sur la construction d'inférence valide à partir des résultats obtenus par DEA. Simar (1992) et Simar et Wilson (1998a; 1998b) jettent les bases de techniques de bootstrap destinées à générer la distribution des scores d'efficacité afin d'opérationnaliser des tests d'hypothèses. Kneip et al. (2008) fournissent une revue complète de ces méthodes. Pour notre part, nous nous contenterons de ce furtif coup d'oeil, car la discussion sur ces techniques est bien au-delà de l'objet de ce rapport de recherche. D'autre part, ajoutons que les méthodes de bootstrap DEA visent à gérer la variabilité des mesures d'efficacité due aux différents échantillonnages et ne concernent pas la prise en compte des erreurs de mesures ou de spécification.

Tout compte fait, l'ajout de composantes stochastiques permet d'assouplir les résultats complètement déterministes obtenus par DEA en considérant que certains bruits peuvent être inclus dans l'évaluation de l'efficacité.

[9] Parmi ces hypothèses notons que l'imposition d'une covariance nulle entre les inputs, $Cov(x_i, x_j) = 0$, $\forall i \neq j$, est celle qui est la plus commune (Cooper et al., 2007).

À notre sens, les méthodes à plusieurs étapes qui ajustent les données pour tenir compte des facteurs exogènes à l'organisation sont plus pertinentes que celles qui se contentent de régresser les scores d'efficacité sur des variables d'environnement. Toutefois, soulignons un problème élémentaire de ces techniques à plusieurs étapes. C'est que les mesures θ^* et les variables d'écart sont dépendantes les unes des autres, puisque l'un des fondements de la méthode DEA repose sur le principe d'évaluation relative. En termes économétriques cela signifie que les résultats sont corrélés entre eux d'une manière non évidente ce qui affecte alors la validité des résultats des régressions (Simar et Wilson, 2007). Enfin, réitérons que l'évaluation même de la performance doit intégrer l'effet de l'environnement et des chocs aléatoires, nous ne pouvons nous contenter d'expliquer comment l'environnement affecte le score d'efficacité sans l'avoir intégré à l'analyse préalablement.

4.2.2 Pertinence pour une analyse en santé

Parmi les facteurs exogènes aux organisations pouvant influencer leur niveau de performance, nous croyons que l'environnement est celui qui ait le plus de résonnance lorsque nous abordons une analyse en santé. Dans les analyses de production typique, il s'agirait d'identifier ce que peut représenter un environnement favorable et défavorable à l'aide de variable telle le niveau de concurrence au sein du marché ou encore la distance des distributeurs et des fournisseurs les plus près. Dans le secteur de la santé, il semble cependant difficile de déterminer des conditions qui puissent influencer l'efficacité de la sorte. La notion d'environnement ne s'exprime pas de la même manière.

En fait, l'environnement doit être considéré en regard de l'influence qu'il exerce sur les possibilités de production, sur ce que les organisations peuvent réaliser dans le cadre de leurs opérations. Pour un système de santé, la matière première des établissements producteurs de soins se constitue d'individus qui présentent des caractéristiques médicales et socio-démographiques diverses. Ainsi, la structure de la population a un impact fondamental sur le type d'interventions qui seront effectuées par chacune des organisations.

Par exemple, un centre hospitalier desservant un bassin d'individus dont la proportion de personnes âgées est plus élevée que le bassin desservi par un autre centre effectuera plus d'hospitalisations que celle-ci et ces hospitalisations seront

d'une durée moyenne plus longue. Est-ce pour autant que l'une est plus efficace que l'autre sur la base des hospitalisations?

La réponse est que nous l'ignorons en l'absence de facteurs tenant compte de la population. Il faut considérer que l'une des caractéristiques constitutives d'un système de santé est que la production est liée de près au caractère stochastique de la demande. Pourtant, même si cette demande de soins reste en partie imprévisible, certains déterminants et certains facteurs de risque peuvent être associés au niveau de soins demandés. Ce sont de tels facteurs qui doivent être traités lorsque nous évoquons la prise en compte de l'environnement.

Certains pourront objecter que n'importe quel producteur, peu importe le bien produit, est tributaire de la demande pour son produit et que les ventes réalisées dépendront des types d'acheteurs et de leurs préférences. Cependant, le lien n'est jamais aussi étroit que dans le cas de soins de santé où la demande détermine directement le nombre d'actes médicaux produits[10].

À la lumière de ce que nous venons de discuter, il importe alors d'intégrer les caractéristiques de la population à l'évaluation même de l'efficacité. Nous insistons encore sur le fait que les méthodes en deux étapes où les scores d'efficacité sont régressés sur un ensemble de variables environnementales sont largement insuffisantes. Quant aux techniques qui ajustent les données en fonction de l'environnement, comme elles demandent de spécifier une forme fonctionnelle entre les données et les variables environnementales, il demeure encore difficile d'en faire un usage adéquat puisque nous ignorons dans la plupart des cas cette forme, et ce, même si nous pouvons avoir une idée de l'influence sur l'état de santé de certains facteurs de risque.

Dans la prochaine section, nous verrons comment l'utilisation de variables non discrétionnaires nous semble être la façon de faire qui est la plus appropriée pour intégrer la structure de la population à l'évaluation de l'efficacité.

[10] Il est important de noter que le nombre d'actes médicaux effectués pourrait également dépendre de la capacité des organisations à produire les soins demandés dans le cas où la demande dépasse l'offre de soins disponibles. Le point que nous voulons faire ici est simplement de mentionner que s'il n'y a pas de malade, alors il n'y aura personne à soigner.

4.3 Modèles à variables non discrétionnaires

Lors de la présentation du modèle CCR, nous avons bien précisé qu'une des hypothèses fondamentales de la technique DEA est la possibilité des gestionnaires de contrôler le niveau d'inputs consommés par leur organisation. La mesure d'efficacité radiale θ repose de manière appréciable sur la validité de ce postulat. Néanmoins, il semble difficile de ne pas s'interroger sur la valeur d'une telle proposition. Bien qu'il puisse exister des secteurs de l'économie où cette hypothèse puisse s'appliquer de façon réaliste, nous croyons que certains facteurs de production demeurent généralement fixes, très certainement à court terme, mais peut-être aussi à long terme. Citons simplement en exemple les immobilisations d'une organisation.

Sur ces considérations, il devient impératif de tenir compte de l'existence de facteurs non discrétionnaires, sans quoi un biais est induit dans les analyses DEA en présumant que tous les gestionnaires peuvent disposer des inputs à leur aise.

4.3.1 Théorie

Le modèle que nous proposons pour tenir compte des contraintes des gestionnaires est tiré de Banker et Morey (1986). Le programme linéaire prend la forme suivante:

$$min_{\theta,\lambda,s^+,s^-} \quad \theta - \varepsilon(\sum_{i\in D} s_i^- + \sum_{r=1}^{s} s_r^+) \tag{42}$$

$$\theta x_{io} = \sum_{j=1}^{n} x_{ij}\lambda_j + s_i^-, \quad i \in D \tag{43}$$

$$x_{io} = \sum_{j=1}^{n} x_{ij}\lambda_j + s_i^-, \quad i \in ND \tag{44}$$

$$y_{ro} = \sum_{j=1}^{n} y_{rj}\lambda_j - s_r^+, \quad r = 1, ..., s \tag{45}$$

$$\lambda_j \geq 0, s_i^- \geq 0, s_r^+ \geq 0 \quad \forall i, j, r. \tag{46}$$

où ε est une très petite constante positive de façon à assurer que la minimisation de θ ne soit pas affectée par la présence des variables d'écart dans l'objectif. La

résolution de ce problème est équivalente à la procédure en deux étapes dont nous avons discuté en présentant le modèle CCR.

Les inputs se trouvent donc divisés en deux catégories, (43) réfère aux inputs discrétionnaires, c'est-à-dire les inputs sur lesquels le gestionnaire a un contrôle et (44) aux inputs non discrétionnaires, c'est-à-dire les inputs qui ne sont pas sous son contrôle. Remarquons également que les inputs non discrétionnaires sont omis de l'objectif du problème. Cette omission ne signifie cependant pas que ces variables n'affecteront pas le score d'efficacité. Regardons la forme dual du problème de minimisation:

$$max_{u,v} \quad z_o = \sum_{r=1}^{s} u_r y_{ro} - \sum_{i \in ND} v_i x_{io} \tag{47}$$

$$\sum_{r=1}^{s} u_r y_{rj} - \sum_{i \in ND} v_i x_{ij} - \sum_{i \in D} v_i x_{ij} \leq 0, \quad j = 1, ..., n \tag{48}$$

$$\sum_{i \in D} v_i x_{io} = 1 \tag{49}$$

$$v_i \geq \varepsilon, \quad i \in D \tag{50}$$

$$v_i \geq 0, \quad i \in ND \tag{51}$$

$$u_r \geq \varepsilon, \quad r = 1, ..., s \tag{52}$$

Cette forme dual fait intervenir uniquement les variables non discrétionnaires dans l'objectif du problème. Rappelons l'interprétation donnée aux poids u et v : ceux-ci représentent la contribution marginale de chacune des variables au score d'efficacité. Ainsi, en ayant recours à la théorie de la dualité mathématique, nous avons $\theta_o^* = z_o^* = \sum_{r=1}^{s} u_r^* y_{ro} - \sum_{i \in ND} v_i^* x_{io}$. Nous constatons alors que les variables non discrétionnaires ont pour effet de réduire le score d'efficacité lorsqu'à la solution optimale elles possèdent des poids positifs ($v_i^* > 0$). En faisant appel à la contrainte du *complementary slackness* selon laquelle $v_i^* s_i^{-*} = 0 \quad \forall\, i$, nous obtenons que $v_i^* > 0$, si $s_i^{-*} = 0$, ou encore $v_i^* = 0$, si $s_i^{-*} > 0$[11]. En d'autres termes, le score d'efficacité n'est pas affecté par les variables non discrétionnaires que si l'organisation dispose d'une quantité plus que nécessaire de ces inputs tel que le démontre la présence d'un slack positif.

[11] Sans oublier, bien sur, le cas où les deux variables seront simultanément nulles.

Dans le contexte de ce modèle, l'efficacité est atteinte si et seulement si $\theta^* = 1$ et tous les slacks apparaissants dans l'objectif en (42) sont nuls, $s_i^{-*} = 0$, $\forall i \in D$ et $s_r^{+*} = 0$, $\forall r$.

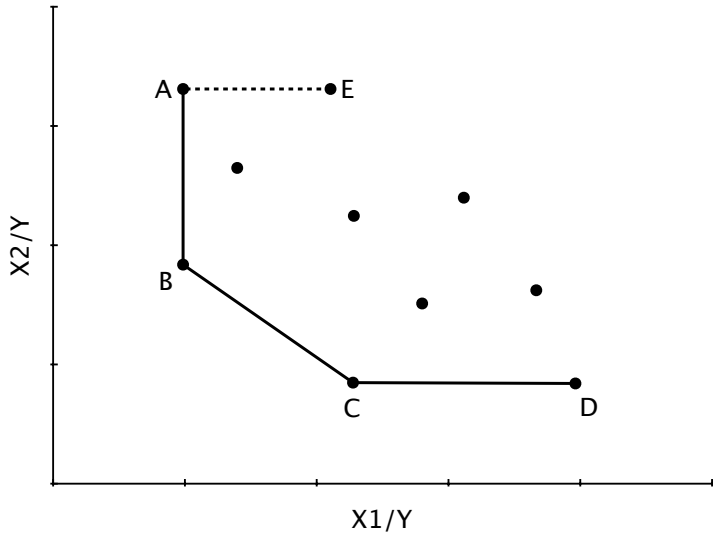

Fig. 14: Deux inputs, X_1 discrétionnaire et X_2 non discrétionnaire

La figure 14 illustre la manière de procéder avec les variables non discrétionnaires. Supposons que deux inputs sont utilisés pour produire un bien Y, X_1 est discrétionnaire et X_2 est non discrétionnaire. En évaluant l'efficacité du point A à l'aide d'un modèle BCC standard, nous trouverons qu'il est situé sur une portion inefficace de la frontière puisqu'il dispose d'une variable d'écart positive en regard de X_2. Toutefois, à l'aide d'un modèle BCC où X_2 est considérée non discrétionnaire, le point A sera efficace puisqu'en étant situé sur la frontière et n'ayant pas de slack positif en X_1, il répond à tous les critères que nous avons exposés pour caractériser l'efficacité avec variables non discrétionnaires. Si nous nous tournons vers le point E, celui-ci sera inefficace peut importe la nature de la variable X_2. Cependant, la nature discrétionnaire ou non discrétionnaire de X_2 influencera l'endroit où se situe sa projection efficace sur la frontière. Conséquemment, la mesure d'efficacité du point B sera différente sous les deux scénarios.

Plusieurs variantes ont résulté de ce modèle, certaines incluant la possibilité d'outputs non discrétionnaires, d'autres permettant un degré de contrôle variable sur les inputs et les outputs ou encore des limites sur la valeur que peuvent

prendre ces variables et finalement des modèles qui imposent des restrictions plus fortes sur les contraintes de façon à obtenir l'égalité stricte entre certaines variables observées et les combinaisons permises de l'ensemble de référence [12].

Nous présentons maintenant un second modèle (Banker et Morey, 1986) qui permet de prendre en considération que les DMUs analysées puissent appartenir à différentes catégories d'organisations et que le type de catégorie à laquelle appartient un établissement est hors de son contrôle.

Supposons un input x qui peut prendre L niveaux différents $(1, 2, ..., L)$, ceux-ci divisant l'ensemble des DMUs en L catégories. Plus précisément, pour l'ensemble de DMUs $N = \{1, 2, ..., n\} = K_1 \cup K_2 \cup ... \cup K_L$, où $K_f = \{j \mid j \in K \; et \; x_j = f\}$ et $K_i \cap K_j = \emptyset, i \neq j$, nous désirons évaluer une organisation $\text{DMU}_o \in K_l$ par rapport aux unités de sa catégorie, mais également des unités de catégories inférieures:

$$min_{\theta,\lambda} \quad \theta \tag{53}$$

$$sujet\ à \quad \sum_{j \in \cup_{f=1}^{l} K_f} x_{ij}\lambda_j = \theta x_{io} - s_i^- \quad i = 1, ..., m \tag{54}$$

$$\sum_{j \in \cup_{f=1}^{l} K_f} y_{rj}\lambda_j = y_{ro} + s_r^+ \quad r = 1, ..., s \tag{55}$$

$$\lambda_j \geq 0 \quad j = 1, ..., n \tag{56}$$

L'objectif de ce genre de modèle consiste donc à évaluer une DMU à partir de la performance des DMUs appartenant à sa propre catégorie, mais également à partir de la performance des DMUs de catégories inférieures puisque celles-ci sont supposées avoir des contraintes plus strictes à la performance dues à leur appartenance à une catégorie inférieure.

Des exemples types de catégories que nous pouvons construire sont donnés par un regroupement en fonction de la taille des villes mesurée par le nombre d'habitants, en fonction de la situation géographique, ou encore en fonction de la structure de marché. Cooper et al. (2004) utilisent une image fort simple pour illustrer l'idée de cette méthode: soit la présence ou non d'un comptoir de service à l'auto pour des restaurants. Dans ce cas, nous pouvons considérer qu'un restaurant avec un service à l'auto est avantagé par rapport à un autre qui n'en possède pas.

[12] Pour une discussion de ces modèles, voir Cooper et al., 2007, pp.215-27.

L'efficacité d'un restaurant sans comptoir à l'auto serait donc évaluée en fonction des performances des autres restaurants qui n'en possèdent pas, tandis qu'un restaurant qui en possède un serait évalué en regard de tous les restaurants, qu'ils aient un comptoir de service à l'auto ou non.

Ces modèles à regroupements variables, lorsqu'ils permettent de comparer une organisation avec des unités appartenant à sa catégorie ou de groupes inférieurs, présument cependant qu'il existe un quelconque ordre hiérarchique entre les différentes catégories. Autrement, si les catégories décrivent des systèmes différents, il est essentiel de mener des analyses séparées pour chacune d'entre elles, afin de ne pas confondre des caractéristiques propres à un groupe d'organisations pour de l'inefficacité.

Ainsi, les modèles à variables non discrétionnaires permettent de relâcher certaines propositions contraignantes du modèle DEA original. Par contre, il faut éviter de négliger les hypothèses sous-jacentes que nous ajoutons lorsque nous travaillons avec un modèle qui fait une distinction entre des variables de différentes natures. Comme nous le mentionnons précédemment, la présence de variables sur lesquelles le gestionnaire a une influence nulle ou limitée affecte l'évaluation de l'efficacité dans des cas bien particuliers. Si les ressources dont dispose une organisation sont disponibles dans une quantité qui dépasse celle qui est nécessaire, l'efficacité ne sera pas affectée. Ce genre d'énoncé suppose que l'excès de ressources ne soit pas nuisible. Il est difficile de croire que cela puisse être possible dans tous les cas. Certainement, des situations existeront où une trop grande disponibilité d'inputs influencera négativement la productivité des autres inputs ou encore diminuera le budget disponible pouvant être alloué à l'achat d'autres types d'inputs.

4.3.2 Pertinence pour une analyse en santé

La possibilité d'intégrer des variables non discrétionnaires est intéressante du point de vue d'une analyse d'efficacité en santé, car, nous le répétons, il est peu réaliste que l'ensemble des variables à inclure dans un modèle DEA soit sous le contrôle complet des gestionnaires des établissements. Dans le contexte du système de santé québécois, l'écho de cette réalité est tout à fait saillant.

Nous avons mentionné précédemment que le Ministère avait une emprise décisionnelle de taille sur le type de services de soins de santé offerts, mais aussi sur

les ressources y étant affectées. En effet, le Ministère alloue les budgets de fonctionnement à 18 agences régionales couvrant l'ensemble du territoire du Québec. Celles-ci doivent ensuite les répartir entre les organismes communautaires, les différents centres de santé et de services sociaux (CSSS)[13], les centres de réadaptation et les centres de protection de l'enfance et de la jeunesse de leur région respective (Québec. Loi sur la santé et les services sociaux. L.R.Q. ch.S-4.2, art. 286, 340, 350, 388 et 463).

Le Ministère autorise également l'allocation des budgets pour le maintien et le remplacement du parc d'équipement médical, tout comme il répartit les médecins dans l'ensemble des CSSS du système en fonction de plans d'effectifs médicaux qu'il élabore avec les agences et les établissements (Québec. Loi sur la santé et les services sociaux. L.R.Q. ch.S-4.2, art. 237-238 et 377). Pour terminer, c'est lui qui détermine la taille des établissements en fixant le nombre de lits au permis qu'ils peuvent opérer (Québec. Loi sur la santé et les services sociaux. L.R.Q. ch.S-4.2, art. 440). Les agences régionales, quant à elles, distribuent les nouveaux diplômés des facultés de médecine et sont aussi responsables d'élaborer des plans stratégiques et d'assurer la coordination des services sur leur territoire (Québec. Loi sur la santé et les services sociaux. L.R.Q. ch.S-4.2, art. 340 et 350-352).

Malgré cette présentation sommaire du système de santé québécois, nous remarquons rapidement que chaque établissement doit composer avec les contraintes que lui imposent le Ministère et l'agence de sa région. L'article 182.4 de la loi sur la santé et les services sociaux est d'ailleurs très révélateur de cette situation:

> *Le directeur général de l'établissement [...] veille au respect de la mission et des orientations stratégiques de l'établissement ainsi qu'à l'atteinte des objectifs annuels de celui-ci à l'intérieur du cadre de gestion qui lui est applicable et des ressources qui lui ont été allouées.*

[13] «En 2003, le gouvernement du Québec confie aux agences la responsabilité de mettre en place un nouveau mode d'organisation des services dans chaque région basé sur des réseaux locaux de services. Au cœur de chacun de ces réseaux locaux de services, on trouve un nouvel établissement appelé centre de santé et de services sociaux (CSSS) né de la fusion de centres locaux de services communautaires (CLSC), de centres d'hébergement et de soins de longue durée (CHSLD) et, dans la majorité des cas, d'un centre hospitalier. La création de ces réseaux locaux de services à l'échelle du Québec a pour objectif de rapprocher les services de la population et de les rendre plus accessibles, mieux coordonnés et continus». Ministère de la Santé et des Services sociaux. En ligne. http://www.msss.gouv.qc.ca/reseau/rls/index.php. (page consultée le 31 juillet 2009).

Ces contraintes ne consistent pas seulement à obliger les établissements à respecter les orientations du Ministère et de contribuer à atteindre les objectifs en matière de santé publique qu'il édicte, elles ont un impact fondamental sur leurs opérations quotidiennes et sur leurs possibilités de production pour reprendre les termes théoriques. De cette façon, ces contraintes opérationnelles doivent être intégrées à l'analyse de l'efficacité de toute organisation du système, puisqu'elles limitent les choix que les gestionnaires ont la possibilité de faire. La prise en compte de variables non discrétionnaires comme les bâtiments, les équipements et les effectifs médicaux est donc assurément nécessaire dans la construction d'un modèle DEA, sans quoi les résultats obtenus n'auront vraisemblablement aucun sens. Comment interpréter une mesure d'efficacité qui soit fondée sur la possibilité de réduire proportionnellement la consommation de ressources si les établissements ont justement des contraintes quant aux ressources dont ils peuvent disposer?

Ensuite, nous avons noté que l'inclusion de variables non discrétionnaires peut s'avérer utile pour tenir compte de l'environnement, plus précisément en intégrant la structure de la population à l'analyse. En considérant certains indicateurs socio-démographiques comme des inputs non discrétionnaires, nous obligeons le problème DEA à considérer le niveau de la variable comme fixée (voir la contrainte (44)). De cette façon, nous obtenons une évaluation de l'efficacité pour chaque DMU en fonction des caractéristiques de la population qui est desservie par ces établissements. Cette façon de procéder semble soulever beaucoup moins de problèmes techniques au niveau de la spécification et de la validité des modèles que les méthodes en plusieurs étapes que nous avons présenté à la section 4.2.1.

Un second point à aborder consiste à savoir s'il peut s'avérer judicieux de différencier certaines catégories d'établissements. Nous avons fait mention dans la section théorique que si différentes catégories d'organisations peuvent exister, évoluant chacune dans un univers où les contraintes à la performance se font de plus en plus restrictives, alors une analyse par DEA doit procéder en évaluant les DMUs de la catégorie la plus restrictive à la moins restrictive selon le programme linéaire présenté en (53).

En nous intéressant à un système de soins de santé, nous pouvons nous demander quels genres de catégories nous pourrions considérer qui puisse présenter ce genre d'emboîtement naturel, cette hiérarchisation des conditions de production. Les premières distinctions qui viennent en tête sont celles entre les établissements

des grands centres urbains et les établissements régionaux, celles entre les établissements à plusieurs missions et ceux n'en possédant qu'une seule, ou encore celles entre les centres hospitaliers universitaires et les centres non universitaires. Dans des contextes différents du contexte québécois, nous aurions pu penser à différencier les organisations sur leur statut (à but lucratif ou non), ou encore sur le type d'assurés qu'ils desservent.

Néanmoins, s'il est réaliste de penser que les possibilités de production puissent différer entre ces ensembles d'établissements, il demeure difficile de déterminer le type d'organisations qui doit faire face aux contraintes les plus rigides et les moins favorables. À défaut de pouvoir se positionner sur la nature des conditions sous lesquelles ces établissements opèrent, il s'avère peut-être plus approprié de considérer des analyses complètement séparées. Par la suite des tests non paramétriques pourront être effectués afin de déterminer si les multiples catégories partagent des frontières similaires ou différentes[14].

4.4 Modèles de qualité

La méthode DEA et ses nombreuses variantes mesurent l'efficacité à partir du processus de transformation des inputs en outputs, sans toutefois prendre en considération ce processus comme une variable du problème. Bien que la technique DEA suppose que l'inefficacité résulte d'un processus de production mal optimisé où les combinaisons de ressources ne sont pas adéquates, elle ignore complètement la nature des biens produits. La seule information pertinente à tirer est de nature quantitative: la quantité de ressources consommées et la quantité de produits livrés. Que les biens produits puissent différer en qualité n'est pas pris en compte dans l'évaluation de l'efficacité. À la limite, si la qualité est plus coûteuse en termes de ressources, l'existence d'un arbitrage entre qualité et efficacité est largement répandue dans l'opinion publique, le score d'efficacité sera plus faible et pourra être interprété comme de l'inefficacité plutôt que de relever la qualité des biens produits.

De sorte à ne pas confondre qualité et inefficacité, la méthode DEA nous suggère alors de procéder à des analyses distinctes pour plusieurs qualités de biens . Cependant, cette façon de faire est certainement contraignante et il serait souhaitable

[14] À ce sujet, le lecteur intéressé pourra consulter Cooper et al. (2007). pp.233-240.

de pouvoir intégrer la notion de qualité à l'analyse d'efficacité, surtout lorsqu'il est question du secteur de la santé où la qualité des services et des soins produits a un impact prépondérant.

4.4.1 Théorie

Certains auteurs sont au fait de cette question. D'ailleurs Ozcan et al. (2008) suggèrent d'intégrer des indicateurs de qualité à une analyse d'efficacité lorsque nous en disposons et proposent de les traiter comme des outputs indépendants, c'est-à-dire en ajoutant les indicateurs de qualité comme des outputs supplémentaires aux s outputs considérés précédemment. La formulation du programme linéaire demeure donc inchangée, il suffit simplement d'ajouter des dimensions.

Shimshak et al. (2009) notent toutefois que cette façon de procéder est insuffisante en regard du fondement de la méthode DEA qui consiste à accorder une liberté à chaque DMU dans la détermination des poids accordés à ses inputs et ses outputs dans la formulation fractionnaire en (1). En effet, il est possible qu'une organisation avec une piètre performance sur le plan de la qualité arrive à atteindre un ratio de 1, donc l'efficacité, en accordant des poids faibles, voire presque nuls, aux outputs mesurant la qualité. De la sorte, des DMUs pourront être caractérisées efficaces sans avoir une performance satisfaisante sur le plan de la qualité et pourront même faire partie de l'ensemble de référence pour d'autres DMUs que la méthode caractérise comme inefficaces.

Parmi les méthodes alternatives, nous comptons le modèle du Q-DEA pour *quality-adjusted DEA* de Sherman et Zhu (2006) qui propose de séparer l'analyse de l'efficacité de «production» et de l'efficacité de «qualité». Dans un premier temps, il s'agit d'estimer un modèle DEA standard et ensuite de transposer graphiquement les scores d'efficacité en fonction d'un indice de qualité sur un axe et de la mesure d'efficacité obtenue sur un autre. Les DMUs efficaces, mais ayant un faible indice de qualité sont par la suite retirées de l'échantillon de DMUs et un nouveau modèle DEA est estimé avec les DMUs restantes. Ainsi, les DMUs efficaces avec un faible indice de qualité se trouvent excluent des ensembles de références des unités inefficaces. Par contre, une telle procédure laisse en plan l'évaluation des DMUs retirées, ce qui est une de ses lacunes importantes.

Une autre série de modèles propose plutôt de conserver les mesures de qualité dans l'ensemble des outputs et d'éviter d'obtenir des poids trop faibles pour celles-ci en imposant des restrictions sur la valeur possible des poids (Allen et al., 1997). Ces restrictions peuvent prendre plusieurs formes, soit en imposant des limites absolues sur la valeur de certains poids, soit en imposant une restriction sur la valeur relative de ceux-ci. Quant à la façon de déterminer ces restrictions, il est possible d'incorporer le jugement des gestionnaires si ceux-ci détiennent de l'information pertinente ou encore d'utiliser des régressions pour s'informer sur la relation existante entre deux variables.

Klimberg et Puddicombe (1999) exposent quant à eux, une tout autre méthode pour tenir compte des aspects de qualité. Ils transforment le modèle original DEA en un modèle à objectifs multiples, c'est-à-dire un modèle où l'optimisation du programme linéaire est effectuée non plus en fonction d'un seul objectif (maximiser le ratio output/input), mais bien en fonction de plusieurs objectifs. Ces objectifs additionnels peuvent être tournés vers la qualité. Il s'agit donc de maximiser la somme de ces objectifs tout en imposant la contrainte que les poids de chacune des variables communes à plus d'un objectif soient près les uns des autres[15].

Somme toute, ces multiples modèles que nous venons d'exposer intègrent la notion de qualité en empêchant des unités d'atteindre la frontière d'efficacité sans performer à la fois autant sur la dimension de la productivité que sur la dimension de la qualité. Toutefois, elles omettent de discuter d'un aspect fondamental, c'est-à-dire l'existence d'une relation entre l'efficacité et la qualité.

Cette perspective mérite que nous reconsidérions sérieusement les façons de faire, puisque l'idée générale qu'il y ait un arbitrage à faire entre qualité et efficacité est très répandue. Afin d'examiner comment il est possible de modéliser un lien entre l'efficacité et la qualité, des modèles dits de congestion développés notamment par Färe et Svensson (1980) et Färe et al. (1989) peuvent être utilisés. Ceux-ci vont essentiellement modéliser le fait qu'il est possible que l'augmentation des inputs puisse entraîner une diminution de certains outputs. Parmi les exemples les plus concrets, pensons à une mine qui embauche des mineurs pour extraire les minéraux, l'augmentation du nombre de mineurs permet d'optimiser les méthodes

[15] Restreindre les poids des variables communes à plus d'un objectif à être semblables permet d'éviter que la solution au problème à objectifs multiples soit décomposée comme la solution aux modèles séparés (Shimshak et al., 2009).

de travail et d'extraire une quantité plus grande de minéraux (Cooper et al., 2007). Toutefois, après un certain nombre, l'ajout de mineurs supplémentaires n'aura plus aucun effet et en viendra même à ralentir le rythme de travail des autres mineurs, réduisant donc la quantité d'outputs produits. Cet exemple illustre l'effet de la congestion d'inputs sur l'output.

Nous pouvons par contre élargir les possibilités et considérer l'effet de congestion d'un output sur un autre output comme le propose Färe et al. (1989). Dans ce cas, la réflexion se pose davantage en termes d'outputs indésirables, c'est-à-dire que l'augmentation d'un output peut entraîner simultanément l'augmentation d'un autre output qui lui est indésirable. En retour, l'augmentation de cet output indésirable peut affecter d'autres aspect de la performance. Nous pouvons considérer la pollution de ce point de vue, étant donné que l'augmentation de la pollution est généralement associée avec un niveau de production plus important. Bien qu'initialement ces modèles de congestion ne se destinent pas à intégrer la dimension de la qualité au modèle DEA, nous illustrerons comment ceux-ci peuvent s'avérer appropriés pour une telle entreprise.

Prenons le modèle DEA initial où l'hypothèse de libre disponibilité forte suppose que l'augmentation de tous les outputs est désirable tout comme la diminution de tous les inputs. Ceci omet alors la possibilité que des outputs indésirables soient associés à la production d'un niveau d'outputs plus important. De manière à prendre en compte cette réalité, le programme linéaire DEA est modifié de la façon suivante (Clément et al., 2008):

$$max_{\tilde{\eta},\mu} \quad \tilde{\eta} \tag{57}$$

$$sujet\ à \quad \sum_{j=1}^{n} x_{ij}\mu_j \leq x_{io} \quad i = 1, ..., m \tag{58}$$

$$\sum_{j=1}^{n} y_{rj}\mu_j \geq \tilde{\eta} y_{ro} \quad r = 1, ..., s \tag{59}$$

$$\sum_{j=1}^{n} y_{tj}\mu_j k_j \geq \tilde{\eta} y_{to} \quad t = 1, ..., p \tag{60}$$

$$\mu_j \geq 0 \quad \forall\, j \tag{61}$$

$$\sum_{j=1}^{n} \mu_j = 1, \qquad 0 \leq k_j \leq 1 \quad \forall\, j \tag{62}$$

où $r = 1, ..., s$ sont les outputs désirables, $i = 1, ..., m$ les inputs et $t = 1, ..., p$ les outpus indésirables. La variable k est une variable d'intensité qui permet de prendre en compte la nature indésirable de certains outputs (Färe et al., 1989, p.92). Notons aussi que le problème DEA est orienté sur la maximisation du vecteur d'outputs étant donné le vecteur d'inputs, une caractéristique importante afin d'analyser les solutions en fonction des types d'outputs.

Ces modèles de congestion ont comme objectif de vérifier l'impact de la production d'outputs indésirables sur l'efficacité. Pour ce faire, il faut alors comparer les résultats obtenus sous le modèle (57) avec les résultats obtenus avec un modèle standard (voir le problème (20)), où la distinction des outputs en fonction de leur nature n'est pas considérée. La mesure de congestion est alors donnée pour chacune des DMUs par la relation qui suit:

$$C_o(x, y) = \frac{\eta^*}{\tilde{\eta}^*} \geq 1$$

Si les scores sont équivalents sous les deux modèles, alors il n'y a pas de congestion, la production d'outputs indésirables n'affecte pas la performance. À l'opposé, si la mesure de congestion est supérieure à 1, l'efficacité d'une DMU est réduite par la présence d'outputs indésirables (Clément et al., 2008).

Discutons maintenant de la pertinence d'utiliser un modèle de ce type pour intégrer la dimension de la qualité à une analyse d'efficacité par DEA. En fait, une grande liberté est laissée à l'analyste dans la définition des outputs qu'il considère indésirables. Il serait possible de considérer différentes mesures étant négativement liées à la qualité selon l'application. Par exemple, le pourcentage de retour de marchandises dans une analyse d'efficacité de magasins de vente au détail, le pourcentage de plaintes dans une application dans le secteur des services, etc. Enfin, nous donnerons plusieurs exemples en santé après avoir discuté de la pertinence d'ajouter une composante pour tenir compte de la qualité dans les analyses de ce secteur.

4.4.2 Pertinence pour une analyse en santé

La notion de qualité revêt un caractère particulier dans le contexte de soins de santé puisque parmi les résultats d'une mauvaise qualité de soins (erreurs médicales, mauvais diagnostics, etc.) nous pouvons compter le décès ou l'invalidité à long terme d'un patient. Il demeure alors essentiel d'intégrer une dimension de qualité aux analyses d'efficacité parce que celle-ci constitue une caractéristique fondamentale de tous les types de soins. Nous proposons d'utiliser la terminologie de Cambell et al. (2000), selon laquelle il existe deux dimensions à la qualité dans un système de santé, soit l'accès au système et aux différents services offerts et l'efficience des actes médicaux prodigués.

Premièrement, l'accessibilité au système se définit comme la capacité des individus à intégrer le système de santé afin de recevoir les soins que nécessite leur condition. D'une part, la notion d'accès fait intervenir la répartition géographique des établissements de soins de santé, à savoir si les individus sur l'ensemble d'un territoire peuvent accéder facilement au système de santé. D'autre part, nous devons aussi considérer la disponibilité des organisations à recevoir (temps d'attente) et à traiter (disponibilité des rendez-vous) des patients de diverses pathologies (Cambell et al., 2000).

La question de l'accessibilité est pertinente pour une analyse d'efficacité dans la mesure où l'accessibilité des services et l'offre de services peuvent être perçue comme un output en soi. Offrir la possibilité d'être soigné et le nombre d'actes médicaux prodigués ne sont pas deux mesures complètement similaires de la production d'un système de santé. Dans le cadre d'une analyse d'efficacité dans le domaine de la santé, nous considérons que l'offre de service doit être reconnue comme un output, et ce, même si ce service n'est jamais matérialisé.

Prenons un exemple très simple, celui d'une salle d'urgence pendant la nuit. Pour opérer cette salle d'urgence, un certain nombre de médecins et d'infirmières sont en service, toutefois certaines nuits aucun individu, ou très peu, peuvent être admis à l'urgence. Si cette salle d'urgence est comparée à une autre qui utilise environ le même nombre de médecins et d'infirmières, mais qui a eu beaucoup plus de cas à traiter durant la nuit, la méthode DEA identifiera la première salle d'urgence comme étant inefficace relativement à la deuxième. Pourtant, ce résultat n'est attribuable qu'au fait que plus d'individus se sont présentés à la seconde

salle d'urgence et n'est aucunement lié à l'efficacité technique et à la façon dont sont organisés les établissements. Devrait-on réduire le niveau des ressources de l'urgence inefficace au minimum, voire même aller jusqu'à fermer l'urgence durant la nuit?

Si cette solution semble inacceptable du point de vue politique, elle demeure plutôt logique du point de vue de l'efficacité qui est mesurée par DEA. Cet exemple élémentaire réitère bien qu'une mesure de l'efficacité technique doit tenir compte de ce qui n'est pas sous le contrôle des organisations telle la demande de soins induite par la structure de la population. Il illustre également que le cadre d'une analyse d'efficacité en santé doit être plus large que le simple rapport des outputs aux inputs.

Intéressons-nous maintenant à la deuxième composante de la qualité que nous avons identifiée: l'efficience des soins. Elle, fait intervenir l'interrogation suivante: une fois que le patient accède au système de santé et qu'il reçoit le traitement jugé nécessaire, les soins sont-ils effectués correctement, soit de manière à améliorer son état de santé? De nouveau, il est important de prendre en compte la qualité des soins puisqu'un nombre plus important d'actes médicaux peut signaler une plus grande efficacité des établissements ou alors que certains soins ayant entraîné des complications ont nécessité que des soins supplémentaires soient prodigués. À la limite, si un département de chirurgie effectue plus d'opérations qu'un autre, mais en y affectant moins de ressources, il est possible que plus de patients décèdent suite à leur opération dans ce premier établissement. Par contre, la méthode DEA l'identifiera comme étant plus efficace techniquement que la seconde, ce qui est relativement absurde lorsque l'on prend en compte la finalité d'un système de soins de santé.

De manière plus conceptuelle, la production d'un système de soins de santé est de «produire de la santé» ou d'améliorer l'état de santé de la population. La comptabilisation d'actes médicaux est une mesure imparfaite de l'output puisqu'elle omet la perspective de l'amélioration de l'état de santé. En ce sens, l'inclusion de la qualité dans l'analyse est un ajout qui permet de compenser cette omission.

Nous venons donc de justifier la valeur ajoutée par l'inclusion de la notion de qualité dans une analyse d'efficacité en santé, il reste cependant à s'interroger sur la façon dont la méthode DEA peut en tenir compte correctement.

D'abord, précisons que la qualité doit être conçue davantage comme un aspect complémentaire à la production des systèmes de santé qu'un aspect indépendant. Sur la question d'accessibilité, un plus grand nombre de soins implique généralement que, marginalement, le temps d'attente augmente et que la disponibilité des rendez-vous diminue. Du côté de l'efficience, des économies d'apprentissage (*learning by doing effects*) sont susceptibles d'émerger avec un niveau d'output supérieur. Cela dit, la plupart des modèles DEA, qu'ils intègrent la dimension qualitative dans un modèle distinct où qu'ils l'ajoutent comme un output indépendant, font abstraction de la notion de complémentarité entre production et qualité en la considérant plutôt comme une composante additive du modèle. La seule exception semble être les modèles de congestion qui considèrent que des outputs indésirables sont indissociables des autres outputs produits. Ce genre de modèle pourrait permettre d'utiliser en tant qu'«outputs indésirables» des indicateurs d'accessibilité, le temps d'attente moyen par exemple, et des indicateurs d'effectivité comme certains taux de maladies, des taux de récidives ou des taux de complications postopératoires.

Pour finir, la prise en compte de la qualité dans les analyses DEA en santé est essentielle parce qu'il faut tenir compte des différences entre les soins produits entre les multiples établissements. L'absence d'un tel contrôle pose un obstacle de taille à l'interprétation des mesures d'efficacité puisque nous pouvons nous retrouver avec des établissements caractérisés comme efficaces du point de vue technique, mais complètement inefficaces du point de vue de l'efficience des soins produits.

4.5 Modèles à ensemble non convexe

La convexité constitue un des fondements importants de la théorie microéconomique. La méthode DEA n'échappe pas à la norme et inclut la définition d'un ensemble convexe des possibilités de production parmi ses nombreuses hypothèses. Dans le modèle CCR, il s'agit d'un cône convexe et dans le modèle BCC nous parlons d'un ensemble convexe à segments linéaires.

Parmi les principes initiaux de la méthode DEA, nous comptons l'idée qui consiste à trouver l'ensemble enveloppant les données qui soit le plus conservateur possible, d'où la notion d'extrapolation minimum à partir des données. Dans cette

optique, certains auteurs comme Deprins, Simar et Tulkens (1984) avancent que l'imposition d'un ensemble convexe ne définit pas l'ensemble le plus près possible des observations. Ils proposent alors le modèle du free disposable hull (FDH).

4.5.1 Théorie

Le modèle FDH définit l'ensemble de production P à partir des observations (x_j, y_j) $j = 1, ..., n.$ de la manière suivante:

$$P_{FDH} = \{(x, y) \mid x \geq x_j, y \leq y_j, x, y \geq 0, j = 1, ..., n\}$$

Un point (x, y) appartient donc à l'ensemble FDH si ses inputs sont au moins aussi nombreux que ceux de n'importe quelle observation j et que ses outputs ne soient pas plus grands que les outputs associés à cette observation j. Un tel ensemble de production est représenté par la frontière de type escalier de la figure 15.

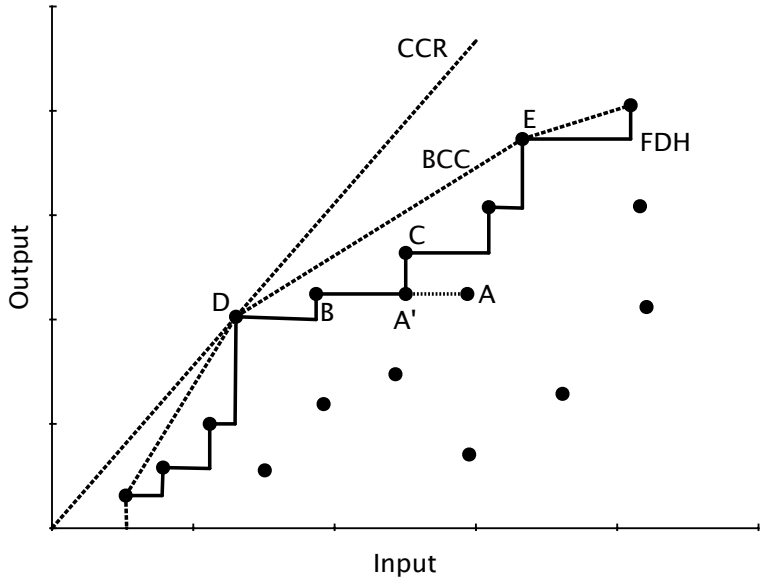

Fig. 15: Ensemble de production et frontière du modèle FDH

Si la frontière FDH semble bien différente des frontières CCR et BCC, le problème à résoudre pour évaluer l'efficacité d'un ensemble de DMUs demeure relativement près des modèles déjà présentés, à la différence que les λ sont transformés en variables binaires.

$$min_{\theta,\lambda} \quad \theta \tag{63}$$

$$sujet\ à \quad \theta x_{io} - \sum_{j=1}^{n} \lambda_j x_{ij} \geq 0 \quad i = 1, ..., m \tag{64}$$

$$y_{ro} - \sum_{j=1}^{n} \lambda_j y_{rj} \leq 0 \quad r = 1, ..., s \tag{65}$$

$$\sum_{j=1}^{n} \lambda_j = 1 \quad \lambda_j \in \{0, 1\} \tag{66}$$

Les modèles CCR, BCC et FDH se distinguent sur la base qu'ils supposent ou non la convexité de l'ensemble de production. Comme nous l'avons évoqué, la convexité entraîne la conséquence suivante: si deux points appartiennent à l'ensemble, alors n'importe quelle combinaison convexe de ceux-ci appartient également à l'ensemble. Dans l'estimation d'un ensemble de production sur la base de données empiriques, cela signifie que n'importe quelle combinaison convexe d'observations est réalisable, que cette combinaison soit observée ou non. Le modèle FDH, en se départissant de l'hypothèse de convexité, positionne l'ensemble de production uniquement à partir de points observés et en exclut les points situés sur un segment entre deux observations qui se situent sur la frontière.

Les effets du passage d'un ensemble convexe à un ensemble de type FDH sont doubles. D'une part, plus d'observations sont caractérisées comme efficaces tel que l'illustre la figure 15. D'autre part, les mesures d'inefficacité sont plus faibles, de par la distance réduite entre les DMUs se situant sur la frontière et les autres DMUs.

Faisons mention d'une dernière différence entre les modèles BCC, CCR et FDH; celle-ci concerne les projections sur la frontière pour les DMUs inefficaces. Prenons le point A de la figure 15, le modèle FDH suggère deux points dont les pratiques peuvent être imitées pour retrouver la frontière, soit B ou C. Le modèle BBC quant à lui suggère une combinaison convexe des points limites D et E qui n'ont pas de contrepartie empirique réelle, puisqu'aucune observation efficace ne se situe

entre ces deux points. Cela met en valeur une limite importante des modèles à ensemble convexe, soit la définition d'ensembles de référence pour les unités inefficaces qui ne sont soutenus par aucune observation de l'échantillon. Il va sans dire qu'il devient alors difficile de justifier qu'une organisation devrait adopter des meilleures pratiques en tentant de reproduire ce qu'une organisation fictive aurait pu atteindre.

4.5.2 Pertinence pour une analyse en santé

L'analyse proposée par le modèle FDH soulève un questionnement probant dans l'évaluation des établissements d'un système de santé en posant la question sur ce qui est possible de réaliser. La production de soins de santé requiert de multiples inputs à combiner de manière plus ou moins complexe en fonction des types d'interventions. Une particularité du domaine de la santé est le recours à une technologie avancée dans la production de nombreux soins tels la neurochirurgie, la chirurgie cardiaque, le traitement des cancers, le recours à l'imagerie médicale, etc. Plusieurs DMUs possèderont possiblement diverses technologies pour effectuer certaines tâches similaires, l'utilisation de celles-ci dépendant de la pratique des médecins, mais aussi des particularités pathologiques des patients. En ce sens, pouvons-nous penser qu'il soit possible de combiner diverses technologies pour produire le même genre de soins comme le suppose un ensemble convexe des possibilités de production?

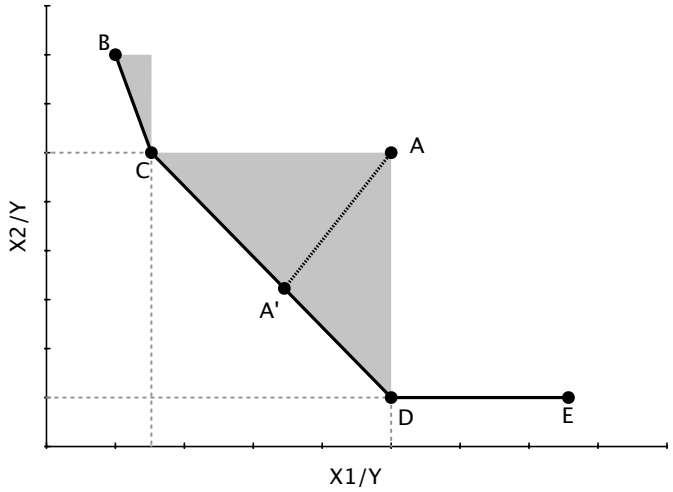

Fig. 16: Combinaison des technologies

Par exemple, à la figure 16, si l'efficacité est évaluée en fonction d'un modèle CCR ou BCC, le point A sera inefficace. Afin d'atteindre la frontière, il sera nécessaire de réduire proportionnellement tous les inputs jusqu'au point A' qui est une combinaison des technologies C et D. Ces technologies de production sont bien différentes en ce qui concerne leur façon de combiner les ressources disponibles, C est très intensive en X_2 et D est plutôt intensive en X_1. Les zones ombragées représentent les parties de l'ensemble de production CCR ou BCC qui sont retranchées de l'ensemble de production FDH. Sous le modèle CCR/BCC, le point A est ainsi projeté dans une zone qui n'est pas réalisable sous le modèle FDH. Il est donc nécessaire de se questionner s'il est possible de combiner les ressources X_1 et X_2 de manière plus équilibrée ou s'il n'y a que les technologies plus intensives dans l'une ou l'autre des ressources qui soient réalisables. En fait, la technologie utilisée au point A pourrait être la même qu'au point C, mais avec une surconsommation de la ressource X_1. Ces cas de figure illustrent bien l'importance qui doit être portée par l'analyste non seulement dans la sélection des variables à inclure dans le modèle, mais aussi dans la construction de celui-ci, c'est-à-dire sur ce qu'il suppose qu'il existe comme technologies.

Nous pouvons poursuivre la discussion sur les technologies dans le secteur de la santé en faisant la remarque que de nombreuses ressources peuvent être complémentaires, au sens où elles doivent être combinées dans une proportion donnée et qu'un niveau excédant cette proportion pour une ou l'autre des ressources ne permet pas d'atteindre un niveau de production supérieur. Dans le domaine des soins de santé, nous pouvons penser qu'une foule de relations de ce genre peuvent exister. Pour ne citer qu'un exemple, pensons à la relation entre le nombre de lits dans un département de chirurgie d'un centre hospitalier et le nombre de médecins qui y travaillent. En effet, augmenter le nombre de lits sans augmenter le nombre de médecins dans le département ne permettra pas d'effectuer nécessairement plus de chirurgies. La figure 17 permet de saisir la difficulté que pose une relation de complémentarité entre certains inputs.

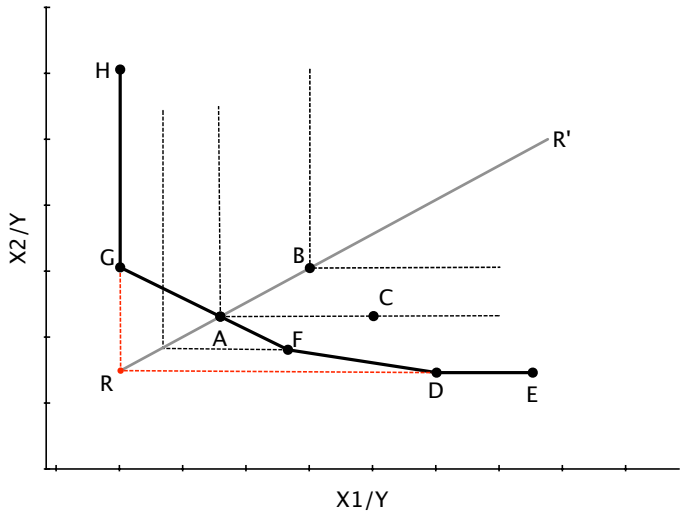

Fig. 17: Technologies complémentaires

La frontière efficace estimée par un modèle BCC est donnée par les segments $\overline{GA}$, $\overline{AF}$ et $\overline{FD}$. Cependant, si les inputs X_1 et X_2 sont complémentaires, la façon efficace de les combiner est donnée par le rayon $\overline{RR'}$. Il n'y a alors que le point A qui soit situé à la fois sur la frontière efficace BCC et sur le rayon de complémentarité. Mais encore, les points H, G, D, E sont situés sur des isoquants inférieurs à l'isoquant sur lequel est situé le point A, démontrant ainsi qu'il ne peut exister qu'un seul point efficace, le point R.

Nous avons poussé l'analyse à l'extrême en terme de simplicité, mais il apparaît bien que la méthode DEA permet difficilement de saisir la complexité amenée par la notion de complémentarité des ressources, une notion qui est pourtant centrale à l'analyse de la production d'un système de santé. En ce sens, comment serait-il possible de traiter les inputs complémentaires dans une analyse DEA? En fait, de façon à investiguer et à examiner les relations de complémentarité entre les diverses ressources qui sont utilisées, nous croyons que l'analyse sur la complémentarité des biens doit constituer une étape préliminaire à l'analyse DEA. Une fois les multiples types de complémentarité mis à jour, le modèle DEA peut être construit en n'utilisant qu'une seule des variables complémentaires, de sorte à éviter de réduire l'analyse d'efficacité en un point comme nous l'avons souligné plus tôt. Inclure une seule des variables complémentaires dans le modèle DEA contribue donc à résoudre la question de la substituabilité entre les inputs[16].

[16] Toutefois, une telle façon de faire impose que plusieurs aspects de l'efficacité devront être

4.6 Modèles intertemporels d'efficacité

Jusqu'à présent, les modèles DEA discutés ont eu pour cadre d'analyse un contexte statique. Cependant, l'intérêt des analyses d'efficacité peut consister à mesurer l'évolution dans le temps de l'efficacité d'un groupe d'organisations. Si le contexte statique convient pour identifier les sources d'inefficacités parmi les établissements d'un groupe, un contexte dynamique est quant à lui tout indiqué dans la perspective de faire un suivi de la performance des DMUs. La méthode DEA permet effectivement d'intégrer le concept d'analyse temporelle à partir d'une série d'observations de panel. Nous aborderons dans les sections qui suivent, deux techniques afin d'évaluer l'efficacité des DMUs dans le temps.

4.6.1 Théorie

La première de ces techniques est connue sous le nom de *Window Analysis.* Nous ne présenterons pas de manière formelle cette dernière puisque le problème d'évaluation de l'efficacité reste tel qu'il est présenté dans les divers modèles DEA et que la technique reste peu appliquée en pratique. Rapportons toutefois qu'il s'agit généralement d'étendre l'échantillon de DMUs évaluées en considérant chacune d'entre elles comme étant une organisation différente à chacune des périodes pour ensuite effectuer les analyses DEA de façon subséquente sur un sous-ensemble des périodes disponibles.

La deuxième technique que nous présentons maintenant demeure la plus utilisée dans l'analyse de l'efficacité avec des données de panel. Il s'agit de la technique des indices de productivité de Malmquist. Plus près d'une technique de statique comparative que d'une analyse dynamique, les indices de Malmquist se destinent essentiellement à évaluer le changement dans la performance d'un établissement entre deux périodes.

Afin d'évaluer l'évolution de l'efficacité d'une organisation, la méthode des indices de Malmquist identifie le mouvement de celle-ci dans le temps comme le produit de deux facteurs: la variation de son efficacité et le progrès technologique de

étudiés à l'extérieur du cadre DEA. Par exemple, la question de savoir si le niveau de complémentarité est le même pour tous les établissements ou encore si les inputs sont utilisés dans la bonne proportion sont des aspects qui devront être abordés dans une discussion portant précisément sur la complémentarité et non comme une modalité additionnelle de l'évaluation DEA.

l'industrie. C'est donc à la mesure de ces différents effets auxquels les indices s'attardent.

Tout d'abord, le changement dans l'efficacité d'une organisation peut être mesuré par le rapport des scores d'efficacité des deux périodes:

$$\Delta efficacité = \frac{\theta_2^*(x_o, y_o)^2}{\theta_1^*(x_o, y_o)^1} \tag{67}$$

$$= \frac{efficacité\ de\ (x_o, y_o)^2\ sujet\ à\ frontière\ de\ la\ période\ 2}{efficacité\ de\ (x_o, y_o)^1\ sujet\ à\ frontière\ de\ la\ période\ 1}$$

Ensuite, le progrès technologique que nous pouvons mesurer par le déplacement de la frontière entre les deux périodes est défini par la moyenne géométrique suivante:

$$\Delta frontière = (\phi_1 \phi_2)^{1/2} \tag{68}$$

$$où \quad \phi_1 = \frac{\theta_1^*(x_o, y_o)^1}{\theta_2^*(x_o, y_o)^1} \tag{69}$$

$$= \frac{efficacité\ de\ (x_o, y_o)^1\ sujet\ à\ frontière\ de\ la\ période\ 1}{efficacité\ de\ (x_o, y_o)^1\ sujet\ à\ frontière\ de\ la\ période\ 2}$$

$$\phi_2 = \frac{\theta_1^*(x_o, y_o)^2}{\theta_2^*(x_o, y_o)^1} \tag{70}$$

$$= \frac{efficacité\ de\ (x_o, y_o)^2\ sujet\ à\ frontière\ de\ la\ période\ 1}{efficacité\ de\ (x_o, y_o)^2\ sujet\ à\ frontière\ de\ la\ période\ 2}$$

L'indice de Malmquist est donc calculé comme étant le produit de ses deux effets:

$$IM = \Delta efficacité \times \Delta frontière \tag{71}$$

$$= \frac{\theta_2^*(x_o, y_o)^2}{\theta_1^*(x_o, y_o)^1} \times \left[\frac{\theta_1^*(x_o, y_o)^1}{\theta_2^*(x_o, y_o)^1} \frac{\theta_1^*(x_o, y_o)^2}{\theta_2^*(x_o, y_o)^1} \right]^{1/2} \tag{72}$$

$$= \left[\frac{\theta_1^*(x_o, y_o)^2}{\theta_1^*(x_o, y_o)^1} \frac{\theta_2^*(x_o, y_o)^2}{\theta_2^*(x_o, y_o)^1} \right]^{1/2} \tag{73}$$

Pour obtenir l'indice de Malmquist, il faut alors calculer deux mesures d'efficacité dans leur période respective, soit $\theta_1^*(x_o, y_o)^1$ et $\theta_2^*(x_o, y_o)^2$ et deux mesures faisant figure de comparaison intertemporelle, soit $\theta_1^*(x_o, y_o)^2$ et $\theta_2^*(x_o, y_o)^1$. Alors que ces deux premiers éléments s'obtiennent facilement en procédant de façon usuelle, il en est autrement pour les éléments faisant intervenir les deux périodes. Il faut porter notre attention sur le fait que nous devons évaluer une DMU à partir de sa performance qui n'est pas comprise dans l'ensemble d'observations servant à définir la frontière. Nous parlons donc d'une analyse exclusive en opposition à une analyse inclusive. Pour préciser $\theta_1^*(x_o, y_o)^2$ s'obtient comme:

$$\begin{aligned}
& \theta_1^*(x_o, y_o)^2 = min \quad \theta \\
sujet\ à \quad & \sum_{j=1}^{n} x_{ij}^1 \lambda_j = \theta x_{io}^2 - s_i^- \quad i = 1, ..., m \\
& \sum_{j=1}^{n} y_{rj}^1 \lambda_j = y_{ro}^2 + s_r^+ \quad r = 1, ..., s \\
& \lambda_j, s_i^-, s_r^+ \geq 0 \quad \forall i, j, r
\end{aligned}$$

où x_{ij}^t et y_{rj}^t avec $t = 1, 2$, $j = 1, ..., n$, $i = 1, ..., m$ et $r = 1, ..., s$ sont respectivement la consommation de l'input i et la production de l'output r de la DMU$_j$ au temps t. Dans le cas où $(x_o, y_o)^2$ se situe sous la frontière de la première période, c'est-à-dire que ce point est inefficace en regard de la première frontière, nous obtenons un résultat standard ($\theta^* \leq 1$). Si par contre, $(x_o, y_o)^2$ se trouve au-dessus de la frontière de la première période, nous obtenons un score d'efficacité qui est alors supérieur à 1.

Un aspect de discussion dans la littérature des indices de Malmquist porte sur cette question d'exclusion/inclusion dans l'évaluation de l'efficacité. La question qui se pose est de savoir si nous devons étendre le principe d'exclusion pour les observations évaluées dans le cadre de leur période respective, soit $\theta_1^*(x_o, y_o)^1$ et $\theta_2^*(x_o, y_o)^2$, puisque nous n'avons pas le choix de le faire pour les comparaisons intertemporelles. Il semble néanmoins qu'aucune des deux façons de faire ne se soit encore imposée (Cooper et al., 2007, p.340).

En somme, les indices de Malmquist ne sont qu'une extension que nous pouvons tirer à partir des résultats de la méthode DEA. La construction des indices ne nécessite pas de se positionner sur le choix de modèles DEA spécifiques (BCC, CCR, FDH, etc.) puisque tous les modèles présentés peuvent s'accommoder de cette analyse temporelle.

Une des exigences importantes pour la construction des indices de Malmquist est de pouvoir disposer de données longitudinales dont la mesure ne varie pas dans le temps, c'est-à-dire qu'il est nécessaire de disposer de données sur les inputs et les outputs qui soient mesurées de la même manière sur toute l'étendue des périodes analysées. À défaut de quoi, les comparaisons entre les périodes n'auront plus de sens.

De plus, il semble que l'analyse proposée par ces indices soit incomplète. En effet, une analyse économique intertemporelle dans un sens strict signifie considérer l'impact de décisions présentes pour le futur, et de décisions passées sur le présent. En ce sens, les indices de Malmquist n'abordent nullement une analyse dans ce cadre. Par exemple, il ne suffit pas d'inclure le capital en tant qu'input dans un modèle DEA, nous devons en tenir compte dans une perspective qui soit beaucoup plus large. D'ailleurs comme le note Jacobs et al. (2006):

> *Capital is by its nature deployed across time. The organisation in year t has enjoyed the benefits of past investments and it also leaves endowment for future periods in the form of investments undertaken in this and preceding periods. This endowment may be an important aspect of both the inputs and outputs of the health system.*

Enfin, ajoutons que les indices de Malmquist évacuent complètement de leur évaluation la notion d'inputs excédentaires et d'outputs déficitaires qui sont représentés par les variables d'écart positives dans la résolution des programmes linéaires

DEA. Ainsi, entre deux périodes, une organisation peut sembler s'être améliorée au niveau de l'efficacité au sens de la mesure d'efficacité radiale, mais cette organisation peut également avoir des inputs excédentaires plus importants à la seconde période. Dans ce cas, les mesures fournies par les indices de Malmquist ne seront pas capables de détecter correctement l'évolution de la performance dans le temps et pourront mener à des conclusions largement incorrectes.

4.6.2 Pertinence pour une analyse en santé

Le suivi de l'efficacité sur une période de temps est sans doute parmi les applications les plus immédiates pour laquelle nous souhaiterions obtenir une mesure de la performance des établissements de santé. La névralgie associée au contrôle des dépenses en santé et à l'optimisation des formes d'organisations sociales et politiques de la production des soins de santé laisse entendre qu'elle pourrait même en être le seul et unique objectif. Avec un tel objectif en tête, quelles significations pouvons-nous tirer des méthodes temporelles associées à la technique DEA?

D'abord, il faut comprendre que des contraintes structurelles se posent de façon tout aussi importante du point de vue dynamique que du point de vue statique. Tout comme il en était question pour l'allocation des ressources, les décisions d'investissements (agrandissements, achats d'équipements, etc.) sont prises par le MSSS en partenariat avec les agences régionales, et ce, même si ce sont les établissements qui expriment la nature de leurs différents besoins (Québec. Loi sur la santé et les services sociaux. L.R.Q. S-4.2, art. 112, 113 et 260-263). La nature publique du système fait en sorte que les investissements, avant d'être entérinés, doivent traverser un long processus constitué de paliers consultatifs qui décideront d'ailleurs de la cédule d'acquisitions et de remplacements des équipements. De cette manière, des établissements pourront bénéficier d'une nouvelle technologie ou d'une enveloppe budgétaire supplémentaire pour la modernisation de leurs installations avant que d'autres établissements puissent en faire autant.

Ces règles d'attribution posent un problème qui devient central dans l'évaluation de l'efficacité à l'aide d'un modèle DEA. Comme l'évaluation de chacun est une évaluation relative en fonction de ce que les autres établissements ont été capables d'atteindre, que certaines organisations puissent disposer de nouvelles technologies ou de budgets supplémentaires avant d'autres peut créer une distorsion dans

l'estimation de la frontière et dans la mesure des performances par rapport à celle-ci. Les organisations rationnées sur le plan technologique pourront apparaître moins efficaces à la seconde période, non pas parce que leur niveau d'efficacité a véritablement diminué, mais parce qu'elles sont comparées à une frontière qui s'est déplacée à cause d'organisations qui ont eu de meilleures opportunités de s'améliorer.

En fait, dès que nous quittons le cadre statique pour le cadre dynamique, la comparaison relative entre les organisations devient incongrue lorsque nous comprenons que leur déplacement dans l'espace des possibilités de production est le produit de leurs décisions et à la fois le produit des décisions ministérielles. Nous pensons qu'une meilleure pratique, afin de suivre l'évolution de l'efficacité d'un établissement, serait plutôt de conserver la frontière de la première période fixe et de mesurer la position de la seconde période par rapport à cette frontière de la première période. De la sorte, nous pourrions disposer d'une mesure d'évolution de l'efficacité entre deux périodes qui soit indépendante du mouvement des autres établissements de l'ensemble.

5 Les composantes des modèles DEA appliqués à la santé

Nous avons présenté l'étendue des possibilités d'analyse que procurent les diverses versions de la méthode DEA et l'écho de ces possibilités lorsque nous nous intéressons à l'évaluation d'établissements producteurs de soins de santé. En conservant ces nombreux aspects théoriques en tête, il est maintenant temps de s'attarder à la mise en oeuvre empirique de la technique DEA dans le domaine de la santé.

Débutons en mentionnant que l'utilisation de la méthode DEA à cette fin est répandue et gagne certainement en popularité. Hollingsworth (2008) reporte que plus de 300 articles et chapitres de livres qui s'y attardent ont été publiés à ce jour. La vaste littérature pose donc un problème de taille à quiconque tente d'effectuer une recension exhaustive de ces études. Néanmoins, nous considérons qu'un tour d'horizon s'avère tout de même fort intéressant en permettant d'identifier les principales tendances dans l'application de la méthode DEA à la santé. C'est donc ce que nous proposons dans cette section. Nous avons recensé plus de 35 articles publiés dans divers journaux que nous avons sélectionnés pour de multiples

raisons, soit parce qu'ils étaient largement cités par les autres articles, qu'ils utilisaient des modèles variés, qu'ils étaient récents, ou encore parce qu'ils amenaient une contribution importante du point de vue de l'avancement théorique.

À partir de ce tour d'horizon, nous souhaitons illustrer la façon dont la méthode DEA a été mise en oeuvre et plus précisément souligner les objectifs qui ont été poursuivis, les modèles utilisés, les résultats obtenus et la façon dont ils ont été interprétés, et finalement les conclusions qui en ont résulté. Enfin, mentionnons que cette section occupe une place essentielle dans ce rapport de recherche en poussant notre réflexion sur la portée et les limites de la méthode DEA.

5.1 Les objectifs poursuivis

Une des questions qui nous intéressent principalement dans ce rapport de recherche est la valeur de la méthode DEA, c'est-à-dire à quelles fins elle peut être utilisée et ce qu'elle permet d'accomplir. En fonction de ce questionnement, l'identification des objectifs et des questions posées par les applications pratiques est l'une des premières choses à mettre en perspective dans notre revue de la littérature.

Notre premier constat est que parmi les études recensées, divers objectifs semblent coexister. Toutefois, ceux-ci peuvent s'organiser autour de deux orientations principales. D'une part, nous retrouvons les articles qui s'intéressent à la mesure de l'efficacité et à sa relation à un certain nombre d'autres variables d'intérêts. Ceux-ci tentent d'évaluer l'efficacité d'un groupe d'organisations bien défini comme des départements de soins intensifs (Puig-Junoy, 1998), des hôpitaux (Grosskopf et Valdmanis, 1987; Harrison et al., 2004; Bilodeau et al., 2004; Chilingerian, 1995; Helmig et Lapsley, 2001), des foyers de soins pour personnes âgées (Fizel et Nunnikhoven, 1992; Kooreman, 1994; Chattopadhyay et Ray, 1996), des centres de santé locaux (Luoma et al., 1996; Crémieux et al., 2001), des mutuelles de santé (HMOs[17]) (Rosenman et al., 1997; Rollins et al., 2001) des centres de santé primaire (Garcia et al, 1999; Kirigia et al., 2004), des centres de santé psychiatrique (Kontodimopoulos et al., 2006), ou encore d'entités régionales (Bardey et Pichetti, 2004; Liu et al., 2006). Une fois la mesure de l'efficacité obtenue, l'intérêt de ces

[17] Les Health maintenance organizations (HMOs) sont une forme d'assurance aux États-Unis offrant une couverture sur des soins offerts seulement par les professionnels de la santé qui sont affiliés aux HMOs.

études est porté sur l'estimation des économies en ressources qui auraient pu être réalisées si toutes les organisations avaient opéré de façon efficace.

Dans cet objectif, nous retrouvons également des études qui tentent de faire la comparaison des niveaux d'efficacité des organisations sur la base de différents critères, le plus commun étant la distinction entre les institutions à but non lucratif et celles à but lucratif (Grosspkof et Valdmanis, 1987; Fizel et Nunnikhoven, 1992; Valdmanis, 1992; Kooreman, 1994; Rosenman et al., 1997; Rollins et al., 2001; Anderson et al., 2003; Bates et al., 2006; Färe et al., 1997; Bardey et Pichetti, 2004; Ferrier et al., 2006; Liu et al., 2006; Chilingerian, 1995; Kirigia et al. 2004; Chattopadhay et Ray, 1996; Harrison et al., 2004). Finalement, un troisième groupe d'études s'intéresse à l'évaluation de l'impact des changements de structure de rémunération ou des méthodes de fonctionnement sur le niveau d'efficacité (Luoma et al., 1996; Maniadakis et Thanassoulis, 2000; Martinussen et Midttum, 2004; Ferrier et al., 2006; Helmig et Lapsley, 2001).

D'autre part, la seconde tendance que nous relevons se distingue par une orientation plutôt théorique. En effet, un nombre important d'articles se concentre à questionner l'utilité et l'apport de la méthode DEA dans l'analyse de la performance. Ils recourent à des applications empiriques de la technique DEA afin d'illustrer son potentiel pour recommander des politiques à mettre en oeuvre, pour identifier les lacunes de fonctionnement des systèmes ou pour analyser la relation entre efficacité et qualité des soins (Salinas-Jiménez et Smith, 1996; Pina et Torres, 1992; Crémieux et al., 2001; Bilodeau et al., 2004; Banker et al., 1986; Clement et al., 2008; Nayar et Ozcan, 2008; Nunamaker, 1983; Grosskopf et Valdmanis, 1987; Shimshak et al., 2009; Kontodimopoulos et al., 2006). Nous relevons également dans cette catégorie, la présence d'études qui tentent d'analyser la sensibilité des résultats en fonction de différentes spécifications de modèles DEA (Grosskopf et Valdmanis, 1987; Valdmanis, 1992; Valdmanis et al., 2008; Spinks et Hollingsworth, 2009; Liu et al., 2006; Nayar et Ozcan, 2008; Puig-Junoy, 1998; Shimshak et al., 2009).

5.2 Les méthodes utilisées

Intéressons-nous maintenant aux méthodes utilisées. Faire état de celles-ci et des différents modèles qui sont employés dans les 35 études recensées demanderait de présenter 35 constructions distinctes. Nous proposons alors de regrouper la discussion en fonction des aspects théoriques présentés à la section 4 précédente et de mettre en évidence les façons de procéder qui nous semblent les plus importantes, soit parce qu'elles sont en accord avec ce que nous avons mentionné dans nos considérations théoriques ou encore parce qu'elles les contredisent fondamentalement.

5.2.1 Remarques générales

Commençons par mentionner que la majorité des études restent bien vagues sur le ou les modèles DEA qu'elles utilisent. En effet, très peu d'articles donnent explicitement le programme linéaire à partir duquel sont obtenus les scores d'efficacité, la plupart se contentant d'énumérer les inputs et les outputs considérés et de préciser l'orientation choisie (input ou output). À ce propos, il est intéressant de constater que la quasi-totalité des articles de notre échantillon considère la minimisation du vecteur d'inputs en maintenant le niveau d'output constant, un choix qui révèle bien l'importance accordée au contrôle des dépenses dans le secteur de la santé[18].

Poursuivons ensuite en remarquant qu'autant le modèle CCR que le modèle BCC est utilisé, mais qu'il existent une préférence nette pour le choix du modèle BCC lorsqu'il est question de caractériser les rendements d'échelle. Ce choix est guidé, comme le notent Valdmanis (1992), Kooreman (1994), Ferrier et al. (2006), et Valdmanis et al. (2008), par l'opportunité de pouvoir décomposer l'efficacité en fonction de l'efficacité technique totale, l'efficacité technique pure et l'efficacité d'échelle selon la formule que nous avons donnée en (37). Nous remarquons également que certains papiers utilisent des restrictions sur les poids de la formulation fractionnaire DEA (Luoma et al., 1996; Puig-Junoy, 1998; Kontodimopoulos et al., 2006), mais que cette pratique demeure limitée, et ce, en dépit du réalisme

[18] Seuls les articles de Salinas-Jiménez et Smith (1996), Färe et al. (1997), Ferrier et al. (2006), Spinks et Hollingsworth (2009), Clement et al. (2008) et Chattopadhyay et Ray (1996) avaient une orientation vers l'output.

qu'elle pourrait introduire dans les analyses comme nous en avons discuté à la section 4.4.1.

Enfin, comme les résultats sont largement tributaires des indicateurs choisis, le coeur de la discussion repose donc sur la sélection des variables qu'il faut intégrer aux modèles DEA. Une place importante est accordée à cette discussion dans les 35 articles. Pour débuter avec les inputs, il semble exister un consensus sur l'inclusion de variables comme les différents types de personnel dont les plus communs sont les médecins, les infirmières et les employés auxiliaires. La capacité à produire des soins est généralement mesurée par une prise en compte du nombre de lits ou parfois de la superficie des installations. La mesure de plusieurs dépenses comme les dépenses en médicaments, les coûts administratifs, les dépenses de fonctionnement sont aussi considérées comme des variables d'inputs pertinentes à inclure.

Du côté des outputs, la discussion se complique aisément, d'abord et avant tout parce que la production d'un système de santé est extrêmement complexe. L'objectif de toute institution d'un système de santé est d'améliorer l'état de santé de ses utilisateurs et non pas de produire un certain nombre d'actes médicaux. Idéalement, l'output d'un système de santé devrait alors refléter l'amélioration de l'état de santé des individus qui entrent en contact avec le système. Cependant, mesurer cette valeur ajoutée est loin d'être évident. Pour cette raison, des mesures alternatives ont été ciblées comme variables «proxy» telle l'utilisation de certains indicateurs de morbidité et de mortalité. Parmi les études de notre échantillon, seulement quelques-unes utilisent des indicateurs de santé de la population comme variables d'outputs (Färe et al., 1997; Spinks et Hollingsworth, 2009; Liu et al., 2006; Bardey et Pichetti, 2004), les autres ayant plutôt recours à une comptabilisation d'actes médicaux en regard de diverses catégories de patients.

En bref, il existe certainement une hétérogénéité dans les études analysées quant à la sélection des variables, surtout lorsqu'il est question de mesurer l'output. Si ces dissemblances sont combinées avec le caractère déterministe de la méthode, cela contribue à justifier l'entreprise menée par plusieurs auteurs qui consiste à considérer plus d'une spécification dans leur article afin de statuer de la robustesse de leurs résultats (Valdmanis, 1992; Luoma et al., 1996; Garcia et al., 1999; Anderson et al., 2003; Kontodimopoulos et al., 2006; Färe et al., 1997; Spinks et

Hollingsworth, 2009; Pina et Torres, 1992; Chilingerian, 1995; Nayar et Ozcan, 2008; Nunamaker, 1983; Parkin et Hollingsworth, 1997).

5.2.2 Tenir compte de l'environnement

De nombreuses études reconnaissent que des facteurs exogènes aux organisations peuvent avoir une influence sur leur performance. Parmi les études sélectionnées, plusieurs façons d'intégrer ces facteurs ont été appliquées. Valdmanis (1992) a décidé de n'inclure que les DMUs répondant à certaines caractéristiques assurant de la sorte l'homogénéité des unités dans son analyse[19]. Salinas-Jiménez et Smith (1996) ont intégré à leurs inputs un indice de maladie pondéré et le taux de chômage et Puig-Junoy (1998), la probabilité de survie à l'admission à l'hôpital et la catégorie de risque de mortalité[20]. Garcia et al. (1999) quant à eux, ont exprimé chaque input en fonction de la population à desservir . Bardey et Pichetti (2004) sont les seuls à avoir ajusté les données de leur analyse comme nous l'avons présenté à la section 4.2.1, c'est-à-dire en utilisant une méthode en trois étapes afin d'ajuster les données en fonction de l'environnement favorable ou défavorable des DMUs.

Cela dit, la méthode la plus commune pour intégrer l'environnement semble véritablement être l'utilisation des mesures radiales d'efficacité en tant que variable dépendante dans une régression, généralement un modèle Tobit, où les variables indépendantes sont représentées par différentes caractéristiques de la population et des caractéristiques de la structure des organisations et de leur marché (Fizel et Nunnikhoven, 1992; Kooreman, 1994; Luoma et al., 1996; Rosenman et al., 1997; Rollins et al., 2001; Anderson et al., 2003; Martinussen et Midttun, 2004; Bates et al., 2006; Bilodeau et al., 2004; Banker et al., 1986; Chilingerian, 1995; Puig-Junoy, 1998). Nous avons précisé qu'une telle façon de faire est questionnable du point de vue économétrique, notamment parce que les scores d'efficacité sont corrélés entre eux. Toutefois, à notre avis, un autre aspect est plus dérangeant

[19] D'autres articles ont bien entendu restreint leur échantillon d'organisations en fonction de certains critères, mais leur choix a été guidé dans la plupart des cas par des considérations qui ne sont pas en lien avec le contrôle des caractéristiques comme principe d'intégration de l'environnement.

[20] Notons que pour Salinas-Jiménez et Smith (1996) le programme linéaire est formulé de telle sorte à maximiser le vecteur d'outputs pour un niveau d'inputs donné et que pour Puig-Junoy (1998) les variables d'environnement sont considérées non discrétionnaires.

encore. Si les variables de la deuxième étape sont supposées être corrélées avec l'efficacité, pourquoi n'ont-elles pas été incluses dans le modèle DEA initialement? En fait, l'objectif de ces deuxièmes étapes qui incluent des régressions auxilliaires n'est pas tellement d'évaluer l'efficacité en contrôlant pour l'hétérogénéité contextuelle, mais plutôt d'identifier les déterminants de l'efficacité une fois celle-ci estimée. C'est ainsi que de nombreuses études procèdent pour évaluer la différence d'efficacité entre des organisations à but lucratif et des organisations à but non lucratif (Fizel et Nunnikhoven, 1992; Rosenman et al., 1997; Rollins et al., 2001; Anderson et al., 2003; Puig-Junoy, 1998), des organisations urbaines et des organisations rurales (Luoma et al., 1996; Bates et al., 2006), ou encore des organisations de taille différentes (Luoma et al., 1996; Rollins et al., 2001). Les articles ayant recours à cette technique se concentrent surtout sur la caractérisation d'un système, ils désirent obtenir une estimation de l'efficacité d'un groupe et de ses composantes et pouvoir ensuite en expliquer la cause. Dans ce sens, ils cadrent dans le premier groupe que nous avons ciblé en discutant des objectifs des applications empiriques.

Quant aux applications cadrant davantage dans le second groupe, nous remarquons que la nécessité de contrôler pour l'hétérogénéité se situe parmi leurs conclusions importantes (Salinas-Jiménez et Smith, 1996; Garcia et al., 1999; Clement et al., 2008; Spinks et Hollingsworth, 2009; Pina et Torres, 1992; Nunamaker, 1983). Cependant, la plupart ne vont pas plus loin et ne proposent pas explicitement de nouvelle méthode de contrôle. Clement et al. (2008), Valdmanis et al. (2008), et Banker et al. (1986) utilisent la comparaison de moyennes entre certains groupes, et Puig-Junoy (1998) a recours à une régression auxiliaire lors d'une seconde étape.

Il se dégage donc deux constats de notre revue de littérature en ce qui concerne la prise en compte de l'environnement dans lequel opèrent les établissements de santé. Premièrement, l'effet de contextes différents est reconnu comme un élément ayant un impact sur la performance des organisations et donc également sur leur efficacité. Deuxièmement, ces variables environnementales ne sont pas spécifiquement incluses dans l'évaluation de l'efficacité, c'est-à-dire de manière à contrôler pour l'hétérogénéité des contextes, puisque la majorité des chercheurs se contentent d'utiliser une régression secondaire visant à expliquer les variations d'un score radial d'efficacité.

5.2.3 Tenir compte des contraintes des gestionnaires

Parmi les hypothèses fondamentales de la méthode DEA, nous avons souligné l'importance jouée par la supposition du plein pouvoir des gestionnaires sur la budgétisation et l'allocation des ressources au sein de leur établissement. Dans la mesure où certains des inputs et des outputs ne sont pas sous le contrôle des organisations, il est logique du point de vue de l'efficacité organisationnelle de considérer la nature de ces variables dans la construction du modèle DEA, sans quoi les résultats feront peu de sens. L'ensemble des études que nous avons consultées fait néanmoins peu de cas de cette réalité, les discussions sur les ressources qui sont ou ne sont pas sous le contrôle des organisations étant réellement négligées.

Quelques articles seulement intègrent la notion de variables discrétionnaires et non discrétionnaires à leur analyse. Par exemple, Pina et Torres (1992) considèrent les coûts en personnel comme un input non discrétionnaire étant donné le taux important de syndicalisation des employés en santé. Puig-Junoy (1998) considère, à son tour, des variables comme le nombre de médecins et d'infirmières et la technologie disponible. L'étude de Crémieux et al. (2001) et celle de Bilodeau et al. (2004) analysent l'efficacité technique et allocative d'établissements de santé québécois et incluent trois inputs non discrétionnaires, soit le nombre de médecins, les équipements disponibles et les bâtiments. La raison invoquée par ces études pour traiter ces variables de cette façon est qu'elles sont hors du contrôle des gestionnaires, du moins à court terme. Cela est donc en accord avec ce que nous avons discuté précédemment.

L'horizon temporel a effectivement un impact dans une analyse d'efficacité, les possibilités à court terme étant bien différentes des possibilités à long terme. À défaut de préciser qu'une analyse est orientée vers une perspective de long terme, il semble que les analyses d'efficacité doivent respecter le caractère imminément fixe à court terme de certaines ressources tel que le capital. Kooreman (1994) et Fizel et Nunnikhoven (1992) reconnaissent cette réalité temporelle et décident d'exclure le capital de leur analyse étant donné l'absence de contrôle des organisations à cet égard. Une telle pratique apparaît courante dans la littérature. Faute de disposer d'un pouvoir discrétionnaire sur des ressources particulières, celles-ci sont simplement exclues du modèle. Par contre, la question qui se pose est de savoir si l'omission de variables de ce genre peut avoir un impact lors de l'évaluation de

l'efficacité. Prenons l'exemple qui nous intéresse le plus, le capital, celui-ci ayant certainement une influence sur les possibilités de production d'une organisation. Si nous l'excluons de l'analyse, la comparaison des performances observées sera faite indépendamment du niveau de capital. Des DMUs pourront être montrées inefficaces à partir des réalisations atteintes par d'autres DMUs potentiellement avec des niveaux de capital différents. De cette façon, la projection théorique que les unités inefficaces auraient dû être capable d'atteindre est biaisée. Sans doute, l'omission de variables pertinentes, et ce, même si ce sont des variables sur lesquelles les organisations ont un contrôle limité, est donc une façon incorrecte de procéder.

Dans un autre ordre d'idées, l'utilisation de variables non discrétionnaires peut également servir à modéliser l'environnement et les caractéristiques des patients auxquelles sont confrontées les organisations. Par exemple, Puig-Junoy (1998) considère deux inputs non discrétionnaires, soit la probabilité de survie d'un patient au moment de son admission à l'hôpital et son risque de mortalité. Chattopadhyay et Ray (1996), quant à eux, incluent parmi leurs outputs un indice d'autonomie des patients dans leur analyse des foyers de soins pour personnes âgées. Bien que ces variables ne sont pas ce qui est habituellement conçu comme variables non discrétionnaires, il semble pourtant pertinent d'intégrer ces données comme étant hors du contrôle des organisations.

Lorsque nous avons discuté des possibilités d'élargir le cadre des hypothèses concernant le pouvoir discrétionnaire des gestionnaires, nous avons abordé le fait que des organisations puissent appartenir à certains groupes ou certaines catégories. Dans ce cas, nous avons souligné l'importance de différencier les analyses en fonction de ces groupes. Des exemples probants en santé sont le type de médecins (spécialistes, généralistes, chirurgiens, etc.), l'orientation vers un but lucratif ou non, la diversité de services offerts par les établissements, etc. En pratique, l'utilisation de programmes différenciés pour des organisations de catégories diverses demeure l'exception plutôt que la norme. En effet, des 35 articles analysés, seulement Chilingerian (1995), Grosskopf et Valdmanis (1987) et Fizel et Nunnikhoven (1992) y ont eu recours. Plutôt, la méthode favorisée consiste à évaluer l'efficacité de l'ensemble des DMUs d'un échantillon ou d'une population, pour ensuite les regrouper en fonction de leur catégorie respective et de comparer leur niveau d'efficacité à partir de la moyenne des différentes catégories. Enfin,

aucune étude ne semble avoir utilisé un programme qui ordonnait les DMUs en fonction d'une hiérarchie des contraintes d'opération auxquelles celles-ci sont confrontées.

5.2.4 Tenir compte de la qualité

Nous en sommes maintenant à analyser la façon dont le concept de qualité a été incorporé empiriquement aux analyses DEA. Tout d'abord, précisons que la mesure de la qualité est une préoccupation notable d'une grande majorité d'articles. Cependant, comme c'était le cas lorsqu'il était question de prendre en compte l'environnement, plusieurs ne s'attardent qu'à mentionner qu'une analyse exhaustive devrait intégrer à son modèle une variable quelconque de qualité.

Poursuivons en soulignant que deux notions de qualité restent récurrentes dans les papiers consultés et que celles-ci ne correspondent cependant pas tout à fait aux deux aspects (accessibilité, efficience) de la qualité que nous avons relevés. D'un côté, nous retrouvons la qualité des services offerts par les établissements, ce que nous pouvons qualifier d'un point de vue sur la satisfaction de la clientèle. De l'autre, nous retrouvons la qualité des soins offerts, ce que nous pouvons davantage considérer comme un point de vue sur la capacité à soigner des établissements. Le point de vue qui sera considéré dépend généralement de l'application dont il est question, jamais les deux visions de la qualité ne seront étudiées en parallèle. Les articles qui s'intéressent à la mesure de la performance des foyers de soins pour personnes âgées (Kooreman, 1994), par exemple, auront tendance à adopter la première définition, tandis que ceux qui s'intéressent à la performance des unités de soins intensifs (Puig-Junoy, 1998) adopteront plutôt la seconde. À notre avis, la façon de définir la qualité revêt de l'importance dans la mesure où elle influencera le choix des variables pour la mesurer et la manière dont elle sera insérée dans l'analyse.

Les articles de notre échantillon ont ainsi choisi plusieurs variables pour mesurer la qualité dans leur analyse: Grosskopf et Valdmanis (1987) proposent de choisir le personnel hospitalier à l'exception des médecins comme un indicateur de la qualité des soins, Garcia et al. (1999) utilisent un indice technique de standards minimaux, Rollins et al. (2001) le nombre d'actes médicaux effectués d'urgence, Nayar et Ozcan (2008) le délai avant de prodiguer certains soins reliés à la pneu-

monie, Puig-Junoy (1998) une variable dichotomique indiquant la survie ou le décès du patient, Kooreman (1994) la présence d'un comité de patients et de procédures pour gérer les plaintes, Anderson et al. (2003) le résultat à l'inspection de l'établissement, Clement et al. (2008) des taux de mortalité ajustés en fonction des caractéristiques du patient, Shimshak et al. (2009) le nombre de patients sans cathéter et le nombre de patients sans handicap physique et Valdmanis et al. (2008) le nombre de cas d'infections postopératoires, de décès et de difficultés respiratoires. Il apparaît que le choix de la mesure de la qualité est dicté par la nature de l'application, mais également par la disponibilité des données[21].

Quant à la manière d'intégrer une mesure de qualité au programme DEA, nous remarquons encore une fois que les approches demeurent assez hétérogènes. Notons en premier lieu, les études qui ont simplement ajouté une dimension pour la qualité aux inputs ou aux outputs de leur modèle. Grosskopf et Valdmanis (1987), Garcia et al. (1999) et Shimshak et al. (2009) l'ont fait au niveau de l'input et Rollins et al. (2001), Nayar et Ozcan (2008) et Puig-Junoy (1998) au niveau de l'output. Ce faisant, nous avons évoqué à la section 4.4.1 que la simple considération de la qualité comme un output ou un input indépendant n'est pas indiquée dans la mesure où il est possible qu'à la solution optimale, la variable de qualité se soit vue attribuer un poids très faible, annulant de la sorte son effet dans le modèle. Parmi les articles que nous venons de citer, seuls Puig-Junoy (1998) et Shimshak et al. (2009) ont reconnu cette possibilité en imposant des restrictions sur les poids des variables de qualité qu'ils ont inclues.

Ensuite, Kooreman (1994), à son tour, incorpore quatre indicateurs de qualité comme des variables indépendantes dans une régression de seconde étape qu'il effectue à partir des mesures radiales. Les coefficients de ces indices lui permettent de déterminer s'il existe un arbitrage entre l'efficacité et la qualité. En dépit de la méthodologie que nous questionnons, cette façon de procéder nous semble inappropriée pour répondre à la question d'arbitrage, car la qualité n'est pas intégrée à l'évaluation même de la performance.

De façon similaire, Anderson et al. (2003) ont eux aussi eu recours à une technique qui séparait l'efficacité et la qualité. Ils ont utilisé une méthode DEA standard et, une fois l'efficacité estimée, ils ont divisé leur échantillon selon quatre niveaux

[21] D'ailleurs, plusieurs auteurs mentionnent que la qualité n'a pas été prise en compte dans leur modèle puisqu'ils ne disposaient pas de données à cet effet.

de qualité afin de statuer sur une relation possible entre la performance moyenne et la qualité.

Par ailleurs, nous avons recherché des papiers qui ont utilisé un modèle de congestion, afin de tenir compte de l'aspect complémentaire entre qualité et quantité, c'est-à-dire en examinant l'effet d'outputs indésirables étant associés à d'autres outputs sur l'efficacité. Bien que nous avons identifié le potentiel des modèles de congestion pour modéliser la relation entre certains aspects de la qualité et de la production, ce ne sont que très peu d'articles qui s'y sont intéressés. Valdmanis et al. (2008) ont évalué les effets de la prévalence de certains évènements comme le nombre d'infections et de troubles postopératoires sur l'efficacité des établissements hospitaliers, tandis que Clement et al. (2008) ont considéré l'impact du nombre d'infarctus, d'hémorragies et de pneumonies également sur la performance des hôpitaux. Ces mesures correspondent tout à fait à ce que nous considérions plus tôt en tant qu'indicateur de l'efficience des soins qui permettent de contrôler pour l'hétérogénéité des soins entre les établissements étudiés.

En dernier lieu, pendant que pour plusieurs, tenir compte de la qualité signifie contrôler pour le niveau de risque des patients et le degré de complexité des actes médicaux, que pour d'autres cela signifie prendre en note la résultante du contact avec le système de santé, nous remarquons que la composante de l'accessibilité n'est, quant à elle, curieusement jamais discutée.

Enfin, mentionnons qu'au sein de la communauté DEA, un consensus s'installe sur le fait que la qualité constitue une composante importante à prendre en considération lorsque l'on utilise la méthode DEA pour évaluer la performance d'établissements dans le domaine de la santé. Cependant, il semble que la façon de le faire soit encore en développement puisque les articles de notre échantillon conçoivent des définitions, et donc des mesures et des méthodes qui demeurent bien divergentes.

5.2.5 Effectuer une analyse temporelle

Une des applications pratiques évidentes de la méthode DEA est sa capacité à suivre l'évolution de l'efficacité dans le temps. Nous avons exposé dans notre section théorique comment les indices de Malmquist et les *window analysis* pouvaient s'avérer utiles dans cette perspective. Au niveau des applications empiriques, nous

remarquons pourtant que plusieurs études font un suivi de l'efficacité sans toutefois recourir à l'une de ces deux méthodes.

Effectivement, des articles comme ceux de Rollins et al. (2001), Martinussen et Midttun (2004), Crémieux et al. (2001), Bilodeau et al. (2004), Helmig et Lapsley (2001) et Nunamaker (1983) se contentent d'effectuer une analyse d'efficacité pour chacune des années de leur échantillon. Parmi ces études, aucune ne prend en considération ou ne mentionne que la technologie et donc la position de la frontière auraient pu changer entre les années étudiées. En s'intéressant de plus près à celles-ci, nous constatons que certaines d'entre elles sont orientées vers un objectif principal qui n'est pas de faire une étude longitudinale de l'efficacité (Martinussen et Midttun, 2004; Nunamaker, 1983). Par contre, pour les autres, les mesures d'efficacité sont explicitement comparées d'une période à l'autre afin de déterminer s'il y a eu des gains ou des pertes d'efficacité. Procéder de la sorte suppose alors que la frontière est restée la même sur la période étudiée et que les mouvements dans les scores d'efficacité sont simplement attribuables aux niveaux d'efficacité changeant des établissements.

En ce qui concerne les indices de Malmquist, notons que le recours à une telle façon de faire est relativement récent, mais que les études qui utilisent les indices de Malmquist commencent à se multiplier. Cependant, comme le fait remarquer Hollingsworth (2008) elles restent encore limitées en nombre, notamment parce que les programmes et les algorithmes pour implémenter la technique ne sont pas facilement disponibles. Dans notre échantillon d'articles, l'étude de Färe et al. (1997) et celle de Spinks et Hollingsworth (2009) utilisent les indices de Malmquist pour comparer l'évolution de l'efficacité pour un certain nombre de pays de l'OCDE. Toutefois, les études ne doivent pas se limiter aux comparaisons internationales, il serait possible de mener bien d'autres études en appliquant la technique de Malmquist à des secteurs plus traditionnels du secteur de la santé comme les hôpitaux ou encore les foyers de soins.

5.3 Les résultats obtenus

La revue des méthodes utilisées est intéressante en soi parce qu'elle permet de comprendre comment la théorie a été appliquée et quels sont les modèles favorisés en pratique. Cependant, cette revue est nécessaire puisqu'elle permet également

de contextualiser les résultats, une possibilité qui nous appelle particulièrement en regard de notre objectif qui est de comprendre ce que nous pouvons retirer de la méthode DEA dans une analyse en santé. Nous en sommes donc à présenter une synthèse des résultats de ces études, à savoir la façon dont ils ont été présentés, traités et interprétés.

Relevons quelques tendances principales. D'abord, la norme semble être de présenter les résultats sous forme de tableaux résumant les mesures radiales d'efficacité. Les scores y sont présentés pour chacune des unités de l'analyse, si le nombre de DMUs de l'échantillon le permet, sinon en valeur moyenne avec des mesures de dispersion comme l'écart-type, la valeur minimale et la valeur maximale. Ensuite, si le modèle utilisé justifie de le faire, l'efficacité est décomposée en fonction de l'efficacité technique, de l'efficacité technique pure et de l'efficacité d'échelle dans le but d'expliquer la contribution de chaque type d'efficacité à l'efficacité totale. Les articles utilisant plusieurs spécifications du modèle DEA présentent aussi les scores radiaux d'efficacité et leurs variations selon leurs multiples modèles.

Par ailleurs, les discussions s'orientent précisément sur la comparaison de ces différents scores, sur le nombre de DMUs ayant atteint un thêta de 1, parfois sur la moyenne des scores pour les unités inefficaces, mais plus probablement sur la moyenne des scores pour l'ensemble des unités. On caractérise généralement le niveau d'inefficacité d'un système soit le pourcentage de ressources qui aurait pu être économisé, la valeur monétaire de ces économies, ou encore des quantités d'inputs, en fonction de la mesure de contraction radiale moyenne obtenue. Une importance de premier plan est donc accordée aux mesures radiales.

Certains articles font néanmoins référence aux variables d'écart du problème dual, les slacks, en mentionnant dans leurs résultats, les inputs excédentaires et les outputs déficitaires. Kirigia et al. (2004) et Chilingerian, (1995) présentent les réductions en inputs et les augmentations en outputs nécessaires pour que chaque DMU atteigne l'efficacité, mais la discussion associée porte davantage sur les économies qui pourraient être réalisées que sur l'importance des variables d'écart dans la notion d'efficacité. Grosskopf et Valdmanis (1987) présentent également les inputs excédentaires moyens de leur modèle. Toutefois, ceux-ci sont utilisés pour déterminer si les hôpitaux à but lucratif consomment plus des différentes ressources lorsqu'ils sont comparés aux autres types d'hôpitaux, et pour conclure que la qualité de soins est plus élevée chez ces premiers. En fait, il n'y a que Valdmanis

et al. (2008) et Rollins et al. (2001) parmi l'ensemble d'études dont nous disposons qui utilisent les inputs excédentaires comme un facteur complémentaire à la mesure radiale pour attester de l'efficacité.

Nous avons déjà mentionné qu'un bon nombre d'articles s'intéresse à la comparaison de l'efficacité entre différents groupes: hôpitaux privés et hôpitaux publics, médecins spécialistes et médecins généralistes, divers types de mutuelles de santé, etc. À cette fin, les unités sont regroupées selon leur groupe de comparaison, afin de calculer la moyenne des mesures radiales d'efficacité. Par la suite, les moyennes des divers groupes sont comparées entre elles pour pouvoir conclure du groupe qui apparaît généralement plus efficace. Un score moyen plus élevé sera un signe d'une plus grande efficacité. Dans bien des cas, des tests non paramétriques de Mann-Whitney ou de Wilcoxon sont effectués afin d'attribuer de la robustesse aux résultats. Cela dit, deux remarques s'imposent.

Premièrement, l'utilisation de la moyenne d'efficacité n'est peut être pas le meilleur indicateur si nous prenons compte de la distribution des scores d'efficacité. En effet, cette distribution est fort probablement asymétrique étant donné le nombre potentiellement important de scores égaux à 1 et l'incapacité de différencier des niveaux d'efficacité pour les unités dont le score atteint ce seuil. L'utilisation du score moyen pour les unités inefficaces seulement semblerait être plus appropriée en constituant une mesure moins biaisée.

Deuxièmement, la procédure qui consiste à estimer un modèle DEA avec un ensemble d'organisations, à obtenir les scores pour chacune d'entre elles, pour ensuite les regrouper en fonction de certaines caractéristiques et discuter des différences moyennes d'efficacité entre ces groupes, semble manquer de rigueur théorique. Les scores DEA sont obtenus relativement aux réalisations de toutes les organisations de l'ensemble. Ainsi, les scores d'efficacité d'un groupe X sont dépendants des réalisations des organisations du groupe Y et du groupe X. Cette dépendance est illustrée dans les ensembles de référence qu'identifie la méthode pour chaque organisation, où des éléments des deux groupes peuvent s'y retrouver et servir de référent. À partir du moment où ces scores sont inextricablement liés, la comparaison de moyenne entre des groupes et éventuellement la conclusion que l'un est plus efficace que les autres fait difficilement du sens.

Ensuite, pour un bon nombre d'études, une fois les résultats des modèles DEA présentés, l'intérêt est porté sur la détermination des facteurs qui influencent

l'efficacité. Tel que noté précédemment, l'utilisation des scores d'efficacité dans une régression sur ces facteurs potentiels est la méthode d'analyse favorisée. Les articles s'attardent donc à présenter les résultats des régressions, il s'en suit alors une discussion sur la significativité des divers facteurs et de leur influence sur le niveau d'efficacité. Étonnamment, cette section occupe une place très importante des papiers, bien souvent plus importante que la discussion sur les résultats DEA (Fizel et Nunnikhoven, 1992; Kooreman, 1994; Luoma et al., 1996; Rosenman et al., 1997; Rollins et al., 2001; Anderson et al., 2003; Martinussen et Midttun, 2004; Bates et al., 2006; Bilodeau et al., 2004; Banker et al., 1986; Chilingerian, 1995; Puig-Junoy, 1998). Ceci nous laisse donc entendre que l'objectif principal de ces études est de se positionner sur les facteurs favorisants ou non l'efficacité et que l'obtention d'une mesure de l'efficacité devient utile seulement en fonction de cet objectif.

Finalement, nous trouvons important de souligner que, puisque la majorité des articles utilise un modèle DEA sous sa forme dual, les poids de la forme fractionnaire ne sont pas directement obtenus. Ils sont de la sorte ignorés lors de la présentation des résultats et des discussions, et ce, en dépit de leur interprétation économique intéressante. Nous pourrions nous attendre alors à ce que les articles présentent, pour les organisations inefficaces, les organisations identifiées comme référents à partir des multiplicateurs λ. Cependant, il n'en est rien, seuls Bardey et Pichetti (2004) exposent les établissements de l'ensemble de référence de chaque unité, ce qui est assez étonnant puisque la mesure de l'efficacité repose fondamentalement sur l'identification de ces référents.

5.4 Conclusions et perspectives

Enfin, nous abordons les conclusions qui sont tirées par les auteurs des articles utilisant la méthode DEA. À cet effet, la conclusion principale de toutes ces études est que la mesure de l'inefficacité est possible, tout comme sa quantification. Bien que quelques articles expriment un certain recul face à de telles conclusions, principalement ceux qui sont destinés à évaluer l'utilité et à explorer les résultats de la méthode DEA, plusieurs se contentent simplement d'énoncer l'ampleur des dépenses qui auraient pu être économisées si tous opéraient à un niveau efficace. Conséquemment, plusieurs se permettent ensuite de recommander ou d'entériner

la mise en oeuvre de politiques visant à optimiser l'utilisation des ressources, à réduire les budgets ou à modifier les structures de compensation. Par exemple, Harisson et al. (2004) suggèrent qu'il serait nécessaire de réorganiser la force de travail des hôpitaux fédéraux américains et qu'il est possible d'identifier les établissements les moins performants qui devraient cesser leurs opérations. Kirigia et al. (2004) vont même jusqu'à proposer que le personnel administratif excédentaire pourrait être envoyé à la retraite afin d'éliminer l'inefficacité qui provient de ressources excessives. Bardey et Pichetti (2004) proposent que les scores DEA peuvent être utiles afin de calculer des cibles pour instaurer un système de péréquation régionale en matière de dépenses de santé en France.

Les études qui ont comparé l'efficacité entre divers groupes, quant à elles, proposent que des structures particulières devraient être adoptées de par leur niveau d'efficacité plus élevé. Bates et al. (2006) suggèrent qu'il faudrait inciter les individus à joindre les plans de protection offert par les HMOs et restructurer le système d'assurance santé aux États-Unis. Autrement, Luoma et al. (1996) proposent la réforme du système de subventions aux établissements afin de profiter de gains d'efficacité.

Si nous nous attardons aux discussions sur les limites posées par l'utilisation d'un modèle DEA, nous remarquons qu'un grand nombre d'articles souligne que l'une des contraintes importantes de leur étude reste que les résultats obtenus ne sont valables que pour leur échantillon particulier (Helmig et Lapsley, 2001; Chilingerian, 1995; Färe et al., 1997; Martinussen et Midttun, 2004). Ces bémols à la généralisation de leurs conclusions semblent alors en accord avec le fait que l'un des objectifs importants de ces études soit la compréhension des déterminants de l'efficacité technique dans sons sens large. Le recours à une analyse par DEA cadre donc avec ce besoin d'obtenir une évaluation de la performance pour ensuite pouvoir s'attarder à l'expliquer.

Certaines des études qui ont testé plusieurs spécifications ont obtenu des résultats différents, alors que d'autres ont obtenu des résultats qui étaient en accord dans les multiples modèles. Alors, parmi les autres limites, nous notons l'expression d'une distanciation en regard des résultats, une distance qui émane de la nécessité de faire un choix sur les variables à inclure dans l'analyse. De cette façon, choisir un vecteur d'inputs et d'outputs semble être de manière appréciable la plus grande difficulté rencontrée.

Étonnamment, bien peu d'articles mentionnent que les hypothèses derrière la méthode DEA peuvent limiter de façon similaire la portée des résultats et des conclusions. Bien que certains auteurs précisent que la technique n'est valide que dans le cadre de la théorie qu'elle expose, il demeure que bien peu de cas est fait de cette théorie. La possibilité de disposer d'une mesure pour évaluer la performance outrepasse souvent les considérations quant au contexte dans lequel celle-ci peut s'avérer valide.

À la lumière de cette revue de multiples études, nous pouvons nous questionner sur ce que permet de faire la méthode DEA lorsqu'elle est appliquée à une analyse d'efficacité en santé. Nous avons noté que les discussions s'orientent presque exclusivement sur l'inefficacité radiale, l'inefficacité non proportionnelle provenant des slacks étant complètement oblitérée des analyses. Pourtant, la définition de l'efficacité technique de Pareto-Koopmans demande que les deux types d'efficacité soient atteints[22]. Si nous ajoutons à cela la difficulté d'intégrer une dimension de la qualité des soins et de contrôler pour l'environnement dans lequel opèrent les établissements, pouvons-nous conclure que nous disposons réellement d'une mesure de l'efficacité technique? Plus important encore, pouvons-nous utiliser cette mesure à des fins de prévisions budgétaires, ou encore à des fins d'allocation des ressources comme l'ont fait les études empiriques? À notre avis, le défi posé par la construction d'un programme DEA valide dans le domaine de la santé est de taille et demeure possiblement irrésolu à ce jour. Pour cette raison, nous pensons qu'un meilleur usage de la méthode DEA et de ses résultats n'est pas en tant que mesure spécifique de l'efficacité, mais plutôt en tant qu'indicateur d'efficacité comme nous le proposerons dans la dernière section de ce rapport de recherche.

[22] Un rappel de la définition: «Une organisation est efficace techniquement si et seulement s'il lui est impossible d'augmenter son niveau d'output pour un niveau de ressources donné, ou encore de diminuer le niveau de ressources consommées pour produire une quantité donnée d'output» (Cooper et al., 2004).

6 Rechercher un indicateur

6.1 Motivations

Avant d'aborder comment un indicateur d'efficacité pourrait être construit, énonçons quelques motivations sous-jacentes à sa recherche. Nous avons mentionné en introduction à ce travail que le contrôle des dépenses de santé est au coeur des préoccupations gouvernementales et oriente bien souvent les politiques qui seront mises en oeuvre; nous n'avons qu'à penser au virage ambulatoire des années 1990 ou encore à la régionalisation du système depuis 2001. Pour les économistes, ce contrôle des dépenses, mais aussi la maximisation des gains de santé doit, entre autres choses, passer par la production efficace de soins de santé par les divers établissements du système. Cette question d'efficacité préoccupe d'ailleurs le ministère de la Santé et des Services sociaux qui a mis en place à travers les années différents groupes de consultation devant s'intéresser à un tel sujet. Parmi les axes d'intérêts de ces groupes, nous comptons la recherche d'indicateurs de mesure de performance au niveau des établissements, des organisations régionales et également du système agrégé.

Pour notre part, nous pensons que la structure du système de santé québécois fait en sorte que l'efficacité relève surtout des établissements, les autres paliers encadrant simplement les actions qu'accomplissent ces premiers. Pour cette raison, la prospection d'indicateurs de performance organisationnelle devrait représenter un objectif de premier plan et orienter la recherche en économie de la santé. Il s'agirait donc de trouver des indicateurs qui permettent de suivre l'évolution des organisations vers une cible: l'atteinte de l'efficacité.

Cela étant dit, la méthode DEA propose spécifiquement une analyse de l'efficacité au niveau organisationnel, c'est donc une caractéristique importante qui nous pousse à évaluer son potentiel en tant qu'indicateur d'efficacité.

6.2 Propriétés souhaitables d'un indicateur

Un indicateur est un outil de mesure construit afin d'observer régulièrement la dynamique et la progression d'un phénomène ou d'un comportement, en le situant en relation avec un ou des objectifs précis. Le choix d'un indicateur repose alors

sur certains critères de sélection, notamment sur des propriétés heuristiques, mais aussi sur des critères propres au contexte d'utilisation de l'indicateur et au concept théorique que nous désirons mesurer.

Parmi les critères de sélection que nous pouvons qualifier de méthodologiques, notons la disponibilité des données nécessaires à la construction de l'indicateur, la fiabilité de celui-ci à produire une mesure avec le minimum de variations aléatoires et finalement sa validité, c'est-à-dire sa capacité à mesurer le phénomène à l'étude. Cette dernière condition est certainement à la fois la plus difficile et la plus fondamentale à respecter. Comme le font remarquer Champagne et al. (2005): «La validité est une caractéristique très subtile des indicateurs». En effet, nous pouvons décliner la validité en cinq aspects distincts: la validité apparente, la validité de contenu, la validité pratique, la validité théorique et la validité causale (Champagne et al., 2005). La validité apparente consiste en la forme la plus générale de validité, elle est destinée à vérifier que l'indicateur semble mesurer le concept approprié. La validité de contenu, quant à elle, vise à examiner que toutes les facettes importantes du concept théorique sont prises en compte dans l'indicateur. Ensuite, la validité pratique réfère à la capacité de l'indicateur à mesurer quelque chose qui soit corrélé avec un comportement ou une action donnée, alors que la validité théorique se penche sur l'opérationnalisation du concept théorique, à savoir sur l'établissement de mesures opérationnelles conforme à la théorie. Pour finir, la validité causale fait référence à la spécification du lien de cause à effet entre l'indicateur et le concept théorique.

En ce qui concerne l'utilisation de la mesure DEA comme indicateur d'efficacité, la notion de validité soulève un bon nombre de questions: la mesure DEA semble-t-elle bel et bien produire une mesure qui s'apparente à l'efficacité d'un établissement? Quels sont les aspects importants de l'efficacité technique en santé, comment pouvons-nous les intégrer au modèle DEA? Est-ce que l'utilisation de ressources et les outputs produits peuvent être des variables pertinentes d'une mesure de l'efficacité d'un établissement? Est-ce que le choix d'inputs et d'outputs d'une organisation est corrélé avec la mesure d'efficacité choisie? Est-ce que l'amélioration de l'indicateur peut être attribuée uniquement à une meilleure efficacité ?

Sans plus attendre, nous pouvons répondre à un certain nombre de ces questions sur la base des discussions que nous avons menées dans les sections qui précèdent

dans ce rapport de recherche. Ainsi, la mesure DEA, de manière générale, semble correspondre à un indicateur de l'efficacité technique d'une organisation de soins de santé. L'idée qui consiste à comparer les établissements de santé entre eux, sur la base de leur consommation en ressources et du nombre et du type de soins produits, s'arrime relativement bien à la définition de l'efficacité. Cette définition est relative certes, mais c'est elle qui demeure tout de même la plus accessible. Enfin, la façon dont la méthode évalue la performance d'une organisation et les variables utilisées, à condition de prendre en considération les remarques que nous avons formulées à travers ce travail, sont également en accord avec les aspects importants à considérer dans une évaluation d'efficacité.

Si la méthode DEA apparaît ainsi opportune afin de disposer d'un indicateur de l'efficacité des organisations de soins de santé, le modèle qui s'avère approprié, un modèle CCR, BCC ou encore une autre variante, ne peut pas être déterminé strictement en fonction des critères de sélection que nous venons de citer. Nous devons, en outre, spécifier certaines propriétés mathématiques que doit posséder un indicateur qui tente de mesurer l'efficacité technique.

Supposons, l'indicateur $\psi(X_o, Y_o)$, où X est un vecteur de m inputs et Y est un vecteur de s outputs pour un établissement donné, nous souhaiterions que $\psi(X_o, Y_o)$ possède les propriétés suivantes (Pastor et al., 1999):

1. $0 < \psi(X_o, Y_o) \leq 1$, l'indicateur s'interprète alors comme un pourcentage d'inefficacité;

2. Si $\psi(X_o, Y_o) = 1$, alors l'établissement est efficace au sens de la définition de Pareto-Koopmans;

3. $\psi(X_o, Y_o)$ est invariant aux unités de mesure des inputs et des outputs;

4. $\psi(X_o, Y_o)$ est strictement monotone croissant dans les outputs et strictement monotone décroissant dans les inputs, ce qui implique *ceteris paribus* que l'indicateur diminue si les ressources consommées augmentent ou que la production diminue, et que l'indicateur augmente si la consommation en ressources diminue ou que la production augmente;

5. $\psi(X_o, Y_o)$ est homogène de degré -1 dans les inputs ce qui implique que pour un facteur $\alpha \Rightarrow \psi(\alpha X_o, Y_o) = \frac{1}{\alpha}\psi(X_o, Y_o)$, si tous les inputs consommés augmentent (diminuent) d'une proportion donnée, l'indicateur d'efficacité diminue (augmente) dans la même proportion;

6. $\psi(X_o, Y_o)$ est homogène de degré 1 dans les outputs ce qui implique que pour un facteur $\beta \Rightarrow \psi(X_o, \beta Y_o) = \beta\psi(X_o, Y_o)$, si tous les outputs produits augmentent (diminuent) d'une proportion donnée, l'indicateur d'efficacité augmente (diminue) dans la même proportion;

7. Si $\psi(X_A, Y_A) > \psi(X_B, Y_B)$ pour deux organisations A et B, alors A est plus efficace que B.

6.3 Le choix de l'indicateur

À travers ce rapport de recherche, nous avons utilisé plusieurs modèles DEA, le modèle à rendements d'échelle constants CCR, le modèle à rendements d'échelle variables BCC, le modèle à ensemble non convexe FDH et quelques variantes de ceux-ci. Comme nous l'avons constaté dans notre revue de littérature, le coeur de la recherche et des applications empiriques DEA reposent sur les modèles BCC et CCR. À cet égard, il semble naturel de vouloir considérer ces modèles afin de construire l'indicateur d'efficacité désiré. Toutefois, nous démontrerons que plusieurs des propriétés mathématiques exposées ne sont pas respectées par ceux-ci, invalidant alors leur utilisation potentielle et suscitant de la sorte le besoin de recourir à un modèle alternatif nommé le modèle ERM.

6.3.1 Vérification des propriétés des modèles BCC et CCR

Débutons par la première propriété, soit la préférence que la mesure soit comprise entre 0 et 1. Les modèles CCR et BCC produisent les mesures θ et $1/\phi$ auxquelles nous avons si souvent fait référence. Lorsque les modèles sont formulés avec une orientation sur l'input, le θ est effectivement compris entre 0 et 1, ce qui est également le cas avec une orientation sur l'output où la mesure est $1/\phi$ avec $\phi \geq 1$. L'interprétation dans les deux cas se fait ainsi en terme de pourcentage. La mesure θ avec une orientation sur l'input s'interprète comme la proportion de la

consommation actuelle d'inputs qui devrait être utilisé afin d'atteindre l'efficacité. La mesure $1/\phi$ avec une orientation sur l'output s'interprète comme la proportion des outputs nécessaires à l'atteinte de l'efficacité qui sont actuellement produits. Ceci nous amène également à valider la propriété (3), soit l'invariance de la mesure aux unités des variables d'inputs et d'outputs.

Parmi les propriétés fondamentales d'un indicateur, nous avons noté sa capacité à mesurer le concept théorique sous-jacent. Ceci fait certainement intervenir la propriété (2), où l'atteinte du meilleur score est associée avec l'atteinte de l'efficacité. Les modèles CCR et BCC rencontrent un problème de taille en regard de cette propriété, c'est-à-dire que la mesure θ peut être égale à 1 sans que l'efficacité au sens de Pareto-Koopmans soit atteinte. Ce résultat provient du fait que le θ exclut les variables de slacks qui constituent pourtant un aspect essentiel de la définition d'efficacité. La mesure d'efficacité des modèles CCR et BCC est alors une mesure partielle de l'efficacité plutôt qu'une mesure globale en ne considérant que l'aspect radial du problème. Ainsi, bien qu'un établissement avec un θ égal à 1 soit situé sur la frontière, celui-ci peut se trouver sur une portion inefficace de cette frontière. La figure 18 permet d'en faire le constat.

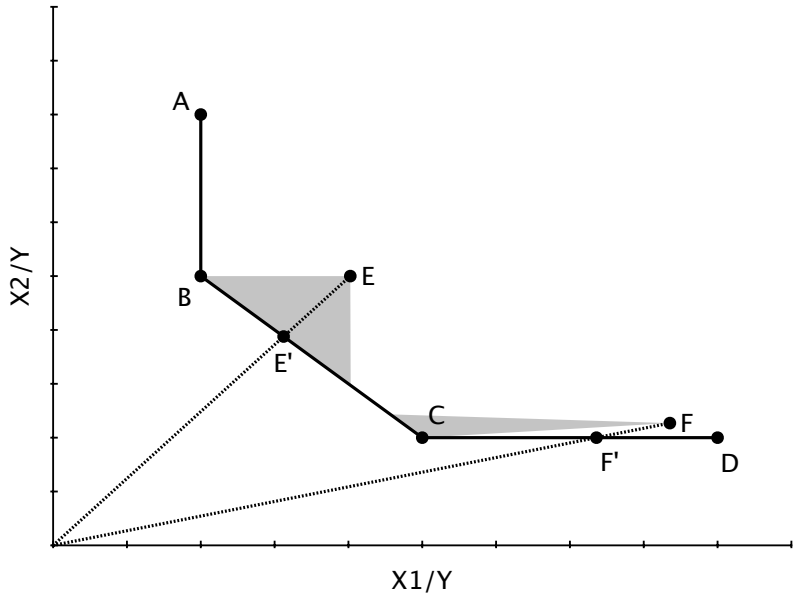

Fig. 18: Illustration des problèmes des modèles CCR et BCC

Les points A et D se situent sur la frontière, mais cela ne signifie pas qu'ils soient efficaces, puisqu'il existe toujours des inefficacités qui pourraient être éliminées.

L'organisation A pourrait réduire sa consommation de l'input x_2 et atteindre une performance efficace comme le point B, l'organisation D pourrait réduire sa consommation de l'input x_1 et atteindre le point C , qui lui est efficace.

De manière analogue, nous pouvons illustrer que la propriété (7) n'est pas respectée dans les modèles d'efficacité radiale. Pour ce faire, utilisons de nouveau la figure 18 et comparons les points E et F. Du point de vue de l'efficacité radiale, en ne s'intéressant qu'à la mesure θ, le point F est plus près de la frontière que le point E, nous avons alors $\theta_E < \theta_F$. Par ailleurs, il n'est pas évident que F soit plus efficace que E, car l'efficacité de F est mesurée par rapport à une portion inefficace de la frontière. Si nous regardons les possibilités (les zones ombragées) pour les deux points afin d'atteindre la portion efficace de la frontière donnée par le segment $\overline{BC}$, il est possible que la distance minimale entre le point E et la frontière soit plus petite que celle entre le point F et la frontière. Par conséquent, E pourrait être plus efficace que F, une réalité qui ne peut être traduite par la mesure θ.

Portons notre attention aux propriétés de monotonicité des modèles CCR et BCC. Nous désirons déterminer si les mesures produites sont monotones croissantes dans les outputs et monotones décroissantes dans les inputs.

Tentons d'abord d'étudier la question en fonction des inputs.

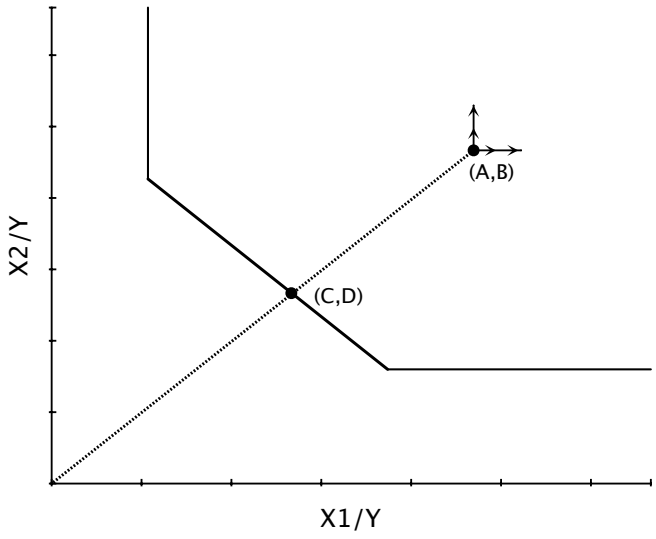

Fig. 19: Monotonicité des modèles CCR et BCC

La figure 19 montre l'évaluation d'un point (A, B) soit par BCC ou CCR, la question est de savoir si la mesure θ sera automatiquement plus petite, témoignant d'un niveau d'efficacité plus faible, si nous augmentons l'un ou l'autre des inputs.

Pour ce faire, nous cherchons à obtenir une expression pour le θ en fonction des points qui le définissent, ici il s'agit de (A, B) et (C, D). Nous avons donc besoin de l'équation du segment de la frontière soutenant le point (C, D) et l'équation du rayon soutenant le point (A, B):

$$i + pC = D \tag{74}$$

$$\frac{D}{C} = \frac{B}{A} \tag{75}$$

où A, B, C et D sont les coordonnées des points de la figure 19 et $i > 0, p < 0$ avec i l'intercept du segment de la frontière soutenant le point (C, D) et p sa pente.

Nous pouvons réexprimer le point (C, D) de la façon suivante:

$$\begin{aligned}
&C = \frac{A}{B}D \\
&et \\
&D = i + p\frac{A}{B}D \\
&\quad \Rightarrow i = D - p\frac{A}{B}D = (1 - p\frac{A}{B})D \\
&D = \frac{i}{1 - p\frac{A}{B}} \\
&\quad \Rightarrow C = \frac{i(A/B)}{1 - p(A/B)}
\end{aligned}$$

Nous pouvons maintenant calculer la mesure CCR ou BCC au point (A, B) comme:

$$\theta(A, B) = \frac{C}{A} = \frac{i}{B - pA}$$

Alors, le calcul des dérivées partielles en fonction des inputs suit facilement:

$$\frac{\partial\theta(A,B)}{\partial A} = \frac{ip}{(B-pA)^2} \leq 0$$

$$\frac{\partial\theta(A,B)}{\partial B} = \frac{-i}{(B-pA)^2} \leq 0$$

car nous avons $p < 0$. Nous avons donc fait la preuve que les modèles CCR et BCC sont monotones décroissants dans les inputs. La preuve pour les outpus suit de façon similaire et nous pouvons établir que les modèles CCR et BCC sont monotones croissants dans les outputs.

En dernier lieu, intéressons-nous aux propriétés d'homogénéité (5) et (6). Notons que nous ne vérifierons pas de manière aussi formelle cette propriété que la monotonicité puisque nous pouvons constater de façon relativement simple que les modèles CCR et BCC ne respectent pas les propriétés énoncées.

Depuis le début de ce rapport, nous avons supposé que tous les inputs et les outputs ne pouvaient prendre que des valeurs positives ou nulles. Pourtant, il est possible que certaines variables pertinentes à inclure dans le modèle DEA puissent prendre des valeurs négatives, nous n'avons qu'à penser à des variables de profits ou encore des indicateurs centrés à 0 dont l'étendue est $[-1, 1]$. Comme les modèles CCR et BCC ne peuvent traiter de telles variables directement, il est nécessaire de faire une transformation de manière à pouvoir les intégrer aux modèles. Regardons de plus près le problème que pose une telle façon de faire pour les propriétés d'homogénéité désirées des modèles.

À partir de la figure 20 de la page suivante, supposons un input x_2 pouvant prendre des valeurs négatives, positives ou nulles.

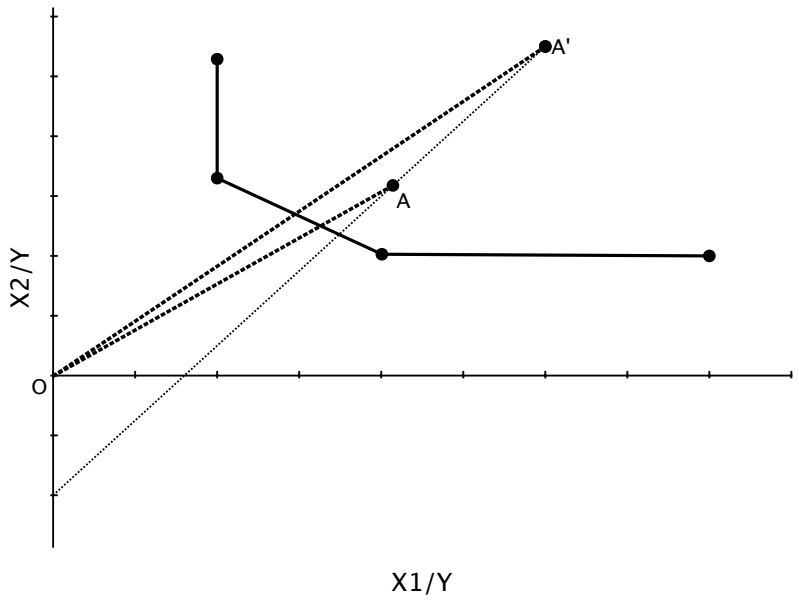

Fig. 20: Homogénéité des modèles CCR et BCC

Si nous évaluons l'efficacité d'un point A à l'aide de CCR ou BCC, nous mesurons le θ en fonction du segment $\overline{OA}$. Afin de regarder les propriétés d'homogénéité, nous augmentons la consommation des deux inputs d'une proportion γ, nous nous déplaçons alors sur le segment pointillé (le plus étroit) jusqu'au point A'. Pour mesurer l'efficacité au point A', nous devons maintenant prendre la mesure θ sur le segment $\overline{OA'}$.

En général, comme les mesure $\theta(A)$ et $\theta(A')$ sont prises sur des rayons différents, nous ne pourrons obtenir l'homogénéité de degré -1 qui était recherchée, nous aurons:

$$\theta(A') = \theta(\gamma A) \neq \frac{1}{\gamma}\theta(A)$$

Encore une fois, la preuve pour les outputs emprunte une logique similaire.

En définitive, nous venons de démontrer que plusieurs des propriétés souhaitables pour un indicateur de performance ne sont pas respectées par les modèles CCR et BCC. Ainsi, les résultats de ces modèles ne peuvent pas constituer des mesures intéressantes en tant qu'indicateur d'efficacité technique, et ce, principalement parce que les méthodes produisent des mesures incomplètes de l'efficacité.

Notre démonstration doit également mettre en perspective les conclusions des nombreuses études qui utilisent les scores d'efficacité des modèles CCR et BCC pour recommander la diminution de budget ou encore de ressources, des recommandations qui pourraient apparaître non seulement insuffisantes pour atteindre l'efficacité, mais aussi qui pourraient s'avérer foncièrement incorrectes puisqu'il existe tout un aspect de l'efficacité qui n'est pas considéré par ces modèles.

6.3.2 Présentation et vérification des propriétés du modèle ERM

Comme nous avons rejeté les modèles CCR et BCC en tant qu'indicateurs d'efficacité, la recherche d'un modèle alternatif qui serait fondé sur les principes de la méthode DEA constitue un objectif important de ce rapport de recherche. Parmi la multitude de modèles qui ont été développés, le modèle ERM proposé par J.T. Pastor, J.L. Ruiz et I. Sirvent en 1999 nous a semblé de loin être le plus satisfaisant. C'est donc vers ce modèle que nous nous tournons désormais. Nous exposerons celui-ci de manière générale, pour ensuite continuer en nous intéressant à ses propriétés mathématiques.

Le modèle ERM est formulé de la façon suivante[23]:

$$min_{\theta_i,\phi_r,\lambda_j} \quad \psi(X_o, Y_o) = \frac{\frac{1}{m}\sum_{i=1}^{m}\theta_i}{\frac{1}{s}\sum_{r=1}^{s}\phi_r} \tag{76}$$

$$\tag{77}$$

$$sujet\ à \qquad \sum_{j=1}^{n} x_{ij}\lambda_j \leq \theta_i x_{io} \quad i = 1, ..., m \tag{78}$$

$$\sum_{j=1}^{n} y_{rj}\lambda_j \geq \phi_r y_{ro} \quad r = 1, ..., s \tag{79}$$

$$\theta_i \leq 1,\ \phi_r \geq 1\ \forall i, r \tag{80}$$

$$\lambda_j \geq 0\ j = 1, ..., n. \tag{81}$$

[23] La formulation en (76) suppose des rendements d'échelle constants. L'ajout de la contrainte usuelle sur les mulitplicateurs lambdas, $\sum_{j=1}^{n}\lambda_j = 1$, permettrait de considérer des rendements d'échelle variables.

Le modèle ERM diffère des modèles DEA que nous avons présentés jusqu'à maintenant parce qu'il révèle l'ensemble des inefficacités qu'il est possible d'identifier. En effet, il considère tant les inefficacités sur les outputs que celles sur les inputs comme le montrent les contraintes (78) et (79). Il serait cependant facile d'imposer une contrainte sur le numérateur ou le dénominateur de sorte à orienter le modèle sur l'output ou sur l'intput comme le font les modèles CCR et BCC.

Un aspect important du modèle ERM est qu'il n'impose pas une unique réduction (augmentation) proportionnelle applicable à tous les inputs (outputs) comme le faisait le θ (ϕ) des modèles BCC et CCR. Plutôt, il permet que le mouvement de chacune des variables soit différent tel que le confirme les indices sur les mesures θ et ϕ. De cette façon, le modèle ERM choisit le chemin le plus court pour se rendre à la frontière d'efficacité et sa mesure représente le ratio de l'efficacité moyenne des inputs sur l'efficacité moyenne des outputs. Ainsi, la mesure ERM peut être scindée en deux composantes, soit l'efficacité moyenne dans la consommation d'inputs donnée par $\frac{1}{m}\sum_{i=1}^{m}\theta_i$ et l'efficacité moyenne dans la production d'outputs donnée par $\frac{1}{s}\sum_{r=1}^{s}\phi_r$.

Les mouvements définis par la réduction proportionnelle θ_i et ϕ_r peuvent être exprimés de la manière suivante:

$$\theta_i = \frac{x_{io} - s_{io}^-}{x_{io}} = 1 - \frac{s_{io}^-}{x_{io}} \quad i = 1, ..., m; \tag{82}$$

$$\phi_r = \frac{y_{ro} + s_{ro}^+}{y_{ro}} = 1 + \frac{s_{ro}^+}{y_{ro}} \quad r = 1, ..., s. \tag{83}$$

Le problème en (76) peut alors être réécrit de la façon qui suit:

$$min_{s_i^-, s_r^+, \lambda_j} \quad \psi(X_o, Y_o) = \frac{1 - \frac{1}{m}\sum_{i=1}^{m}\frac{s_{io}^-}{x_{io}}}{1 + \frac{1}{s}\sum_{r=1}^{s}\frac{s_{ro}^+}{y_{ro}}} \tag{84}$$

$$sujet\ à \qquad \sum_{j=1}^{n} x_{ij}\lambda_j = x_{io} - s_{io}^- \quad i = 1, ..., m \tag{85}$$

$$\sum_{j=1}^{n} y_{rj}\lambda_j = y_{ro} + s_{ro}^+ \quad r = 1, ..., s \tag{86}$$

$$s_{io}^-, s_{ro}^+ \geq 0\ \forall i, r \qquad \lambda_j \geq 0, j = 1, ..., n \tag{87}$$

Ce programme fractionnaire ne comporte pas de solution unique, cependant Pastor et al. (1999) définissent une transformation afin de linéariser le problème permettant ainsi d'obtenir la solution optimale au modèle ERM aussi facilement que les solutions aux modèles CCR et BCC.

Posons,

$$\beta = {}^{1}\!/_{(1+\frac{1}{s}\sum_{r=1}^{s}\frac{s_{ro}^{+}}{y_{ro}})} \tag{88}$$

$$t_{io}^{-} = \beta s_{io}^{-},\ i = 1, ..., m \tag{89}$$

$$t_{ro}^{+} = \beta s_{ro}^{+},\ r = 1, ..., s \tag{90}$$

$$\mu_j = \beta \lambda_j,\ j = 1, ..., n \tag{91}$$

Nous pouvons réexprimer (84) comme:

$$min_{t_i^-, t_r^+, \mu_j, \beta} \quad \beta - \frac{1}{m}\sum_{i=1}^{m}\frac{t_{io}^{-}}{x_{io}} \tag{92}$$

$$sujet\ à \qquad \beta + \frac{1}{s}\sum_{r=1}^{s}\frac{t_{ro}^{+}}{y_{ro}} = 1 \tag{93}$$

$$\sum x_{ij}\mu_j = \beta x_{io} - t_{io}^{-} \quad i = 1, ..., m \tag{94}$$

$$\sum_{j=1}^{n} y_{rj}\mu_j = \beta y_{ro} + t_{ro}^{+} \quad r = 1, ..., s \tag{95}$$

$$\beta \geq 0 \tag{96}$$

$$t_{io}^{-}, t_{ro}^{+} \geq 0\ \forall i, r \tag{97}$$

$$\mu_j \geq 0\ j = 1, ..., n. \tag{98}$$

N'importe quelle solution optimale pour (92) avec $\beta > 0$ produit également une solution optimale pour (84) par les changements de variables (88)-(91). Comme aucune solution avec $\beta = 0$ n'est réalisable pour le problème (92), si nous nous intéressons seulement à la mesure d'efficacité, nous pouvons simplement résoudre (92) sans passer ensuite par les transformations (88)-(91).

La figure 21 de la page suivante compare les chemins que prendront les modèles d'efficacité radiale et le modèle ERM pour projeter un point inefficace sur la frontière afin d'obtenir leur mesure respective d'efficacité. Le point E est projeté sur la frontière au même endroit par les modèles radiaux et le modèle ERM, alors

la mesure d'efficacité sera équivalente entre ceux-ci. Quant au point F, la mesure d'un modèle radial sera calculée à partir d'une projection sur la portion inefficace de la frontière, par conséquent cette mesure ne prendra pas en considération la réduction supplémentaire en x_1 qu'il serait possible d'effectuer afin d'éliminer toutes les inefficacités. Par ailleurs, le modèle ERM, parce qu'il n'est pas contraint à appliquer la même proportion de réduction à tous les inputs, sera capable de produire une mesure qui prenne en considération cette réduction supplémentaire. Généralement, nous aurons $\psi(X_o, Y_o) \leq \theta$.

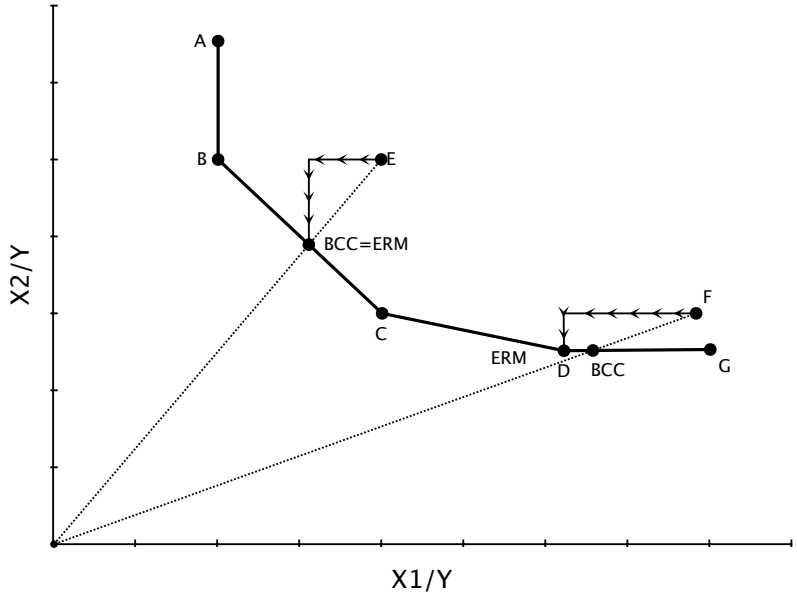

Fig. 21: Modèle ERM

Le modèle ERM produit donc cette mesure globale d'efficacité que nous ne retrouvions pas avec les mesures dérivées des modèles CCR et BCC. La mesure $\psi(X_o, Y_o)$ obtenue par ERM représente ainsi déjà un meilleur indicateur de l'efficacité, puisqu'elle vérifie les propriétés (2) et (7) que nous avons énoncées pour obtenir un indicateur valide et fiable.

Examinons maintenant si les autres propriétés identifiées sont respectées. Les propriétés (1) et (3) le sont naturellement en fonction de la définition du programme linéaire en (84), c'est-à-dire que l'indicateur est compris entre 0 et 1 et qu'il est invariant aux unités de mesure des variables.

Au niveau de la monotonie de l'indicateur, la propriété (4) précise qu'un indicateur devrait être strictement monotone croissant dans les outputs et strictement monotone décroissant dans les inputs. Pastor et al. (1999) font la preuve que la mesure donnée par ERM respecte cette condition:

Supposons deux DMUs, la DMU$_o$ avec le vecteur (X_o, Y_o) =$(x_{1o}, ...x_{ko}, ...x_{mo};\ y_{1o}, ..., y_{so})$ et la DMU$_a$ avec le vecteur (X_a, Y_a) =$(x_{1a}, ...x_{ka}, ...x_{ma};\ y_{1a}, ..., y_{sa})$ qui diffèrent dans la consommation d'un seul input, soit $x_{ka} = x_{ko} + a$ avec $a > 0$. Posons de plus, P_o et P_a comme le programme linéaire en (84) et $\psi(X_o, Y_o)$ et $\psi(X_a, Y_a)$ les solutions optimales de la DMU$_o$ et de la DMU$_a$.

À partir de la solution $(\hat{\lambda}_{1o}, ..., \hat{\lambda}_{no},\ \hat{s}^-_{1o}, ..., \hat{s}^-_{mo},\ \hat{s}^+_{1o}, ..., \hat{s}^+_{so})$ de la DMU$_o$, nous pouvons vérifier que $(\hat{\lambda}_{1a}, ..., \hat{\lambda}_{na},\ \hat{h}^-_{1a}, ..., \hat{h}^-_{ma},\ \hat{s}^+_{1a}, ..., \hat{s}^+_{sa})$ avec $\hat{h}_{ia} = \hat{s}^-_{io},\ i \neq k$ et $\hat{h}_{ka} = \hat{s}^-_{ko} + a$ est une solution réalisable de P_a. Dans ce cas, nous aurons $\psi(X_o, Y_o) > \psi(X_a, Y_a)$ montrant ainsi la monotonie décroissante dans les inputs. La preuve pour l'output suit alors de manière similaire.

En ce qui concerne les caractéristiques d'homogénéité du modèle ERM, utilisons le propramme en (76) pour vérifier si les propriétés (5) et (6) sont respectées. Supposons $\alpha > 1$ et $(\theta^*_i, \phi^*_r, \lambda^*_j)$ l'optimum pour une DMU$_o$. Nous pouvons vérifier que $(\theta^*_i/\alpha, \phi^*_r, \lambda^*_j)$ qui correspond à $1/\alpha\psi^*(X_o, Y_o)$ est une solution réalisable du problème (76) pour un vecteur $(\alpha X_o, Y_o)$ puisque toutes les contraintes demeurent satisfaites. Nous avons alors la relation qui suit:

$$\psi^*(\alpha X_o, Y_o) \leq \frac{1}{\alpha}\psi^*(X_o, Y_o) = \frac{\frac{1}{m}\sum_{i=1}^{m} \theta^*_i/\alpha}{\frac{1}{s}\sum_{r=1}^{s} \phi^*_r}$$

Cependant, une relation plus générale d'égalité entre les deux mesures ne peut pas être établie. Pastor et al. (1999) donnent également la preuve pour la relation entre $\psi(X_o, \beta Y_o)$ et $\beta\psi(X_o, Y_o)$ où nous trouvons :

$$\psi(X_o, \beta Y_o) \geq \beta\psi(X_o, Y_o)$$

Les propriétés d'homogénéité du modèle ERM ne sont donc pas exactement celles que nous aurions souhaitées. Toutefois, en les combinant avec la propriété de

la monotonicité qui elle, est respectée, nous nous assurons d'avoir un indicateur d'efficacité qui soit corrélé négativement avec la consommation d'inputs et positivement avec la production d'outputs. Si nous ajoutons à cela le fait que la mesure ERM soit une mesure d'efficacité technique globale, le modèle ERM devient alors valide afin de constituer l'indicateur que nous recherchons pour observer l'évolution de l'efficacité technique des établissements de soins de santé.

6.4 Comment obtenir un indicateur à partir d'un modèle DEA

Notre choix d'un indicateur d'efficacité s'arrête ainsi sur la mesure ERM en fonction des propriétés désirables que celle-ci possède. Il est toutefois nécessaire de préciser que les remarques que nous avons formulées en ce qui concerne les variables discrétionnaires et non discrétionnaires, l'inclusion de composantes pour tenir compte de la qualité et de l'environnement et les autres commentaires sur la complémentarité des ressources et les technologies restent toutes aussi pertinentes en regard de ce modèle et doivent rester des éléments centraux lors de la construction et de la résolution des programmes linéaires.

De plus, nous avons suggéré à la section 4.6.2 que le suivi de la performance des établissements devrait peut-être s'effectuer dans un cadre DEA statique plutôt que dynamique étant donné les contraintes d'investissement auxquelles sont confrontées les organisations du système de santé québécois. Effectivement, puisque le ministère de la Santé et des Services sociaux est l'acteur central en matière de production des soins de santé, les mouvements des établissements dans le temps sont conditionnés par les décisions ministérielles et ne reposent plus uniquement sur les méthodes de gestion et d'organisation des ressources que prennent ces derniers. Nous réitérons cette proposition une fois encore, maintenant que nous discutons de l'utilisation de la mesure ERM en tant qu'indicateur de performance.

Concrètement, lorsque nous nous intéressons à l'évolution de l'efficacité d'un établissement entre deux périodes, cela signifie que nous proposons de conserver la frontière de la première période lors du calcul de la mesure ERM pour la seconde période. Une telle comparaison permet d'examiner si un établissement donné s'est amélioré au niveau de l'efficacité entre deux points dans le temps, car de cette manière, nous pouvons comparer les mesures $\psi(X_o, Y_o)$ des deux périodes

sans tenir compte des mouvements des autres organisations. Ce n'est qu'une fois que l'indicateur atteint son maximum qu'il sera alors utile d'établir une nouvelle frontière.

Nous pouvons également mettre en valeur un aspect de la mesure ERM qui en fait un indicateur de performance utile dans un contexte où les établissements de santé opèrent dans le cadre d'objectifs prescrits par le Ministère. En effet, contrairement aux modèles CCR et BCC, la forme fractionnaire du modèle ERM en (76) ou en (84) impose une pondération équivalente à tous les inputs et tous les outputs. Cela signifie alors que l'impact marginal de chacune des variables sur la mesure d'efficacité est d'égale importance. Cette absence de poids ou plutôt leur neutralité fait en sorte qu'il est possible pour l'analyste de modifier et d'imposer lui-même une structure de poids reflétant certaines des priorités organisationnelles et ministérielles. En accordant ainsi, une plus haute importance relative à certains inputs ou certains outputs, le niveau de l'indicateur d'efficacité pourrait refléter la bonne ou moins bonne performance des établissements sur des aspects établis prioritaires de la production de soins. Nous devons toutefois nous assurer que les pondérations se somment à 1. Par exemple, l'objectif à minimiser en (84) pourrait ressembler à ceci, où les pondérations sont données par les variables α et ω :

$$\psi(X_o, Y_o, \alpha, \omega) = \frac{1 - \frac{1}{m}\sum_{i=1}^{m} \alpha_i \frac{s_{io}^-}{x_{io}}}{1 + \frac{1}{s}\sum_{r=1}^{s} \omega_r \frac{s_{ro}^+}{y_{ro}}}$$

$$avec \quad \sum_{i=1}^{m} \alpha_i = 1$$

$$\sum_{r=1}^{s} \omega_r = 1$$

Notons d'ailleurs que lorsqu'il s'agit d'intégrer un quelconque jugement de valeur, cette façon de faire semble plus naturelle que la méthode qui consiste à imposer des restrictions sur les poids de la forme fractionnaire des modèles CCR et BCC étant donné que ces poids sont les variables sur lesquelles les problèmes optimisent.

Concluons, finalement, notre discussion sur la mesure ERM et sa pertinence en tant qu'indicateur de performance en mentionnant que ce modèle ne réfute pas les principes initiaux de la méthode DEA telle l'utilisation d'un ensemble

d'observations pour inférer les possibilités de production et la comparaison entre différents établissements procurant des services de même nature pour établir le niveau des meilleures performances. Plutôt, le modèle ERM s'éloigne de la tradition DEA instauré par les modèles CCR et BCC au niveau des résultats, dans le sens où les mesures produites reflétent sans réserve la notion d'efficacité relative de Pareto-Koopmans qui a été à la source de la technique.

7 Conclusions

L'objectif de ce rapport de recherche était de mener une discussion sur la mesure de l'efficacité d'un système de soins de santé à savoir comment nous pouvons construire une mesure d'efficacité et en quoi les hypothèses sous-jacentes à la méthode utilisée influencent les résultats que nous obtenons. Dans cette optique, nous nous sommes concentrés sur la méthode DEA afin d'étudier en profondeur la façon dont celle-ci définit, mesure et évalue l'efficacité. De plus, comme le titre de ce rapport l'indique, nous étions également intéressés à attester de la portée et des limites de la méthode DEA en nous interrogeant sur le sens des mesures obtenues et sur la manière dont nous pourrions en faire usage dans une analyse d'efficacité en santé.

D'emblée, soulignons que la méthode DEA couvre un spectre théorique très large. Si le modèle CCR apparaît un peu trop contraignant, nous avons cependant montré quelles sont les possibilités pour complexifier l'analyse en ajustant le modèle pour tenir compte de rendements d'échelle variables et pour y intégrer la notion de qualité, des variables non discrétionnaires et des composantes aléatoires. Notons de plus que le cadre DEA s'adapte particulièrement bien à une analyse temporelle de l'efficacité. Enfin, nous avons présenté dans ce rapport de recherche les modèles qui constituent ce que nous appelons le coeur de la méthode DEA, bien que nous avons omis certaines des extensions possibles, il demeure que les résultats et les conclusions auxquels nous parvenons ne s'en trouvent pas altérés.

Nous avons cherché à mettre de l'avant les hypothèses de la méthode pour ensuite montrer comment certaines considérations pratiques peuvent commander des modifications parfois subtiles et parfois considérables. Destinées à relâcher des hypothèses exigeantes du modèle initial, ces extensions théoriques ne s'éloignent

pas pour autant complètement du cadre strict DEA au sein duquel plusieurs suppositions s'imposent en force. À cet égard, nous avons noté la difficulté d'intégrer des notions de complémentarité que ce soit au niveau des facteurs de production ou au niveau du lien entre la qualité et l'output et la difficulté de relâcher la substituabilité entre les inputs.

À travers cette discussion sur les hypothèses des modèles DEA, nous avons surtout aussi traité de leur pertinence dans le cadre d'un système de soins de santé ce qui nous permet alors de tirer quelques conclusions sur la manière dont une analyse DEA dans le domaine de la santé se doit d'être menée.

Premièrement, il semble que le contexte dans lequel opèrent différents établissements de soins de santé soit le facteur le plus important à comprendre pour construire un modèle DEA valide. Ce contexte est essentiel parce qu'il influence ce qui est réalisable, c'est-à-dire qu'il détermine la forme de l'ensemble des possibilités de production. En ce sens, il demeure nécessaire de comprendre ce que les établissements ont la possibilité de faire, une compréhension qui doit s'articuler à travers plusieurs questions. Quelles sont les technologies disponibles? Quelles décisions sont sous le contrôle des gestionnaires? Comment la demande de soins influence les soins qui seront produits? Quels sont les effets sur l'accessibilité et la qualité des soins de la quantité de soins dispensés?

Pour intégrer les réponses de ces diverses interrogations à l'analyse DEA nous devons adopter certaines façons de faire. Dans un premier temps, il est nécessaire de contrôler pour la demande de soins qui est dérivée de la structure de la population, afin de ne comparer que des établissements qui sont susceptibles d'avoir à produire le même type de soins en des quantités comparables. Ajoutons que l'utilisation d'inputs non discrétionnaires semble être la meilleure façon d'effectuer ce contrôle. Les études empiriques ont également démontré qu'il était aussi important de contrôler les soins prodigués en fonction du niveau de risque des patients.

Ensuite, comme les établissements de santé sont intégrés dans un cadre plus large qui est le système de santé, nous devons considérer le mode de fonctionnement de ce système et ses paliers décisionnels pour décider de ce que les gestionnaires ont la possibilité de réaliser. Dans le contexte québécois qui nous intéresse principalement, nous avons montré que plusieurs décisions, notamment à propos de l'allocation des ressources humaines et financières, relèvent du ministère de la

Santé et des Services sociaux. Cela dit, une analyse DEA doit inclure ce genre de contraintes à la libre disposition des ressources en utilisant des variables non discrétionnaires.

En plus, le caractère particulier de la production de soins de santé et la notion de service public y étant attachée font en sorte que nous ne pouvons pas dissocier l'aspect qualitatif de l'efficacité. Autant l'accessibilité des soins que leur efficience doivent être considéré comme des aspects complémentaires à la production. Ceci demande de recourir à des modèles DEA de congestion qui sont les seuls modèles que nous avons identifiés pouvant traiter la qualité en ce sens.

Au terme de ce travail, nous pouvons donc présenter de manière assez concise la façon de mener une analyse d'efficacité en santé en insistant sur les aspects qui ont de l'importance dans de ce domaine. Cependant, il semble que les considérations importantes que nous soulignons ne sont que partiellement, parfois même aucunement abordées par les études empiriques, la pratique demeurant plutôt la négligence de ces diverses perspectives que nous avons soulevées. Que l'objet de ces études soit la mesure du niveau d'efficacité ou la comparaison entre des établissements, le traitement des résultats obtenus par DEA est surtout mis en valeur dans l'optique d'obtenir une mesure spécifique de l'efficacité et d'y faire suivre de multiples prescriptions en terme d'allocation des ressources et de budgétisation.

Pourtant, nous le répétons, la méthode DEA offre un cadre d'analyse de l'efficacité qui est rigide d'un certain point de vue puisque les hypothèses sont fortement structurantes. Le jugement de l'analyste et sa réflexion sur le contexte précis dans lequel il mène une analyse d'efficacité doit occuper une place centrale et non substituable afin d'obtenir des résultats qui ont du sens. Nous avons mentionné qu'en vue d'évaluer les composantes d'un système de soins de santé, des analyses sur les différents types de complémentarité des ressources ou encore sur les rendements d'échelle sont nécessaires avant d'entreprendre une analyse DEA, ceci dans le but de construire un modèle valide. Sans ces réflexions préalables, la mesure θ obtenue ne peut s'interpréter comme une mesure de l'efficacité, du moins de la manière dont en font usage les études empiriques en tant que niveau d'efficacité.

D'ailleurs, nous avons fait la preuve que les modèles DEA les plus communs, les modèles CCR et BCC, ne sont même pas dotés des propriétés d'homogénéité du premier degré et que la mesure θ rend compte de l'efficacité de façon incomplète,

ce qui rend difficile, voire incorrect, d'utiliser les résultats DEA en tant que mesure spécifique de l'efficacité.

En fait, nous concluons que l'interprétation de la mesure DEA doit se faire plutôt comme celle d'un indicateur d'efficacité technique, c'est-à-dire qu'elle permet de suivre l'évolution des établissements de santé vers l'atteinte de l'efficacité. À cette fin, nous avons proposé un modèle qui reste peu utilisé, le modèle ERM, mais qui respecte pourtant de nombreux critères de validité contrairement aux modèles CCR et BCC. Notons aussi la principale force de la mesure ERM qui réside dans son habileté à détecter l'ensemble des sources d'inefficacité, un aspect qui échappe aux modèles standards.

Concevoir la mesure obtenue comme un indicateur préférablement à une mesure d'efficacité est opportun en regard de l'incertitude entourant l'estimation d'un modèle DEA, à savoir si les variables pertinentes ont toutes été incluses, et cela, de manière exhaustive et exclusive. Ainsi, contrairement à ceux qui suggèrent l'estimation de plusieurs modèles de façon à statuer de la constance des résultats, nous pensons que les analyses préalables et un positionnement réfléchi sur les variables à inclure sont des pratiques suffisantes afin d'obtenir un indicateur à l'aide de la méthode DEA.

En ce qui concerne l'indicateur proposé, mentionnons divers usages que nous pourrions en faire. D'abord, du point de vue des établissements de soins, la possibilité de disposer d'un indicateur d'efficacité peut constituer un outil de gestion adéquat afin de les éclairer sur la valeur de certains des choix qu'ils ont à faire et de prendre des décisions qui vont dans le sens de l'amélioration de l'efficacité. Le cheminement jusqu'à l'atteinte de l'efficacité serait ainsi guidé par un indicateur dont nous savons qu'il révèle véritablement les améliorations et les détériorations de la performance.

Au niveau agrégé, l'indicateur d'efficacité que nous proposons peut également contribuer à établir, pour le compte des agences régionales, un facteur d'évaluation annuelle des établissements. Un critère d'évaluation qui ne se destine pas à déterminer la part du budget devant être retranchée ou encore à imposer des limites sur la consommation de certaines ressources, mais davantage à donner une indication du progrès réalisé sur l'efficacité par les organisations. Au niveau du Ministère, l'indicateur pourrait orienter les politiques en permettant de simuler les gains en efficacité des établissements sous différents scénarios notamment en

ce qui concerne l'attribution de nouveaux équipements et la façon de structurer le système. Dans le contexte où les demandes d'imputabilité sont croissantes, tant de la part des élus que de la part de la population, nous croyons que l'indicateur DEA s'avère un outil qui ne peut être négligé.

Somme toute, la contribution de notre travail aura été de préciser les aspects fondamentaux associés à la production de soins de santé et de statuer de leur importance lors d'une analyse d'efficacité. Nous aurons également mis en valeur comment certains problèmes se posent lors de l'application d'un modèle DEA et comment ceux-ci peuvent être résolu tant à l'intérieur du cadre de la méthode que de façon parallèle. Par-dessus tout, nous aurons proposé l'utilisation d'un modèle alternatif aux modèles réguliers, reflétant mieux la notion d'efficacité posée par Pareto et Koopmans. Finalement, nous aurons souligné que la portée de la méthode DEA doit se limiter à une interprétation en terme d'indicateur d'efficacité et non pas à cette mesure exacte à laquelle les études empiriques accordent tant d'importance.

Enfin, rappelons que dans ce rapport de recherche, nous avons opté pour une définition technique de l'efficacité. Ce choix étant dicté par un point de vue orienté sur les établissements, nous avons, de la sorte, écarté la notion allocative. Néanmoins, les modèles DEA permettent de mener une analyse d'efficacité allocative. Il pourrait alors être pertinent de s'attarder au genre de réflexion que nous avons proposé dans ce travail en fonction de la dimension allocative du système de santé québécois en adoptant un point de vue plus général de l'efficacité. Une discussion sur le problème allocatif du système de santé permettrait ensuite de s'attarder aux questions de son financement.

Finalement et plus généralement, si nous insistons sur l'utilité des analyses de productivité telles que proposées par les modèles DEA afin de comprendre des phénomènes complexes comme la production de soins de santé, nous devons également y voir que ces modèles ne représentent pas une panacée. S'ils peuvent effectivement éclairer les gestionnaires sur les décisions favorisant l'efficacité des soins, l'évaluation de l'efficacité et son amélioration ne sauraient reposer uniquement sur de tels indicateurs. Des objectifs politiques non partisans en termes de santé publique doivent également encadrer les analyses techniques en prenant en compte certains aspects comme les progrès technologiques qui font en sorte que nous sommes maintenant capables de traiter des maladies en réduisant le temps

de séjour dans les établissements, les transitions épidémiologiques dans le sens où la population développe des maladies chroniques et de longue durée plutôt que des maladies infectieuses et épidémiques et la transition démographique en particulier le vieillissement de la population. Ce n'est qu'en intégrant les indicateurs d'efficacité à ces dimensions que de nouvelles méthodes de production de soins plus efficaces pourront être développées et que nous pourrons innover sur la forme et dans la nature des services offerts en matière de santé à la population.

8 Bibliographie

1. Agrell, P.J. et P. Bogetoft. (2001). *DEA-Based Regulation of Health Care Systems*. Seventh European Workshop on Efficiency and Productive Analysis. Oviedo (Espagne).

2. Aigner, D.J., C.A.K. Lovell et P. Schmidt. (1977). «Formulation and Estimation of Stochastic Frontier Production Function Models». *Journal of Econometrics*. 6. 21-37.

3. Allen, R.A., A. Athanassopoulos, R. G. Dyson et E. Thanassoulis. (1997). «Weights restrictions and value judgements in data envelopment analysis: evolution, development and future directions». *Annals of Operations Research*. 73. 13–34.

4. Anderson, R.I., H. S. Weeks, B.K. Hobbs et J. R. Webb. (2003). «Nursing Home Quality, Chain Affiliation, Profit Status and Performance». *Journal of Real Estate Research*. 25(1). 43-60.

5. Badillo, P.Y. et J.C. Paradi. (1999). *La méthode DEA: Analyses des performances*. Hermès Science Publications. Paris.

6. Banker, R.D. (1984). «Estimating Most Productive Scale Size Using Data Envelopment Analysis». *European Journal of Operational Research*. 17. 35-44.

7. Banker, R.D., A. Charnes et W.W. Cooper. (1984). «Some Models for Estimating Technical and Scale Inefficiencies in Data Envelopment Analysis». *Management Science*. 30(9). 1078-92.

8. Banker, R.D., R.F. Conrad et R.P. Strauss. (1986). «A Comparative Application of Data Envelopment Analysis and Translog Methods: An Illustrative Study of Hospital Production». *Management Science*. 32(1). 30-44.

9. Banker, R. D. et R.C. Morey. (1986),.« The Use of Categorical Variables in Data Envelopment Analysis ». *Management Science*. 32(12). 1613-27.

10. Banker, R.D. et R.M. Thrall. (1992). «Estimation of Returns to Scale Using Data Envelopment Analysis». *European Journal of Operational Research.* 62. 74-84.

11. Banker, R.D., I. Bardhan et W.W. Cooper. (1996). «A Note on Returns to Scale in DEA». *European Journal of Operational Research.* 88(3). 585-85.

12. Bardey, D. et S. Pichetti. (2004). «Estimation de l'efficience des dépenses de santé au niveau départemental par la méthode DEA». *Économie et Prévision*. 5(166). 59-69.

13. Bates, L.J., K. Mukherjee et R.E. Santerre. (2006). «Market Structure and Technical Efficiency in the Hospital Services Industry: A DEA Approach». *Medical Care Research and Review*. 63 (4). 499-524.

14. Bilodeau, D., P.Y. Crémieux, B. Jaumard, P. Ouellette et T. Vovor. (2004). «Measuring Hospital Performance in the Presence of Quasi-Fixed Inputs: An Analysis of Qubec Hospitals». *Journal of Productivity Analysis*. 21. 183-99.

15. Butler, T.W. et L. Li. (2005). «The Utility of Returns to Scale in DEA Programming: An Analysis of Michigan Rural Hospitals». *European Journal of Operational Research*. 161(2). 469-77.

16. Campbell, S.M., M.O. Roland et S.A. Buetow. (2000). «Defining Quality of Care». *Social Science and Medicine*. 51. 1611-25.

17. Champagne, F., A.P. Contandriopoulos, J. Picot-Touché, F. Béland et H. Nguyen. (2005). «Un cadre d'évaluation de la performance des systèmes de services de santé: Le modèle EGIPSS». Résumé du rapport technique. Groupe de Recherche Interdisciplinaire en Santé. N05-02. Université de Montréal.

18. Charnes, A., W.W. Cooper et E. Rhodes. (1978). «Measuring the Efficiency of Decicion Making Units». *European Journal of Operational Research*. 2. 429-44.

19. Chattopadhay, C. et S.C. Ray. (1996). «Technical, Scale and Size Efficiency in Nursing Home Care: A Nonparametric Analysis of Connecticut Homes». *Health Economics*. 5. 363-73.

20. Chilingerian, J.A. (1995). «Evaluating Physician Efficiency in Hospitals: A Multivariate Analysis of Best Practices». *European Journal of Operational Research*. 80. 548-74.

21. Clement, J.P., V.G. Valdmanis, G.J. Bazzoli, M. Zhao et A. Chukmaitov. (2008). «Is More Better? An Analysis of Hospital Outcomes and Efficiency with a DEA Model of Output Congestion». *Health Care Management Science*. 11. 67-77.

22. Coelli, T.J., D.S. Prasada Rao, C.J. O'Donnell et G. Battese. (2005). *An Introduction to Efficiency and Productivity Analysis*. 2e édition. Springer.

23. Cooper, W.W., L.M. Seiford et K. Tone. (2007). *Data Envelopment Analysis A Comprehensive Text with Models, Applications, References and DEA-Solver Software*. Springer. New York.

24. Cooper, W.W., L.M. Seiford et J. Zhu. (2004). *Handbook on Data Envelopment Analysis*. Kluwer Academic Publishers. Boston.

25. Crémieux, P.Y., D. Gadbois, B. Jaumard, P. Ouellette et T. Vovor. (2001). «Évaluation de l'efficience budgétaire des CLSC au Québec à partir de la méthode DEA». *L'actualité Économique*. 77(3). 409-24.

26. Deprins, D., L. Simar et H. Tulkens. (1984). «Measuring Labor Efficiency in Post Offices». Dans M. Marchand, P. Pestieau et H. Tulkens (dir.). *The Performance of Public Enterprises: Concepts and Measurements*. Amsterdam. North Holland.

27. Färe, R. et L. Svensson. (1980). «Congestion of production factors». *Econometrica*. 48(7). 1745-53.

28. Färe, R., S. Grosskopf, CAK. Lovell CAK et C. Pasurka. (1989) «Multilateral productivity comparisons when some outputs are undesirable: a nonparametric approach». *Review of Economics and Statistics*. 71.90–98.

29. Färe, R.S., S. Grosskopf, B. Lindgren et J.P. Poullier. (1997). «Productivity Growth in Health-Care Delivery». *Medical Care*. 35(4). 354-66.

30. Farrell, M.J. (1957). «The Measurement of Productive Efficiency». *Journal of the Royal Statistical Society*. 120(3). 253-90.

31. Ferrier, G.D., M. Rosko et V.G. Valdmanis. (2006). «Analysis of Uncompensated Hospital Care Using a DEA Model of Output Congestion». *Health Care Management Science*. 9. 181-88.

32. Fizel, J.L. et T. Nunnikhoven. (1992). «Technical Efficiency of For-profit and Non-profit Nursing Homes». *Managerial and Decision Economics*. 13. 429-39.

33. Førsund, F.R., C.A.K. Lovell et P. Schmidt. (1980). «A Survey of Frontier Production Functions and of Their Relationship to Efficiency Measurement». *Journal of Econometrics*. 13. 5-25.

34. Førsund, F. R. et N. Sarafoglou. (2002). «On the Origins of Data Envelopment Analysis». *Journal of Productivity Analysis*. 17. 23-40.

35. Fried, H.O., C.A.K. Lovell, et al. (2002). «Accounting for Environmental Effects and Statistical Noise in Data Envelopment Analysis». *Journal of Productivity Analysis*. 17. 157-74.

36. Garcia, F., C. Marcuello, D. Serrano et O. Urbina. (1999). «Evaluation of Efficiency in Primary Health Care Centers: An Application of Data Envelopment Analysis». *Financial Accountability and Management*. 15(1). 67-83.

37. Grosskopf, S. et V. Valdmanis. (1987). «Measuring Hospital Performance: A Non-Parametric Approach». *Journal of Health Economics*. 6. 89-107.

38. Harrison, J.P., M.N. Coppola et M. Wakefield. (2004). «Efficiency of Federal Hospitals in the United States». *Journal of Medical Systems*. 28(5). 411-22.

39. Helmig, B. et I. Lapsley. (2001). «On the Efficiency of Public, Welfare and Private Hospitals in Germany Over Time: A Sectoral Data Envelopment Analysis Study». *Health Services Management Research*. 14. 263-74.

40. Hollingsworth, B. (2008) «The Measurement of Efficiency and Productivity of Health Care Delivery». *Health Economics*. 17. 1107-28.

41. Jacobs, Rowena, Smith, Peter C., Street, Andrew. (2006). *Measuring Efficiency in Health Care: Analytics Techniques and Health Policy.* Cambridge University Press. New York.

42. Kirigia, J.M., A. Emrouznejad, L.G. Sambo, N. Mungunti et W. Liambila. (2004). «Using Date Envelopment Analysis to Measure the Technical Efficiency of Public Health Centers in Kenya». *Journal of Medical Systems.* 28(2). 155-66.

43. Klimberg, R.K., et M. Puddicombe. (1999). «A multiple objective approach to data envelopment analysis». *Advances in Mathematical Programming and Financial Planning*. 5. 201–32.

44. Kneip, A., L. Simar et P. W. Wilson. (2008). «Asymptotics and Consistent Bootstraps for DEA Estimators in Non-Parametric Frontier Models». *Econometric Theory*. 24. 1663-97.

45. Kontodimopoulos, N., T. Bellali, G. Labiris et D. Niakas. (2006). «Investigating Sources of Inefficiency in Residential Mental Health Facilities». *Journal of Medical System.* 30. 169-76.

46. Kooreman, P. (1994). «Nursing Home Care in The Netherlands: A Nonparametric Efficiency Analysis». *Journal of Health Economics*. 13. 301-16.

47. Land, C.K., C.A.K. Lovell et S. Thore. (1993). «Chance-constrained data envelopment analysis». *Managerial and Decision Economics.* 14. 541-54.

48. Liu, C., B. Ferguson et A. Laporte. (2006). «Ranking the Health System Efficiency among Canadian Provinces and American States». Document de travail. Université de Guelph.

49. Luoma, K., M. L. Järvio, et al. (1996). «Financial Incentives and Productive Efficiency in Finnish Health Centers». *Health Economics*. 5. 435-45.

50. Maniadakis, N. et E. Thanassoulis. (2000). «Assessing Productivity Changes in UK Hospitals Reflecting Technology and Input Prices». *Applied Economics*. 32. 1575-89.

51. Martinussen, P.E. et L. Midttun. (2004). «Day Surgery and Hospital Efficiency: Empirical Analysis of Norwegian Hospitals, 1999-2001». *Health Policy*. 68. 183-96.

52. Mas-Colell, A., Michael D. Whinston et Jerry R. Green. (1995). *Microeconomic Theory*. Oxford University Press. New York.

53. Meeusen, W. et J. van den Broeck. (1977). «Efficiency Estimation from Cobb-Douglas Production Functions With Composed Error». *International Economic Review*. 18. 435-44.

54. Nayar, P. et A.Y. Ozcan. (2008). «Data Envelopment Analysis Comparison of Hospital Efficiency and Quality». *Journal of Medical System*. 32. 193-99.

55. Nunamaker, T.R. (1983).«Measuring Routine Nursing Service Efficiency: A Comparison of Cost per Patient Day and Data Envelopment Analysis Models». *Health Services Research*. 18(2). 183-205.

56. Ozcan, Y.A. (2008). *Health Care Benchmarking and Performance Evaluation*. Springer. Newton.

57. Parkin, D. et B. Hollingsworth. (1997). «Measurinf Production Efficiency of Acute Hospitals in Scotlan, 1991-94: Validity Issues in Data Envelopment Analysis». *Applied Economics*. 29. 1425-1433.

58. Pastor, J.T., J.L. Ruiz et I. Sirvent. (1999). «An Enhanced DEA Russel Graph Efficiency Measure». *European Journal of Operational Research*. 115. 596-607.

59. Pina, V, et L. Torres. (1992). «Evaluating the Efficiency of Nonprofit Organizations: An Application of Data Envelopment Analysis to the Public Health Service». *Financial Accountability and Management*. 8(3). 213-24.

60. Puig-Junoy, J. (1998). «Technical Efficiency in the Clinical Management of Critically Ill Patients». *Health Economics*. 7. 263-77.

61. Québec. Loi sur la Santé et les Services sociaux. L.R.Q. chapitre S-4.2.

62. Québec. Secrétariat du Conseil du Trésor. «Budget et Dépenses 2009-2010». En ligne. http://www.tresor.gouv.qc.ca/fr/publications/budget/09-10/graphiques_FR.pdf (page consultée le 9 août 2009).

63. Ramanathan, R. (2003). *An Introduction to Data Envelopment Analysis: A Tool for Performance Measurement*. Sage Publications. New Delhi.

64. Rollins, J., K. Lee, Y. Xhu et A. Ozcan. (2001). «Longitudinal Study of Health Maintenance Organization Efficiency». *Health Services Management Research.* 14. 249-62.

65. Roseman, R., K. Siddhartan et M. Ahern. (1997). «Output Efficiency of Health Maintenance Organizations in Florida». *Health Economics*. 6. 295-302.

66. Salinas-Jiménez, J. et P. Smith. (1996). «Data Envelopment Analysis Applied to Quality in Primary Health Care». *Annals of Operations Research*. 67. 141-61.

67. Sherman, H.D. et J. Zhu. (2006). «Benchmarking with Quality-Adjusted DEA (q-dea) to Seek Lower-Cost High-Quality Service: Evidence from a U.S. Bank Application». *Annals of Operations Research.* 145. 301–19.

68. Shimshak, D.G., M. L. Lenard et R.K. Klimberg. (2009). «Incorporating Quality into Data Envelopment Analysis of Nursing Home Performance: A Case Study». *Omega The International Journal of Management Science*. 37, 672-85.

69. Simar, L. (1992). «Estimating efficiencies from frontier models with panel data: A comparison of parametric, non-parametric and semi-parametric methods with bootstrapping». *Journal of Productivity Analysis*. 3. 167–203.

70. Simar, L., et P. W. Wilson. (1998a). «Sensitivity Analysis of Efficiency Scores: How to Bootstrap in Nonparametric Frontier Models». Management Science 44(11), 49–61.

71. Simar, L., et P. W. Wilson. (1998b). «Nonparametric Tests of Returns to Scale, Discussion paper #9814». Institut de Statistique. Université Catholique de Louvain, Louvain-la-Neuve, Belgique.

72. Simar, L. et P.W. Wilson. (2007). «Estimation and Inference in Two-Stage, Semi-Parametric Models of Production Processes». *Journal of Econometrics*. 136. 31-64.

73. Spinks, J. et B. Hollingsworth. (2009). «Cross-country Comparisons of Technical Efficiency of Health Production: A Demonstration of Pitfalls». *Applied Economics*. 41. 417-27.

74. Tone, K. (1996). «A Simple Characterization of Returns to Scale in DEA». *Journal of the Operations Research Society of Japan.* 39. 604-613.

75. Tulkens, H. et P. Vanden Eeckaut. (1999). *Mesurer l'efficacité: avec ou sans frontières?* In: La Méthode DEA: Analyse des Performances. ed: Patrick-Yves Badillo et Joseph C. Paradi. Hermes Science Publications. Paris.

76. Turgeon, J., H. Anctil et J. Gauthier. (2003). «L'évolution du Ministère et du réseau: continuité ou rupture». dans V. Lemieux, P. Bergeron, C. Bégin et Gé Bélanger, dir. *Le système de santé au Québec: Organisations, Acteurs et Enjeux*. pp.93-117.

77. Valdmanis, V. (1992). «Sensitivity Analysis for DEA Models: An Empirical Exemple Using Public vs. NFP Hospitals». *Journal of Public Economics*. 48. 185-205.

78. Valdmanis, V., M.D. Rosko et R.L. Mutter. (2008).«Hospital Quality, Efficiency, and Input Slack Differentials». *Health Services Research*. 45(5). 1830-1848.