

# Tarification des hôpitaux : la prise en compte des hétérogénéités \*

Brigitte Dormont et Carine Milcent

2 février 2005

Mots clés : coûts hospitaliers, régulation hospitalière, hasard moral, panels non cylindrés.

*Classification JEL* : C3, D8, L5

---

\*Correspondance à : Théma, Université Paris X, Bâtiment G, 200 avenue de la république 92001 Nanterre cedex, France. Tel : 00 33 1 40 97 78 36. e-mail : dormont@u-paris10.fr ou milcent@u-paris10.fr

<sup>0</sup>Cette étude a été financée en partie par une convention avec la DREES (Ministère de l'emploi et de la solidarité). Nous remercions Werner Antweiler (University of British Columbia) pour ses précieux commentaires, ainsi que Bruno Crépon, Alberto Holly, Nathalie Destais et les deux rapporteurs anonymes de la revue. Nous avons aussi profité des remarques des participants du séminaire santé CREST-LEI, du séminaire Théma, du séminaire Fourgeaud et du Summer Workshop du NBER (Boston, 2002).

# 1 Introduction

Tous les acteurs du système hospitalier se plaignent de la pénurie des moyens à leur disposition. Dans le même temps, un consensus se dégage pour constater que l'activité hospitalière se caractérise par des inégalités de dotations budgétaires et de grandes inefficacités (Mougeot (1999<sub>a</sub>)). Une réforme de la tarification hospitalière est à l'ordre du jour, avec l'expérimentation d'un système de tarification par pathologie. L'objet de cet article est d'étudier, dans cette perspective, la variabilité des coûts des hôpitaux publics français. Nos applications empiriques sont réalisées sur des données stratifiées à trois dimensions hôpital-séjour-année concernant des patients traités pour infarctus du myocarde aigu. Notre approche consiste à essayer d'identifier les inefficacités repérables, à définir des formules de paiement correspondant à des coûts de traitement efficace, puis à effectuer des simulations permettant d'évaluer les gains budgétaires qui pourraient être associés à l'application d'une telle tarification.

Cet article est organisé de la façon suivante. Dans la section 2 on fournit des éléments sur la tarification hospitalière, avec une mise en perspective à la fois théorique et historique, en évoquant les expériences française et américaine. La section 3 est consacrée à l'analyse économétrique des coûts de traitement hospitaliers, avec une présentation de la stratégie empirique, des données, des spécifications adoptées et des estimations obtenues. Celles-ci permettent de simuler, dans la section 4, les économies budgétaires potentielles qui pourraient découler de l'application d'une tarification par pathologie.

## 2 La régulation des hôpitaux : mises en perspective historiques et éléments de théorie

Le secteur des soins hospitaliers doit faire l'objet d'une régulation. Cet impératif ne découle pas du statut des établissements, qui peut être privé, public ou privé à but non lucratif. Il est lié au fait que la demande qui s'adresse aux hôpitaux, solvabilisée par l'assurance maladie, est peu influencée par les mécanismes de marché. Les établissements hospitaliers peuvent être régulés par des tutelles aux statuts forts différents. Il peut s'agir de l'Etat (hôpitaux publics français), d'un assureur universel obligatoire (le programme *Medicare*

pour les retraités américains) ou encore d'un assureur privé (cas des HMO aux Etats-Unis). Cet article s'intéresse aux hôpitaux publics français.

Le régulateur doit encourager les hôpitaux à répondre aux besoins, sans induire d'activité injustifiée. Il doit inciter les établissements à l'efficacité, c'est à dire à fournir les soins au moindre coût, sans sacrifier la qualité ni sélectionner les patients. Enfin, le régulateur doit distribuer les ressources de manière équitable entre les hôpitaux et, dans cette perspective, être capable d'évaluer l'activité de production de soins des établissements et les contraintes spécifiques auxquelles certains d'entre eux peuvent être confrontés.

Plusieurs types de régulation peuvent être adoptés. Le paiement à l'acte finance de manière rétrospective les soins au niveau du coût effectivement observé. Le système du budget global consiste à attribuer à l'établissement, selon des critères qui peuvent varier, un budget pour l'ensemble de son activité annuelle. La tarification par pathologie consiste à financer chaque cas traité par un forfait calculé de façon prospective pour chaque pathologie. Le principe de l'achat de soins, enfin, consiste à sélectionner par un mécanisme d'enchères le prestataire d'un panier de soins prédéfinis.

Dans ce qui suit, nous faisons une brève présentation des évolutions récentes de la régulation des hôpitaux en France et aux Etats-Unis, dans le système *Medicare*. Ce dernier exemple est convoqué à titre d'expérience du mode de régulation qui pourrait être envisagé pour la France. Puis nous donnons des repères théoriques permettant de comprendre les principes de la tarification par pathologie et son évolution vers un concept de paiement mixte.

## **2.1 Le budget global en France : inefficacités et iniquité**

Comme dans la plupart des pays de l'OCDE, les hôpitaux français ont tout d'abord été financés par un système de paiement rétrospectif. Ce système permet un haut niveau de qualité des soins mais n'incite pas à l'efficacité.

Entre 1945 et 1983, le paiement rétrospectif prend la forme du calcul d'un prix de journée qui fonde la procédure d'allocation des ressources. L'absence de contrainte budgétaire qui prévaut durant cette période se traduit par une forte dérive des dépenses hospitalières mais aussi par le développement d'une médecine de pointe (Moison et Tonneau (1997)).

En 1983 est instauré le système du budget global. De façon générale, un tel système permet de contrôler la dépense hospitalière, mais risque d'entraîner une diminution du volume de services rendus et de la qualité des soins.

Dans le cas français, le budget est établi par l'application d'un taux directeur à la dotation de l'exercice précédent. Ce dispositif a contribué à figer les inégalités de dotations initiales indépendamment de l'évolution de l'activité des hôpitaux.

Concrètement, la sévérité de la contrainte budgétaire subie par un hôpital dépend de trois éléments : (i) le niveau d'efficacité de l'établissement lors de l'année servant de base historique au calcul du budget global ; (ii) son dynamisme en matière d'activité et d'acquisition des innovations techniques ; (iii) le pouvoir de négociation du directeur de l'hôpital auprès de la tutelle afin d'influencer le budget alloué. Il est clair que l'efficacité initiale et le dynamisme sont en faveur du niveau et de la qualité du service rendu : dans ce système, ils jouent dans le sens d'un resserrement de la contrainte budgétaire, toutes choses égales par ailleurs. Ainsi, le système du budget global, tel qu'il est appliqué actuellement, permet des inefficacités dans l'activité des établissements et conduit à une allocation des ressources inéquitable.

Conscientes de ces imperfections, les autorités de régulation ont très vite cherché à mettre en place une collecte d'information sur l'activité de production de soins des établissements. Le lancement en 1982 du Programme de médicalisation des systèmes d'informations (PMSI) répond à cet objectif. Opérationnel à partir de 1994 seulement, le PMSI permet d'évaluer les services rendus par les hôpitaux en nombre de points ISA (indice synthétique d'activité). Les ordonnances de 1996 réforment le système hospitalier en créant les agences régionales de l'hospitalisation (ARH), tutelles uniques au niveau régional. Les directeurs des ARH sont dotés d'une position statutaire qui leur garantit en principe une indépendance vis-à-vis des pressions de diverses natures. Ils ont toute latitude pour appliquer une logique d'égalisation des dotations en effectuant les arbitrages budgétaires en fonction du nombre de points ISA réalisé par les établissements.

Le bilan des ordonnances de 1996 est mitigé : les redéploiements de ressources n'ont pas eu lieu, peut-être à cause d'un contexte difficile de restriction budgétaire. Le budget global est toujours en place, avec les inefficacités et les allocations inéquitables afférentes. La tutelle examine sérieusement l'opportunité de se tourner vers un autre mode de régulation : la tarification à la pathologie a tout d'abord été envisagée avec une expérimentation prévue dans l'article L716.2 du code de la santé publique. Il a maintenant été déci-

dée qu'à partir de Janvier 2004 serait appliquée une tarification "à l'activité" dont le principe - à l'étude - s'inspire de la tarification par pathologie.

## **2.2 L'expérience de Medicare aux Etats-Unis : de multiples amendements à la tarification forfaitaire**

Système d'assurance maladie des retraités américains de 65 ans et plus, le programme *Medicare* rompt dès 1983 avec le paiement rétrospectif en instaurant un système de tarification forfaitaire pour rembourser les séjours hospitaliers de ses ayants droit. Le but recherché est de parvenir à une maîtrise des dépenses hospitalières.

Très rapidement, on s'est écarté de l'application stricte de paiements prospectifs. Des ajustements ont été introduits pour tenir compte des disparités des coûts salariaux sur le territoire américain, des différences de statut entre les établissements (mission d'enseignement) ou encore de l'existence d'une forte proportion de patients à bas revenus ou souffrant de maladies chroniques. Des remboursements additionnels ont été créés pour aider les hôpitaux à supporter les coûts dans les cas extrêmes, par exemple dans le cas d'une durée de séjour particulièrement longue. Avec la mise en place d'instances de contrôle, ces nombreux aménagements au système initial ont été introduits afin de limiter les comportements déviants auxquels peut inciter un paiement forfaitaire trop rigide : sélection de clientèle, baisse de la qualité des soins ou encore manipulation du codage des séjours (pour aboutir à une pathologie mieux rémunérée).

Ainsi, la tarification a intégré de plus en plus de paiements à caractère rétrospectif (McClellan, 1997). Cette évolution peut être invoquée comme explication à l'efficacité limitée de la tarification à la pathologie sur la maîtrise des coûts.

## **2.3 Principe de la tarification par pathologie et évolution vers un concept de paiement mixte**

Le pragmatisme de l'administration de *Medicare* l'a conduite à introduire ces multiples aménagements pour assouplir le système initial. Dans le même temps, de nombreuses contributions théoriques faisaient évoluer le modèle de base en donnant des fondements théoriques à des paiements mixtes, combinant forfait et remboursements rétrospectifs.

La modélisation retenue suppose en général que le coût de traitement d'un patient atteint d'une pathologie particulière dans l'hôpital  $h$  est donné<sup>1</sup> par :  $C_h = c_h - e_h$ .  $c_h$  et  $e_h$  sont des informations privées de l'hôpital;  $c_h$  est un paramètre de productivité, d'autant plus faible que la productivité de l'hôpital est élevée et  $e_h$  est l'effort de réduction du coût. Fournir l'effort  $e_h$  implique pour l'hôpital une désutilité égale à  $\xi(e_h)$ . La fonction  $\xi(\cdot)$  est supposée continue, croissante et convexe. Les services rendus engendrent un surplus égal à  $S_h > 0$ , en échange duquel il reçoit un transfert  $P_h$  de la part de la tutelle. Les hôpitaux maximisent leur utilité  $\Pi_h = P_h - C_h - \xi(e_h)$ . Le régulateur cherche à définir des transferts qui permettent de maximiser le surplus net des consommateurs de soins, sous la contrainte que les hôpitaux ne fassent pas faillite :  $Max_h (S_h - P_h)$ , s.c.  $\Pi_h > 0 \forall h$ . Chaque hôpital est supposé être en situation de monopole local. Il n'y a pas de comportement de collusion entre les établissements hospitaliers.

La tarification par pathologie est d'abord un contrat à prix fixe. Le paiement proposé étant un forfait défini indépendamment de son coût, l'hôpital a intérêt à réduire ses coûts au minimum : il fournit l'effort optimal de premier rang tel que  $\xi'(e^*) = 1$ . A ce stade, une partie seulement du problème est résolue. En effet,  $c_h$  est une information privée de l'hôpital : la définition du forfait par la tutelle peut le conduire à la faillite ou lui attribuer des rentes excessives.

Le problème de la tutelle est donc de trouver le niveau de paiement correspondant à la production efficace, autrement dit de définir une formule de tarification permettant d'extraire la totalité, ou de réduire le plus possible la rente informationnelle.

- Le modèle de Shleifer (1985) résout ce problème informationnel à l'aide d'une hypothèse simplificatrice qui consiste à supposer que les paramètres de productivité des hôpitaux sont identiques :  $c_h = c \forall h$ . Dans ce cas les disparités de coûts sont exclusivement dues à l'aléa moral :  $C_h = c - e_h$  et la tutelle peut mettre en place un mécanisme de concurrence par comparaison. Celui-ci consiste à proposer à chaque hôpital un paiement défini sur la base

---

<sup>1</sup>De façon générale, les modèles définissent le paiement d'un séjour pour une pathologie, avec des soins dont la qualité est fixée. On peut enrichir la modélisation en considérant aussi une variable représentant la qualité des soins (Ma, 1994 et Mougeot, 1999<sub>b</sub>) : celle-ci vient augmenter le coût de production mais aussi la demande de soins qui s'adresse à l'hôpital. Les modèles correspondants étudient alors, non pas le paiement d'un séjour, mais celui du volume de soins prodigués par l'hôpital, volume endogénéisé *via* la qualité.

de la moyenne des coûts observés pour les autres hôpitaux en fin d'exercice

$$P_h = \xi(e^*) + \bar{C}_h, \text{ où } \bar{C}_h = \frac{C_k}{H-1}, H \text{ désignant le nombre d'hôpitaux régulés.}$$

Ici,  $\bar{C}_h$  est défini de telle sorte qu'il ne peut pas être influencé par  $C_h$  : le contrat qui en résulte a les propriétés d'un contrat à prix fixe. Comme la règle de paiement est annoncée en début d'exercice, la moyenne correspond *in fine* au coût lié à l'effort optimal :  $C_h = c - e^* = \bar{C}_h, \forall h$ . Au total, les transferts  $P_h$  sont tels que chaque hôpital atteint l'équilibre budgétaire :  $P_h = c - e^* + \xi(e^*)$ . L'expression de  $P_h$  permet de constater qu'il s'agit bien d'un forfait qui correspond au coût de production efficace<sup>2</sup>.

- La représentation idéale de Shleifer constitue le fondement théorique de la tarification par pathologie. Elle repose sur des hypothèses peu réalistes : homogénéité des hôpitaux, homogénéité des patients pour une pathologie donnée, qualité des soins fixée. Or, les risques associés à une tarification trop homogène sont connus : sélection ou discrimination des patients, baisse de la qualité des soins délivrés (Newhouse (1996)).

Les contributions théoriques ont cherché à améliorer le modèle de base en remettant en cause l'hypothèse d'homogénéité des patients (Keeler (1990), Pope (1990), Ma (1994, 1998), Ellis (1998)), ou en remettant en cause l'hypothèse d'homogénéité des hôpitaux (Auriol et Laffont (1992), Laffont et Tirole (1993), Ellis (1998)). A partir d'hypothèses et d'approches théoriques très différentes, ces modèles conduisent à définir des tarifications optimales correspondant à des paiements mixtes, combinant forfait et coûts observés.

Le principe d'un système de paiement mixte fait maintenant l'objet d'un consensus, rejoignant ainsi la pratique de la régulation aux Etats-Unis. Mais le coefficient de partage entre forfait et coût effectif est défini de façons très différentes selon le modèle théorique auquel on se réfère, son corps d'hypothèses et sa paramétrisation. De plus, ce coefficient peut dépendre de variables ou de fonctions non observables (comme par exemple la désutilité de l'effort chez Laffont et Tirole (1993)). Le problème débouche alors sur des questions largement empiriques que nous abordons dans le cas français : comment identifier le niveau des coûts correspondant à une activité efficace ? Comment intégrer les hétérogénéités des patients et celle des établissements dans la tarification des séjours hospitaliers ?

---

<sup>2</sup> Auquel s'ajoute la valeur de la désutilité de l'effort de premier rang (la fonction  $\xi(\cdot)$  est supposée identique  $\forall h$ ).

## 3 Analyse économétrique des coûts de traitement hospitaliers

### 3.1 Logique de la démarche empirique adoptée

Nos observations des coûts se situent au niveau individuel du séjour hospitalier<sup>3</sup>. Ces coûts résultent d'une activité financée selon le système du budget global. Ils reflètent donc en partie le mode de fixation du budget : application d'un taux directeur au budget précédent, négociation plus ou moins fructueuse du gestionnaire de l'hôpital auprès de la tutelle.

Sur le terrain, les établissements disposant d'une bonne comptabilité évaluent les coûts par séjour et transmettent cette information en vue de la constitution de la base de coût PMSI, que nous utilisons pour les évaluations empiriques. Dans ce cadre, la variabilité des coûts observés par séjour est influencée par plusieurs éléments : (i) des caractéristiques des patients traités qui impliquent effectivement des coûts de traitement différenciés, (ii) des caractéristiques des établissements (infrastructures, économies de gamme ou d'échelle), (iii) des inefficacités de production (rendues plus ou moins possibles selon les largesses obtenues lors de la négociation budgétaire).

Cet article vise à évaluer l'effet potentiel de l'adoption d'un système de tarification par pathologie sur la dépense consacrée aux hôpitaux publics en France. Nous proposons une démarche en deux étapes.

- La première étape consiste à effectuer une analyse économétrique des coûts hospitaliers de façon à repérer, dans la variabilité des coûts, ce qui peut être attribuable à des inefficacités. La méthode retenue correspond dans son principe aux approches en termes de frontières de production : des différences de coût, repérées toutes choses égales par ailleurs, entre deux établissements, sont identifiées comme reflétant différents niveaux d'efficacité<sup>4</sup>.

- Dans la seconde étape, on évalue, sur les échantillons utilisés, le niveau de budget qui aurait suffi dans le cas où les coûts auraient correspondu à une activité efficace. Sous l'hypothèse qu'une tarification par pathologie permettrait d'obtenir ces niveaux d'efficacité (Shleifer, 1985), nous interpré-

---

<sup>3</sup>Notre approche s'écarte donc de la littérature, relativement abondante, qui évalue l'efficacité des hôpitaux à partir de données sur les coûts moyens par établissement. Un tour d'horizon très synthétique de cette littérature peut être trouvé dans Linna (1998).

<sup>4</sup>En réalité l'approche retenue est plus complexe, puisqu'elle nous amène à distinguer les différences permanentes et transitoires (voir la section 3.3).

tons les résultats obtenus comme les effets simulés des gains potentiels liés à l'instauration d'un tel système de tarification. Outre une évaluation des gains budgétaires potentiels, les coûts de traitement efficaces estimés permettent de définir les formules de paiements que nous préconisons.

### **3.2 Des données à trois dimensions : 7 314 séjours pour infarctus du myocarde dans les hôpitaux publics français sur les années 1994-1997**

On dispose pour l'étude d'un échantillon de 7 314 séjours observés dans 36 hôpitaux sur les années 1994-1997. Ces données sont issues de la base de coût constituée par le Programme de médicalisation des systèmes d'informations (PMSI). Dans le PMSI, les séjours sont classés en "Groupes Homogènes de Malades" (GHM) à l'aide d'une arborescence sur la base des différents diagnostics et actes pratiqués sur le patient. Un GHM définit la pathologie pour laquelle la tutelle est supposée fixer une tarification. Afin que les séjours observés soient les plus homogènes possible en termes de pathologies, nous avons sélectionné les séjours de patients âgés d'au moins 40 ans, pour lesquels le diagnostic principal est l'infarctus du myocarde aigu et qui ont été classés dans le même GHM : Infarctus du myocarde sans complication (GHM 179).

Pour chaque séjour, on observe le coût du séjour et les diagnostics secondaires, les actes pratiqués, le mode d'entrée (en provenance du domicile ou d'un autre hôpital), le mode de sortie (sortie vers le domicile, un autre hôpital ou un autre service), la durée du séjour, l'âge et le sexe du patient.

Les données du PMSI permettent d'accéder à une information d'une richesse considérable : les observations sont effectuées au niveau de chaque séjour et sont très détaillées en ce qui concerne les diagnostics, les traitements et les coûts associés. Le PMSI offre toutefois une vision restrictive de la production hospitalière. Les seules observations sur l'output sont constituées des modalités de la variable "mode de sortie", c'est-à-dire le décès, le transfert ou le retour au domicile. On ne dispose pas d'informations relatives à la qualité de vie du malade après le séjour, à une réhospitalisation précoce, ni à la contraction d'une éventuelle infection nosocomiale. Enfin, on ne dispose pas de renseignement sur la qualité du service rendu en termes d'accueil, de confort et de prise en charge de la douleur.

Si les données du PMSI couvrent exhaustivement les hôpitaux publics français, les données de la base de coût sont particulières dans la mesure où

la participation à sa constitution est définie sur le principe du volontariat. Les établissements, en nombre limité, qui ont accepté de fournir les informations sur leurs coûts par séjour disposent par définition d'une bonne comptabilité analytique. Des analyses comparatives avec la base exhaustive<sup>5</sup> révèlent toutefois que la base de coût est représentative des séjours effectués dans les hôpitaux publics français pour la pathologie étudiée et les actes pratiqués.

La structure des données est assez complexe. Tout d'abord, il s'agit de données stratifiées à trois dimensions : séjour-hôpital-année. Ensuite, les données sont non cylindrées dans plusieurs dimensions<sup>6</sup> : non seulement le nombre de séjours observés varie selon l'hôpital et l'année considérés, mais la période d'observation des hôpitaux est de longueur variable. Dans ce qui suit, nous procédons à une brève description de nos données au niveau des séjours, puis au niveau - supérieur - des hôpitaux où ces séjours ont eu lieu.

### **3.2.1 Les séjours pour infarctus du myocarde sans complication**

Le GHM 179 que nous avons sélectionné est très majoritairement représenté parmi les séjours pour infarctus du myocarde aigu : il en représente environ les deux-tiers. Les principales caractéristiques des patients correspondant aux séjours classés dans le GHM 179 figurent dans le tableau 1. Les trois-quarts sont des hommes, relativement jeunes. 89 % des patients proviennent de leur domicile. 64 % des sorties s'effectuent vers le domicile et 36 % vers un autre hôpital<sup>7</sup>.

#### **Insérer ici tableau 1**

Dans le tableau 1 figurent aussi les taux d'application des actes innovants : cathétérisme, angioplastie et pose de stents. Ces actes sont des innovations techniques améliorant la performance du traitement, notamment en termes de diagnostics, de soin et de qualité de vie du patient. Ils ne feraient pas l'objet d'un paiement spécifique dans le cadre d'une tarification par pathologie puisque leur réalisation ne conduit pas au classement du séjour dans un GHM spécifique. Or, ils sont particulièrement coûteux : nous avons estimé

---

<sup>5</sup>Pour laquelle les coûts ne sont pas observés.

<sup>6</sup>Comme on le verra plus loin, cette caractéristique a des conséquences sur les méthodes utilisables.

<sup>7</sup>Les infarctus du myocarde avec décès sont classés dans un GHM spécifique (GHM 180). Dans l'ensemble des infarctus du myocarde aigu observés dans la base de coût, le taux de décès est de 9 % environ.

que leur réalisation impliquait un surcoût de 34 % (cathétérisme) à 61 % (angioplastie) par rapport au coût moyen du GHM 179 (Milcent (2001)). Une tarification ne prenant pas en compte ces actes pénaliserait donc les hôpitaux les pratiquant.

### 3.2.2 Les hôpitaux observés

La répartition par année des hôpitaux de la base de coûts apparaît dans le tableau 2. Toutes années confondues, nous observons 36 hôpitaux. L'année 1996 ne comprend que 17 hôpitaux car de nombreuses observations ont été perdues, cette année-là, à cause d'un changement de nomenclature dans le codage des diagnostics<sup>8</sup>.

#### Insérer ici tableau 2

La mise en oeuvre des actes innovants évoqués ci-dessus requière des équipements et des compétences techniques spécifiques. De fait, une part non négligeable des hôpitaux ne pratiquent jamais ces actes. Nous avons classé les établissements en deux catégories : les hôpitaux techniques et les hôpitaux non techniques. Un hôpital est défini comme technique une année donnée s'il a pratiqué des cathétérismes sur au moins 2 % de ses séjours et/ou au moins une angioplastie<sup>9</sup>. Sur la base de cette définition, on obtient 20 hôpitaux techniques sur les 36 hôpitaux observés. Les hôpitaux techniques correspondent à 71,5 % des séjours observés (tableau 2).

Pour compléter les données, nous avons intégré des informations fournies par la Statistique Annuelle des Etablissements de santé (SAE)<sup>10</sup> : la catégorie de l'établissement, le nombre de lits, de journées-lits et le taux d'occupation des lits en médical (respectivement, en chirurgical), le nombre de disciplines pratiquées dans l'établissement. On a défini trois modalités pour la catégorie de l'établissement : CHR désigne un centre hospitalier régional, qui assume dans la quasi-totalité des cas des missions d'enseignement et de recherche ; PUB correspond à un hôpital public sans caractéristique particulière ; PRIV désigne un établissement privé, à but non lucratif, participant au service public hospitalier (PSPH). Ces derniers établissements présentent la particularité de n'avoir été soumis que progressivement à la contrainte de budget

---

<sup>8</sup>Passage du CIM9 (Code International des maladies 9) au CIM10.

<sup>9</sup>Accompagnée ou non de la pose de stents.

<sup>10</sup>L'enquête SAE couvre tous les établissements de santé publics et privés installés en France (métropole et DOM).

global et d’obéir à un impératif de rentabilité pour une partie de leurs patients (notamment les longs séjours). Les données du tableau 3 utilisent les 95 observations se situant au niveau des hôpitaux-années. On constate que toutes les observations relatives aux CHR correspondent à des hôpitaux techniques, ainsi que la majorité des observations relatives aux hôpitaux privés PSPH.

### Insérer ici tableau 3 et 4

Le tableau 4 présente les corrélations entre la catégorie de l’hôpital et différents indicateurs moyens calculés pour les 95 hôpitaux-années de la base de coût. Les CHR ont un faible taux de sortie par transfert. Les hôpitaux privés PSPH se caractérisent par un fort taux d’entrée par transfert, et des taux d’actes techniques élevés. Les hôpitaux publics standards (PUB) sont peu techniques, ont un faible taux d’entrée par transferts, ”exportent” beaucoup leurs patients et ont un faible taux d’actes. Les hôpitaux techniques ont un faible taux de sortie par transfert et - ce n’est guère une surprise - de forts taux d’actes innovants. Les flux de patients vers ces hôpitaux capables de pratiquer les actes techniques apparaît clairement dans les corrélations positives entre taux d’entrée par transfert et taux d’application des actes innovants et négative entre taux de sortie par transfert et taux de cathétérisme<sup>11</sup>.

### 3.3 Spécification économétrique des coûts des séjours

Soit  $C_{i,h,t}$  le coût du séjour  $i$  effectué dans l’hôpital  $h$  pendant l’année  $t$ . Nous avons retenu pour la fonction de coût la spécification suivante :

$$C_{i,h,t} = X'_{i,h,t}\gamma_t + W'_{h,t}\alpha + Q'_h\lambda + a + c_t + \eta_h - \varepsilon_{h,t} + u_{i,h,t} \quad (1)$$

$X'_{i,h,t}$  représente les caractéristiques individuelles observables des patients : effets croisés âge-sexe, mode d’entrée, mode de sortie, durée de séjour. Les variables  $W'_{h,t}$  et  $Q'_h$  sont les caractéristiques observables de l’hôpital, constantes

---

<sup>11</sup>L’enquête SAE n’est pas bien renseignée pour les variables indicatrices de la taille et de la diversification des activités des établissements. De ce fait, nous n’avons pas pu calculer, pour ces variables, de corrélations pour toutes les observations hôpitaux-années, mais seulement pour 73 (sur 95). Ces corrélations montrent que les CHR sont de grande taille et ont une activité très diversifiée. Les hôpitaux privés PSPH, en revanche, concentrent leur pratique sur un faible nombre de disciplines.

ou variables dans le temps : catégorie de l'établissement, technicité, taux d'entrée et de sortie par transfert, taux d'application des actes innovants.  $a$  est une constante.

A caractéristiques des patients données, la variabilité des coûts peut trouver sa source dans les caractéristiques des hôpitaux : existence d'économies d'échelle ou de gamme, missions qui peuvent être assurées (CHR, hôpital de proximité), diversification des pathologies traitées, qualité des soins délivrés (mise en oeuvre des technologies innovantes, accueil, prise en charge de la douleur, etc.), niveau de capital humain du personnel soignant, qualité de la gestion des ressources humaines, rigueur dans la gestion de l'hôpital. Certains de ces déterminants sont observables, d'autres non observables.

Dans cet article, nous supposons que la tutelle est économètre. Plus précisément, nous supposons que la tutelle dispose des données du PMSI pour établir les paiements hospitaliers : le partage entre variables observables et non observables est alors le même pour la tutelle et l'économètre<sup>12</sup>. Les caractéristiques observables sont, concernant les patients, les variables  $X'_{i,h,t}$ , concernant les hôpitaux, les variables  $W'_{h,t}$  et  $Q'_h$ . A caractéristiques observables données, la variabilité des coûts dépend, dans la spécification (1), du terme

$$c_t + \eta_h - \varepsilon_{h,t} + u_{i,h,t}.$$

$c_t$  est un effet fixe temporel, commun à tous les établissements, qui est lié à la progression générale des budgets hospitaliers<sup>13</sup> et reflète l'évolution des prix, des salaires et du progrès technique.

L'hétérogénéité non observée des patients est spécifiée par la perturbation  $u_{i,h,t}$ , supposée iid  $(0, \sigma_u^2)$ .  $\varepsilon_{h,t}$  est une perturbation supposée iid  $(0, \sigma_\varepsilon^2)$  et non corrélée avec  $u_{i,h,t}$ , dont nous étudierons l'interprétation ci-dessous.

#### *a) Interprétation de l'effet spécifique hôpital $\eta_h$*

L'hétérogénéité non observée des établissements est spécifiée comme un effet spécifique hôpital, qui peut être fixe ou aléatoire.  $\eta_h$  peut être conçue

---

<sup>12</sup>En réalité, la tutelle dispose d'informations supplémentaires : celles, plus ou moins informelles, qui résultent des contacts directs liés aux négociations avec la direction de l'hôpital. Cependant, l'information émanant du PMSI est vérifiable et celle émanant des contacts directs plus manipulable. Dans la perspective de régulation retenue dans cet article, il est raisonnable d'écarter cette source d'information des calculs des paiements.

<sup>13</sup>Dans le contexte institutionnel français en vigueur depuis les ordonnances de 1996,  $c_t$  est lié à la progression des budgets hospitaliers qui découle du vote de l'ONDAM (Objectif Quantifié de Dépenses en Assurance Maladie), dans le cadre de la loi de financement de la sécurité sociale.

comme la résultante de trois composantes :

$$\eta_h = \eta_h^{as} + \eta_h^{hm} + \eta_h^q.$$

Dans le cadre théorique d'une relation d'agence entre la tutelle et l'hôpital, où la tutelle observe mal l'effort de réduction du coût fourni par le gestionnaire de l'hôpital (aléa moral) et les caractéristiques de l'établissement expliquant sa productivité (antisélection), les composantes de  $\eta_h$  peuvent s'interpréter de la façon suivante :

$\eta_h^{as}$  est un paramètre d'antisélection : les infrastructures de l'hôpital conduisent à un fonctionnement plus ou moins coûteux, il peut exister des économies d'échelle ou de gamme<sup>14</sup>.  $\eta_h^{hm}$  représente le hasard moral de long terme : la gestion de l'hôpital peut être dispendieuse de façon permanente. Enfin,  $\eta_h^q$  correspond à la qualité : performances des soins, accueil, prise en charge de la douleur, etc.

*b) Interprétation du terme  $\varepsilon_{h,t}$*

La variabilité des effets spécifiques  $\eta_h$  reflète les écarts de coûts moyens entre les établissements, à caractéristiques observables  $X'_{i,h,t}$ ,  $W'_{h,t}$  et  $Q'_h$  données. Comment maintenant interpréter la perturbation  $\varepsilon_{h,t}$  ? À caractéristiques observables et à choc conjoncturel ( $c_t$ ) donnés, elle est définie comme l'écart, une année  $t$ , de l'hôpital  $h$  à son niveau de coût moyen. En tant que tel,  $\varepsilon_{h,t}$  peut donc être compris comme un indicateur de l'effort transitoire de réduction du coût, autrement dit, de l'aléa moral correspondant à l'effort fourni par le gestionnaire de l'hôpital en matière de réduction des coûts<sup>15</sup>. Concrètement, on peut penser à la plus ou moins grande rigueur qui peut être adoptée par celui-ci dans les procédures de négociation pour les marchés des consommables et pour les tarifs des différentes activités assurées par des intervenants extérieurs.

En principe,  $\varepsilon_{h,t}$  contient aussi les composantes ordinaires de toute perturbation : variables omises et erreurs de mesure. Ces dernières doivent toutefois être d'une importance modérée.

---

<sup>14</sup>En utilisant les données de l'enquête SAE, nous avons essayé de saisir ces effets directement en intégrant dans le modèle la taille de l'établissement et la diversification de son activité, sans obtenir de résultat significatif, sans doute à cause du fait que l'on étudie ici le coût des séjours pour une seule pathologie.

<sup>15</sup>Pour faire à nouveau référence au contexte institutionnel du budget global,  $\varepsilon_{h,t}$  correspond à l'inefficacité *autorisée*, compte tenu de la contrainte budgétaire plus ou moins lâche subie par l'hôpital  $h$  en  $t$ .

Par définition,  $\varepsilon_{h,t}$  est en effet la perturbation d'une équation qui explique le coût de tous les séjours : dans ce cadre, une erreur de mesure figurant dans  $\varepsilon_{h,t}$  devrait affecter systématiquement tous les enregistrements des séjours pour l'hôpital  $h$  à l'année  $t$ . Il ne peut donc s'agir d'une erreur de retranscription pour un patient, mais d'un biais systématique, une année donnée, dans les critères utilisés pour définir la durée d'un séjour par exemple, ou encore d'une erreur sur la catégorie d'établissement à laquelle appartient l'hôpital. Cette dernière hypothèse est peu vraisemblable.

Examinons maintenant l'hypothèse des variables omises qui pourraient figurer dans  $\varepsilon_{h,t}$  : elles correspondent à des chocs ayant affecté l'hôpital  $h$  et lui seul, une année  $t$ . Il peut s'agir d'une panne touchant par exemple son système d'eau courante. Nous pensons que la tutelle a intérêt à assimiler *a priori* ces accidents à du hasard moral, afin de pousser les établissements à les déclarer, dans le cas où ils estimeraient que les écarts de coûts auxquels ils conduisent sont exceptionnels et justifiables<sup>16</sup>.

On peut donc interpréter la perturbation  $\varepsilon_{h,t}$  comme un indicateur de l'aléa moral transitoire, lié aux efforts fournis par le gestionnaire de l'hôpital en matière de réduction des coûts.

Au total, puisqu'il y a vraisemblablement de l'aléa moral de long terme dans les caractéristiques des hôpitaux  $\eta_h$ , notre estimation ne permettra pas de repérer *tout* l'aléa moral en estimant la variance de  $\varepsilon_{h,t}$ , mais seulement l'aléa moral transitoire. En tout état de cause, les développements qui précèdent permettent raisonnablement de penser que la variance de  $\varepsilon_{h,t}$  est *exclusivement* due à de l'aléa moral<sup>17</sup>.

### *c) Forme fonctionnelle de l'équation de coût*

Avant d'aborder l'estimation proprement dite de l'équation de coût, soulignons que la spécification (1) est linéaire, avec pour variable expliquée  $C_{i,h,t}$

---

<sup>16</sup>Ce raisonnement revient à effectuer une distinction entre la position de la tutelle et celle des économètres auteurs de cet article. Mais dans l'hypothèse d'une tutelle-économètre qui appliquerait les paiements définis ici, il n'y aurait plus de distinction : la tutelle-économètre observerait les accidents de parcours, vérifiables, que les hôpitaux auraient intérêt à les révéler.

<sup>17</sup>Dans la littérature sur les frontières de production, l'identification des inefficacités productives a tout d'abord été réalisée à partir de spécifications de perturbations sous la forme de lois demi-normales permettant d'effectuer la distinction avec la perturbation du modèle. Ici, le fait de disposer de données à plusieurs dimensions permet d'adopter une spécification semi-paramétrique en termes d'effets spécifiques, beaucoup moins contraignante (cf. Wagstaff, 1989).

et non  $\text{Log}(C_{i,h,t})$ . Ce choix, est à la fois en rapport avec l’objet de cet article et justifié par les tests réalisés. Notre propos est tout d’abord de définir une formule de paiement suffisamment lisible pour qu’une nouvelle tarification soit compréhensible et donc recevable par les acteurs du système hospitalier. Ensuite, sur un plan plus technique, les tests réalisés montrent que la distribution de  $C_{i,h,t}$  est plus proche d’une loi normale que d’une loi lognormale<sup>18</sup>. Par ailleurs, les variables explicatives sont toutes qualitatives, à l’exception de la durée de séjour. Pour étudier l’éventualité d’une relation non linéaire entre le coût et la durée de séjour, nous avons introduit des termes quadratiques qui se sont révélés globalement non significatifs<sup>19</sup>. Notons enfin que l’introduction d’une variable comme l’âge sous une forme qualitative, avec des tranches fines (quatre tranches) et de nombreux effets croisés permet de spécifier, dans le cadre d’un modèle linéaire comme (1), de nombreuses non-linéarités dans la relation entre  $C_{i,h,t}$  et ses déterminants.

### 3.4 Estimation de la fonction de coût

#### 3.4.1 L’effet spécifique hôpital : fixe ou aléatoire ?

Considérons à nouveau le modèle (1).  $\eta_h$  peut être spécifié comme un effet fixe ou aléatoire.

- Supposer que  $\eta_h$  est aléatoire implique que l’hétérogénéité non observée ne joue sur les coûts qu’au deuxième ordre (sur leur variance) et est non corrélée avec les caractéristiques observables  $X'_{i,h,t}$ ,  $W'_{h,t}$  et  $Q'_h$ . En supposant que  $\eta_h$  est iid  $(0, \sigma_\eta^2)$ , on devrait obtenir une estimation convergente et efficace en appliquant les moindres carrés quasi-généralisés à :

$$C_{i,h,t} = X'_{i,h,t}\gamma_t + W'_{h,t}\alpha + Q'_h\lambda + a + c_t + \underbrace{\eta_h - \varepsilon_{h,t} + u_{i,h,t}}_{v_{i,h,t}} \quad (2)$$

En pratique, cette méthode d’estimation ne s’applique pas directement, et ce pour deux raisons : (i) la matrice de variance-covariance de la perturbation

---

<sup>18</sup>Ce résultat est vraisemblablement lié au fait que la sélection des séjours classés dans le GHM 179 implique une troncature à droite de la distribution pour les séjours plus coûteux, lesquels sont classés dans d’autres GHM (avec complications, ou avec pontage).

<sup>19</sup>Sauf dans un cas : pour l’effet croisé durée<sup>2</sup>-année 1996, lorsque l’on admet que les coefficients de la durée de séjour et de la durée de séjour au carré varient selon les années.

$v_{i,h,t} = \eta_h - \varepsilon_{h,t} + u_{i,h,t}$  a une structure particulière du fait de la nature stratifiée des données (séjours-hôpitaux-année) ; (ii) les données sont non cylindrées dans plusieurs dimensions : non seulement le nombre de séjours observés varie selon l'hôpital et l'année considérés, mais la période d'observation des hôpitaux est de longueur variable. De ce fait, notre problème est différent de celui de Baltagi, Song et Jung (2001), qui étudient l'estimation d'un modèle à erreurs composées stratifiées dans un panel à trois dimensions, non cylindré dans une seule dimension. Dans notre cas, Antweiler (2001) montre que l'on ne peut pas trouver de transformation simple des données qui permette d'obtenir l'estimateur des moindres carrés quasi-généralisés (*via* l'application des moindres carrés ordinaires aux données transformées) et propose d'utiliser l'estimateur du maximum de vraisemblance (MV). En supposant que l'on peut retenir une hypothèse de normalité des perturbations, et que les différentes composantes de  $v_{i,h,t}$  sont indépendantes entre elles, cette procédure permet d'obtenir une estimation convergente et asymptotiquement efficace (voir annexe).

- On peut aussi supposer que  $\eta_h$  est un effet fixe. Dans ce cas, le modèle comporte des variables indicatrices des hôpitaux (pour estimer les effets fixes  $\eta_h$ ) et on ne peut pas identifier les paramètres  $\lambda$ . Une estimation convergente et efficace est obtenue par l'application des moindres carrés quasi-généralisés (MCQG) au modèle :

$$C_{i,h,t} = X'_{i,h,t}\gamma_t + W'_{h,t}\alpha + a + c_t + \eta_h - \underbrace{\varepsilon_{h,t} + u_{i,h,t}}_{\xi_{i,h,t}} \quad (3)$$

### 3.4.2 Tests de spécification

- Il est assez logique de penser que l'effet spécifique  $\eta_h$ , qui reflète l'aléa moral de long terme et les caractéristiques de l'hôpital en termes d'infrastructures et de qualité des soins, puisse être corrélé avec des variables explicatives comme celles qui expriment son statut ou ses capacités techniques. Pour étudier cette possibilité et trancher entre les hypothèses d'un effet hôpital fixe ou aléatoire, nous avons défini un test de spécification en étendant au cas de nos données non cylindrées à trois dimensions l'approche retenue par Mundlak (1978) pour le modèle à erreurs composées standard. L'hypothèse nulle testée est celle d'absence de corrélation entre  $\eta_h$  et les variables explicatives du modèle. En supposant qu'une éventuelle corrélation entre  $\eta_h$  et ces variables peut

s'écrire comme une régression affine de la forme<sup>20</sup> :  $\eta_h = X'_{\cdot,h,\cdot} \pi_1 + W'_{h,\cdot} \pi_2 + \beta_h$ , avec  $\beta_h$  iid  $(0, \sigma_\beta^2)$ , non corrélé avec  $\varepsilon_{h,t}$  ni avec  $u_{i,h,t}$ , le test d'indépendance de l'effet spécifique hôpital revient à tester la nullité des coefficients  $\pi_1$  et  $\pi_2$  dans le sur-modèle :

$$C_{i,h,t} = X'_{i,h,t} \gamma_t + W'_{h,t} \alpha + X'_{\cdot,h,\cdot} \pi_1 + W'_{h,\cdot} \pi_2 + a + c_t + \underbrace{\beta_h - \varepsilon_{h,t} + u_{i,h,t}}_{\zeta_{i,h,t}} \quad (4)$$

Compte tenu de la structure des données, le modèle (4) est estimé par la méthode du maximum de vraisemblance définie par Antweiler. Nous avons alors testé l'hypothèse  $H_0 : \pi_1 = \pi_2 = 0$ , par un test de rapport de vraisemblance<sup>21</sup>.

- Comme nous le verrons plus loin, le test conduit à rejeter l'hypothèse d'indépendance entre  $\eta_h$  et les variables explicatives du modèle, ce qui nous conduit à privilégier par la suite le modèle (3), où l'effet hôpital  $\eta_h$  est fixe. Ce modèle a une perturbation de la forme :  $-\varepsilon_{h,t} + u_{i,h,t}$ . Il s'agit d'un modèle à erreur composées standard, qui peut être estimé de façon convergente et asymptotiquement efficace par les moindres carrés quasi-généralisés dès lors que les perturbations  $\varepsilon_{h,t}$  sont non corrélées aux variables explicatives. Cette hypothèse est testée à l'aide d'un test d'Hausman (Dormont (1989)). Nous verrons ci-dessous que ce test permet de valider le modèle où les effets  $\varepsilon_{h,t}$  sont aléatoires et non corrélés aux variables explicatives.

- Enfin, les procédures précédemment décrites sont valides à condition que les variables explicatives soient aussi non corrélées avec l'élément résiduel de la perturbation,  $u_{i,h,t}$ , qui reflète l'hétérogénéité non observée des patients. Le coût du séjour dépend de caractéristiques observables, telles que l'âge, le sexe du patient, la durée du séjour, le fait que l'hôpital soit technique ou non, etc. Il dépend aussi de particularités de l'état de santé ou des préférences du patient, observables par le médecin et inobservables par l'économètre. N'étant pas captées par les variables explicatives du modèle, ces particularités se retrouvent dans la perturbation  $u_{i,h,t}$ . Si elles influencent la durée du séjour ou l'affectation du patient dans un hôpital technique, les variables correspondantes seront non exogènes. Nous avons utilisé un test d'Hausman pour tester l'hypothèse nulle d'exogénéité de ces variables explicatives dans le modèle (3). L'estimateur convergent et asymptotiquement efficace sous

<sup>20</sup>  $X'_{\cdot,h,\cdot}$  et  $W'_{h,\cdot}$  sont les moyennes par hôpital des variables  $X'_{i,h,t}$  et  $W'_{h,t}$ .

<sup>21</sup> On retient l'hypothèse de normalité des perturbations, ce qui est plus restrictif qu'un test d'Hausman.

l'hypothèse nulle est l'estimateur des moindres carrés quasi-généralisés ; l'estimateur convergent sous l'hypothèse alternative et sous  $H_0$  est l'estimateur des double moindres carrés à erreurs composées développé par Baltagi (1981). Les instruments utilisés pour mettre en oeuvre cet estimateur sont l'âge, le sexe (variables explicatives supposées exogènes), les diagnostics secondaires, coronariens ou autres et des variables de proximité géographique permettant d'expliquer l'affectation d'un patient à tel ou tel hôpital. Pour gagner en précision en augmentant le nombre d'instruments utilisés, nous avons élevé au carré et croisé certaines variables instrumentales. Les tests conduisent à ne pas rejeter l'exogénéité des variables explicatives testées pour la fonction de coût. Le modèle (3) sera donc estimé de façon convergente et asymptotiquement efficace par les moindres carrés quasi-généralisés.

### 3.4.3 Résultats obtenus

Les résultats des estimations des modèles (2) et (3) et ceux des tests figurent dans les tableaux 6 et 6bis.

#### Insérer ici tableaux 5 , 6 et 6bis

Deux modèles ont été estimés, qui correspondent à différentes listes de variables  $W'_{h,t}$  décrivant les caractéristiques des hôpitaux. Le modèle *A* retient les indicateurs en rapport avec les missions assignées par la tutelle à l'hôpital et avec les stratégies d'investissement de long terme de celui-ci : la catégorie de l'établissement, sa technicité et ses taux moyens d'entrée et de sortie par transfert. Le modèle *B* incorpore des variables supplémentaires comme les taux annuels moyens d'utilisation des actes innovants.

Pour ne pas alourdir sa présentation, nous ne donnons pas dans le tableau 6 les estimations des coefficients des caractéristiques individuelles  $X'_{i,h,t}$  des séjours (les différents effets croisés conduisent à un total de 32 variables), des indicatrices temporelles  $c_t$  ou des effets spécifiques hôpitaux  $\eta_h$ . Les effets spécifiques hôpitaux estimés dans le cadre du modèle à effets fixes feront l'objet d'une représentation graphique. On peut appréhender l'effet des caractéristiques individuelles, des indicatrices temporelles et de la durée de séjour à l'aide des données du tableau 5, qui présente l'estimation par les moindres carrés quasi-généralisés d'un modèle à effet fixe hôpitaux,

où l'on a simplifié<sup>22</sup> les effets croisés pour les variables  $X'_{i,h,t}$ . Les estimations du tableau 5 confirment des résultats bien connus en matière de pathologie cardiaque : le coût diminue avec l'âge ; il est de façon générale plus élevé pour les hommes. Une journée supplémentaire d'hospitalisation coûte en moyenne, toutes choses égales par ailleurs, 380 € environ. Par ailleurs, l'estimation d'une spécification incomplète n'incluant que les caractéristiques individuelles des patients  $X'_{i,h,t}$  comme variables explicatives permet de constater que 54,2 % de la variance des coûts des séjours peut être expliquée par l'hétérogénéité observable des patients.

Le test du rapport de vraisemblance nous conduit à rejeter, pour les modèles A et B, l'indépendance des effets spécifiques hôpitaux  $\eta_h$  au profit du modèle à effets fixes (voir tableau 6bis). Nous commentons toutefois tout d'abord les résultats obtenus sur les deux spécifications, avant de privilégier par la suite le modèle à effets fixes.

Les coefficients estimés des caractéristiques observables des hôpitaux figurent dans le tableau 6. L'estimation du modèle à effets aléatoires montre que les coûts des CHR et des hôpitaux privés PSPH ne sont pas significativement différents de ceux des autres hôpitaux publics. Ce résultat est surprenant. Concernant les CHR, les experts de la Mission PMSI évaluent un surcoût de + 13 % lié aux activités d'enseignement et de recherche dans les hôpitaux français (Direction des Hôpitaux, 1996). Sur un échantillon d'hôpitaux espagnols, Lopez-Casasnovas et Saez (1999) estiment un surcoût significatif, de 9 % environ. Quant aux hôpitaux privés PSPH, une étude de la Fédération des établissements hospitaliers et associations privées à but non lucratif déclare un écart de 14 % sur les charges salariales des PSPH qui se traduirait par un alourdissement de 7 % de leurs budgets (Apparizio, Brocas et Moisdon, 1999).

En revanche, l'estimation du modèle à effets aléatoires conduit à un coefficient positif et significatif (431 €) pour la variable TECH (tableau 6, modèle A). Ce coefficient peut s'interpréter en calculant le surcoût relatif correspondant, par rapport au coût moyen du séjour de référence (2 902 €). Les coûts des hôpitaux techniques, c'est-à-dire de ceux qui ont fait les investissements nécessaires pour la réalisation de cathétérismes ou d'angioplasties, sont donc supérieurs de 14,8 %. Cet effet positif est observé lorsque

---

<sup>22</sup>Les résultats obtenus sont qualitativement similaires selon la spécification retenue en termes d'effets croisés pour les caractéristiques individuelles des patients. Le tableau 5 permet d'en avoir un résumé plus lisible.

l'on estime le modèle  $A$  mais devient négatif lorsque l'on introduit les taux moyens d'application des actes dans le modèle  $B$ . En réalité, l'effet négatif de la technicité est alors largement compensé, et au-delà, par les effets positifs des taux d'actes. Au total, être un hôpital technique conduit toujours à un surcoût<sup>23</sup>. Le résultat concernant l'absence d'impact significatif sur les coûts des activités d'enseignement et de recherche doit être interprété au regard de l'effet positif de cette variable TECH. En effet, tous les CHR de notre échantillon sont des hôpitaux techniques sur toute leur période d'observation (tableau 3). Le résultat obtenu signifie donc que des hôpitaux qui n'ont pas le statut de CHR mais pratiquent les activités innovantes ont aussi des coûts plus élevés : l'existence d'un surcoût est plus directement liée à la réalité d'une activité de pointe (TECH) qu'au statut.

L'estimation du modèle à effets aléatoires révèle aussi un effet positif de la variable TI : les hôpitaux dont une forte proportion de patients sont admis par transferts ont des coûts plus élevés.

Notre procédure d'estimation permet d'identifier, dans la variance des coûts non expliquée par les variables explicatives, les composantes attribuables au hasard moral transitoire et à l'hétérogénéité non observée. En effet, la méthode du maximum de vraisemblance permet d'estimer  $\sigma_\eta$ , écart-type des effets spécifiques hôpitaux  $\eta_h$ , lorsque ceux-ci sont supposés aléatoires. Pareillement, on estime  $\sigma_\varepsilon$ , écart-type de la perturbation  $\varepsilon_{h,t}$  qui peut être considérée comme un indicateur du hasard moral transitoire. L'influence de ce hasard moral sur la variance des coûts est loin d'être négligeable : l'estimation de l'écart-type correspondant (410 ou 429) représente 50 % environ de l'estimation de  $\sigma_\eta$  (897 ou 785).

Les résultats commentés ci-dessus doivent être relativisés car nos tests conduisent à conclure que les effets spécifiques hôpitaux  $\eta_h$  sont corrélés aux variables explicatives : l'estimateur du maximum de vraisemblance perd donc ses propriétés de convergence. Nous privilégions maintenant le modèle où les  $\eta_h$  sont supposés fixes. Ce modèle est estimé par les moindres carrés quasi-généralisés, procédure justifiée par le fait que le test d'Hausman d'indépendance des effets  $\varepsilon_{h,t}$  nous conduit à ne pas rejeter l'hypothèse nulle (voir tableau 6bis).

---

<sup>23</sup>En se basant sur le taux d'application des actes observé en moyenne pour les hôpitaux techniques, on obtient un surcoût de la technicité de 25,4 % environ pour le modèle  $B$ , évalué comme résultante des effets des variables TECH, TxCath, TxAngio et TxStent.

L'estimation du modèle à effets fixes ne permet pas d'identifier les coefficients des variables constantes dans le temps  $Q'_h$ . Par ailleurs, la variable TECH n'est plus significative, une fois que les effets fixes ont pris en compte les différences permanentes dans les coûts moyens par hôpital.

La spécification avec des effets fixes hôpitaux permet d'obtenir maintenant des estimations convergentes des termes  $\eta_h$  et  $\varepsilon_{h,t}$  et de leurs écarts-types respectifs. Tout d'abord, on constate que la corrélation entre les  $\hat{\eta}_h$  et les  $\hat{\varepsilon}_{h,t}$  est très faible (-0,001 pour les modèles A et B) et non significative. La valeur estimée pour  $\sigma_\varepsilon$  est très proche de celle qui a été obtenue par le maximum de vraisemblance : 399 ou 446 (modèle A ou B). Concernant  $\sigma_\eta$ , on trouve des valeurs plus élevées<sup>24</sup> de mille francs environ : 1 057 ou 993 (modèle A ou B). L'importance de la variabilité des coûts due au hasard moral transitoire est encore considérable :  $\sigma_\varepsilon$  représente environ 50 % de  $\sigma_\eta$ .

### Insérer ici graphiques 1 et 2

Pour avoir une idée de l'importance des écarts-type  $\sigma_\varepsilon$  et  $\sigma_\eta$ , on peut les rapporter à l'écart-type des coûts de nos séjours (4 198 €, pour un coût moyen de 2 863 €). On peut aussi, comme dans les graphiques 1 et 2, mettre en rapport les effets estimés  $\hat{\eta}_h$  et  $\hat{\varepsilon}_{h,t}$  avec les coûts moyens par hôpital  $C_{.,h.}$  ou par hôpital-année  $C_{.,h,t}$  correspondants<sup>25</sup>. Les observations sont rangées par ordre croissant de coûts moyens. On constate que les effets spécifiques hôpitaux sont en rapport avec la hiérarchie des coûts moyens, mais sont très loin de l'expliquer entièrement (graphique 1). Le graphique 2 montre bien la régularité des coûts, face aux fluctuations de l'aléa moral transitoire : face à une réalité fluctuante de cas traités, plus ou moins coûteux en moyenne par hôpital-année, la régularité des budgets autorise de nombreuses inefficacités.

## 4 Effet potentiel d'une tarification par pathologie sur les coûts de traitement

L'importance des inefficacités mises en évidence encourage l'instauration d'une tarification à la pathologie. Nous avons vu dans le paragraphe 2.3 que le

---

<sup>24</sup>Cet écart peut s'interpréter comme l'effet du statut de l'hôpital, qui bien que non significatif, était capté par les variables  $Q'_h$  dans le modèle à effets aléatoires et est nécessairement intégré aux  $\eta_h$  dans le modèle à effets fixes.

<sup>25</sup>Ces graphiques ont été réalisés pour le modèle A.

bon paiement doit se situer au niveau du coût correspondant à la production efficace.

#### 4.1 Paiements proposés : deux formules selon la prise en compte ou non des hétérogénéités entre hôpitaux

Considérons le modèle (3) à effets hôpitaux fixes :

$$C_{i,h,t} = X'_{i,h,t}\gamma_t + W'_{h,t}\alpha + a + c_t + \eta_h + \xi_{i,h,t}$$

avec  $\xi_{i,h,t} = -\varepsilon_{h,t} + u_{i,h,t}$ .

##### a) Première formule de paiement

Le coût est d'autant plus faible que l'effort de réduction du coût fourni par le gestionnaire de l'hôpital,  $\varepsilon_{h,t}$ , est maximal. Pour définir, à partir des données observables, un paiement qui réduise au maximum l'aléa moral transitoire, la tutelle-économètre peut retenir la formule suivante :

$$P^1_{i,h,t} = X'_{i,h,t}\hat{\gamma}_t + W'_{h,t}\hat{\alpha} + \hat{a} + \hat{c}_t + \hat{\eta}_h + \underset{h,t}{\text{Min}} \hat{\xi}_{.,h,t} \quad (5)$$

Les estimations sont ici obtenues par les MCQG appliqués au modèle (3), dont les résultats, commentés dans la section précédente, figurent dans le tableau 6.  $\hat{\xi}_{i,h,t}$  est un estimateur convergent de  $\xi_{i,h,t} = -\varepsilon_{h,t} + u_{i,h,t}$ . Soit les moyennes par hôpital définies par  $\hat{\xi}_{.,h,t} = \frac{1}{N_{h,t}} \sum_{i=1}^{N_{h,t}} \xi_{i,h,t}$ , où  $N_{h,t}$  est le nombre de séjours observés dans l'hôpital  $h$  en  $t$ . Calculer les moyennes par hôpital permet de définir le paiement comme l'espérance du coût efficace, en évitant de prendre en compte la distribution d'échantillonnage des séjours à l'intérieur de chaque hôpital. Comme  $\hat{u}_{.,h,t} \xrightarrow{P} 0$  quand  $N_{h,t}$  est grand<sup>26</sup>,  $\underset{h,t}{\text{Min}} \hat{\xi}_{.,h,t}$  est un estimateur convergent de  $\underset{h,t}{\text{Min}} (-\varepsilon_{h,t})$ , c'est à dire de la valeur correspondant à l'effort maximal de réduction du coût.

Par la tarification  $P^1$ , la tutelle tient compte des caractéristiques observables  $X'_{i,h,t}$  et  $W'_{h,t}$  et des hétérogénéités non observées constantes dans le

---

<sup>26</sup>En moyenne,  $N_{h,t}$  est égal à 77, avec un minimum de 19 et un maximum de 250.

temps  $\eta_h$ , quelle qu'en soit l'origine (aléa moral de long terme, antisélection ou qualité des soins). Cette tarification incite à l'efficacité dans l'activité hospitalière dans la mesure où elle ne rémunère pas les variations de coûts dues à l'aléa moral transitoire.

*b) Seconde formule de paiement*

La première formule de tarification que nous proposons peut être considérée comme laxiste. Plus exactement, elle peut être critiquée comme trop respectueuse des écarts  $\eta_h$ , dont on a vu qu'ils pouvaient recouvrir autant une meilleure qualité des soins que des avantages acquis en matière budgétaire permettant une moindre efficacité de façon permanente (aléa moral de long terme). On peut alors envisager une seconde formule de paiement qui tienne compte des caractéristiques observables des patients et de l'hôpital, mais "écrase" les hétérogénéités non observées  $\eta_h - \varepsilon_{h,t}$  :

$$P_{i,h,t}^2 = X'_{i,h,t} \hat{\gamma}_t + W'_{h,t} \hat{\alpha} + \hat{a} + \hat{c}_t + \overset{1/2}{\underset{3/4}{\text{Min}}}_{h,t} \hat{\eta}_h + \hat{\xi}_{.,h,t} \quad (6)$$

Calculer le paiement en utilisant  $\overset{1/2}{\underset{3/4}{\text{Min}}}_{h,t} \hat{\eta}_h + \hat{\xi}_{.,h,t}$  revient pour la tutelle à se caler sur l'hôpital dont la somme de la caractéristique inobservable et de la perturbation représentative de l'aléa moral transitoire est minimale. En effet,  $\hat{\xi}_{.,h,t}$  étant un estimateur convergent de  $-\varepsilon_{h,t}$ , on a :

$$\overset{1/2}{\underset{3/4}{\text{Min}}}_{h,t} \hat{\eta}_h + \hat{\xi}_{.,h,t} \xrightarrow{P} \underset{h,t}{\text{Min}} \{ \eta_h - \varepsilon_{h,t} \}.$$

Appliquer cette seconde forme de paiement revient, pour la tutelle à supposer que toute l'hétérogénéité entre les établissements correspond à de l'aléa moral.

*c) Poids du rétrospectif dans la première formule de paiement  $P^1$*

On a vu que l'hétérogénéité des patients expliquait plus de 50 % de la variance des coûts. Ne pas la prendre en compte dans les formules de tarification serait une incitation à la sélection des patients par les hôpitaux. Les formules proposées limitent ce risque grâce à une prise en compte des caractéristiques individuelles des patients. Cette prise en compte correspond à un paiement prospectif à travers une grille de tarification de la forme : "pour un homme,

dans telle tranche d'âge, vous recevrez X € en supplément". Dans cette perspective, la formule  $P^2$  peut être considérée comme un paiement prospectif, amendé par la forme d'ajustement au risque que constitue la prise en compte des hétérogénéités observées des patients. En revanche, la formule  $P^1$  comporte une part de paiement rétrospectif dans la mesure où l'on rémunère les écarts de coûts dus à  $\eta_h$ .

Au total on peut, selon ces critères, distinguer dans la formule  $P^1$  les éléments prospectifs et rétrospectifs de la façon suivante :

$$P_{i,h,t}^1 = \underbrace{X'_{i,h,t}\hat{\gamma}_t + W'_{h,t}\hat{\alpha} + \hat{a} + \hat{c}_t}_{\text{Prospectif}=F_{i,h,t}} + \underbrace{Min_{h,t}\hat{\xi}_{.,h,t}}_{\text{Rétrospectif}} + \underbrace{\{\hat{\eta}_h\}}_{\text{Rétrospectif}} .$$

Considérons l'expression classique des paiements mixtes comme une moyenne pondérée du forfait  $F$  et du remboursement au coût observé  $C$  :  $P = \mu F + (1 - \mu)C$ . A partir du calcul des  $P_{i,h,t}^1$  et de la décomposition effectuée ci-dessus, on peut calculer  $\mu_{i,h,t} = \frac{P_{i,h,t}^1 - C_{i,h,t}}{F_{i,h,t} - C_{i,h,t}}$  et évaluer la moyenne  $\hat{\mu}$  au niveau de l'échantillon. On obtient (pour le modèle A) :  $\hat{\mu} = 44,7\%$ , avec un écart-type de 12,8 %. Soulignons que cette évaluation n'est en aucune manière une formule de paiement. Elle est effectuée *a posteriori* à partir des simulations de paiements réalisées sur notre échantillon. Elle permet de connaître, à titre indicatif seulement, l'ordre de grandeur de la prise en compte du rétrospectif que représenterait un paiement défini selon la formule  $P^1$  que nous proposons (équation 5), qui incorpore une rémunération des écarts  $\eta_h$ .

## 4.2 Simulation des économies budgétaires potentielles

Les estimations permettent de simuler l'application des tarifications  $P^1$  et  $P^2$  proposées. La tarification  $P^1$  est plus souple qu'une tarification de type  $P^2$ , laquelle est basée sur l'hypothèse que toute hétérogénéité entre les établissements correspond à de l'aléa moral. En effet,  $P^2$  "écrase" toutes les hétérogénéités non observées  $\eta_h - \varepsilon_{h,t}$ , alors que  $\eta_h$  peut correspondre pour partie à de "bonnes" caractéristiques, une certaine qualité des soins par exemple. Avec la tarification  $P^1$ , la tutelle tient compte des hétérogénéités non observées constantes dans le temps  $\eta_h$ , qu'elles soient dues à une mauvaise gestion ou à une qualité particulière des soins. Cette tarification reste incitative car elle ne rémunère pas les variations de coûts moyens hospitaliers dus à l'aléa moral purement transitoire  $\varepsilon_{h,t}$ . Comme nous l'avons vu

dans le paragraphe précédent,  $P^1$  peut être considéré comme un paiement mixte qui intègre les hétérogénéités entre établissements. Du point de vue du poids accordé au paiement des hétérogénéités entre établissements, la tarification optimale est vraisemblablement située entre  $P^2$  (trop homogène) et  $P^1$  (laxiste puisqu'elle rémunère toutes les hétérogénéités non observées constantes dans le temps).

Dans le tableau 7 figurent les économies budgétaires que l'on peut espérer de la mise en oeuvre des tarifications proposées. Ces simulations sont réalisées à niveau d'activité et à comportements constants<sup>27</sup>. Nous supposons que les hôpitaux rejoignent leurs coûts de production efficaces<sup>28</sup>, définis par les paiements. Les économies budgétaires *globales* sont mesurées par l'écart

relatif entre le total des coûts et le total des paiements :  $ebg = \frac{\sum_{i,h,t} (C_{i,h,t} - P_{i,h,t})}{\sum_{i,h,t} C_{i,h,t}}$ .

Nous avons aussi calculé dans le tableau 7 les moyennes et écarts-type des économies budgétaires par hôpital-année,  $eb_{h,t} = \frac{C_{h,t} - P_{h,t}}{C_{h,t}}$ . Nous commentons essentiellement les valeurs des économies budgétaires globales, lesquelles sont affectées par les tailles des établissements, mais le lecteur peut constater aisément que ces valeurs sont proches des moyennes des économies budgétaires par hôpital-année.

On constate que l'intervalle constitué par les paiements  $P^1$  et  $P^2$  est relativement large : la tarification  $P^1$  conduit à des économies potentielles de 16 % environ ; la tarification  $P^2$  conduit à des gains budgétaires de 46 ou 42 %, selon le modèle considéré ( $A$  ou  $B$ ).

$P^1$  est la tarification la plus "laxiste", puisqu'elle rémunère toutes les hétérogénéités entre les établissements : les différences de qualité mais aussi les écarts de coûts qui seraient dus à des inefficacités permanentes de gestion. Elle conduit quand même à un résultat non négligeable en matière d'économie budgétaire (16 %) car elle est suffisamment incitative pour réduire une grande part de l'aléa moral. Cette stratégie de tarification ( $P^1$ ) doit être re-

---

<sup>27</sup>On omet donc d'intégrer dans les simulations d'éventuelles réactions stratégiques des hôpitaux à la nouvelle tarification.

<sup>28</sup>Nous supposons ainsi que le paiement proposé conduit les hôpitaux à faire les efforts nécessaires pour ajuster leur coûts au niveau de la tarification. Ceci doit être vérifié sous deux conditions : (i) des déficits doivent constituer une menace crédible de fermeture d'un établissement ; (ii) il n'y a pas de comportement stratégique consistant à sélectionner de la clientèle ou à ajuster la qualité. La prise en compte dans les paiements des hétérogénéités des patients et des hôpitaux joue en faveur de la vérification de la condition (ii).

commandée. Prudente, elle permet d'éviter de niveler les paiements sur un établissement dont la qualité des prestations (variable difficilement vérifiable par la tutelle) serait la plus mauvaise.

Nous avons interprété l'aléa moral transitoire  $\varepsilon_{h,t}$  comme reflétant l'effort fourni par le gestionnaire de l'hôpital en matière de réduction des coûts, en se référant à la plus ou moins grande rigueur dans la négociation des marchés pour les consommables et les prix des activités externalisées. L'ordre de grandeur de l'économie budgétaire simulée, 16 %, peut-il être considéré comme excessif, au regard de l'interprétation retenue pour  $\varepsilon_{h,t}$ ? En aucune manière : le prix des consommables est un élément très important des coûts de traitement. Pour prendre un exemple, la réalisation d'une angioplastie entraîne un surcoût de 60 % qui est presque exclusivement dû au matériel utilisé.

Une fois la formule de tarification choisie ( $P^1$ ), quel modèle faut-il retenir ?

- Pour éviter les biais d'estimation, il convenait d'intégrer la durée de séjour dans les spécifications considérées. En revanche, il est impératif que le paiement mis en place soit le plus prospectif possible : il doit être établi par séjour et non par jour d'hospitalisation. De ce fait, nous proposons de rembourser la durée de séjour sur la base de son coefficient estimé dans la fonction de coût, multiplié par un indicateur adéquat de la durée de séjour moyenne, pour le type de patient et la pathologie considérés.

- Quelle liste des caractéristiques hospitalières  $W'_{h,t}$  faut-il retenir entre les options représentées par les modèles  $A$  et  $B$ ? La différence essentielle entre ces deux possibilités réside dans le fait que le modèle  $B$  incorpore des caractéristiques manipulables par les établissements comme le taux d'actes techniques. L'intérêt d'incorporer les taux d'actes dans les paiements est d'éviter les stratégies de sélection de patients ou de discrimination. *A contrario*, le risque encouru est d'introduire une forme de paiement à l'acte encourageant une pratique excessive des actes. Or on constate que les taux d'actes ne sont pas significatifs dans le modèle à effets fixes et que les différences entre les économies budgétaires obtenues par l'application des tarifications aux modèles  $A$  et  $B$  sont négligeables quelle que soit la formule de paiement retenue ( $P^1$  ou  $P^2$ ). Ces évaluations sont obtenues sur une période où les hôpitaux sont principalement régulés par le système du budget global, et où des considérations d'ordre financier doivent peu intervenir sur les décisions de mise en oeuvre des actes. Dans ce contexte, il apparaît que des paiements calculés sur la base du modèle  $A$ , qui intègre la variable TECH mais non les taux

d'actes, sont aussi proches des coûts que des paiements calculés à partir de TECH et des taux d'actes (modèle  $B$ ).

Ce résultat suggère de calculer les paiements à partir du modèle  $A$  qui considère la seule variable TECH. Cet indicateur présente l'intérêt d'être une caractéristique observable : la tutelle a les moyens de savoir si l'hôpital a la capacité de pratiquer des cathétérismes et angioplasties ; l'établissement n'a aucun intérêt à gonfler abusivement son taux d'actes pour obtenir un supplément de paiement ; sur une période *a priori* sans manipulation sur les taux d'actes, les paiements sont aussi proches des coûts que l'on inclue ou non les taux d'actes dans leur calcul.

Dans le tableau 8 figurent des corrélations entre les coûts et les paiements qui confirment nos commentaires. Une corrélation élevée signale que l'incitation à la sélection de patients est limitée. On constate que les économies budgétaires non négligeables calculées dans le tableau 7 vont de pair avec des corrélations très élevées.

### 4.3 Conclusion

Un système de tarification hospitalier pertinent doit impérativement prendre en compte l'hétérogénéité entre les établissements. Les inconvénients d'un système purement prospectif sont connus : sélection des patients, baisse de la qualité des soins. Pour les éviter, de nombreux auteurs ont préconisé l'utilisation d'une tarification mixte, qui combinerait forfait et remboursement du coût observé. Dans cet article, on adopte une approche économétrique pour définir une tarification qui tienne compte de l'hétérogénéité des hôpitaux.

Celle-ci est prise en compte par des caractéristiques observables des établissements et des effets spécifiques hôpitaux. Ces derniers peuvent être interprétés comme les résultantes de trois effets : la sélection adverse, le hasard moral de long terme et les différences permanentes dans la qualité des soins. On obtient deux formules de paiements possibles. La première formule de paiement prend en compte toute l'hétérogénéité non observée entre les hôpitaux, pourvu qu'elle soit constante dans le temps. La seconde ignore toute hétérogénéité non observée.

La première formule de paiement semble préférable : elle présente le grand avantage de procurer un surcroît de financement aux hôpitaux délivrant des soins de haute qualité. Par ailleurs, nous montrons qu'elle est susceptible de

conduire à des économies budgétaires substantielles, car elle est suffisamment incitative pour réduire la partie - non négligeable - des coûts attribuable au hasard moral transitoire. Ces économies sont obtenues par une pression ne jouant que sur le gestionnaire de l'hôpital. En principe, le gain est acquis sans contrainte sur les patients ni sur le personnel soignant.

Cette tarification est facile à mettre en oeuvre, dès lors que la tutelle a accès à l'information sur les coûts hospitaliers. A cet égard les économies budgétaires que nos simulations mettent en évidence doivent être mise en regard avec les coûts de gestion d'un système de tarification à la pathologie, notamment le coût du système d'information qu'il nécessite. Mais il faut bien comprendre qu'en France, ce système d'information est déjà installé (le PMSI) et qu'il n'y aurait qu'un coût marginal à le compléter pour qu'il contienne des informations sur les coûts par séjours.

Un inconvénient de ce système est qu'il conduit à attribuer des financements plus élevés à des hôpitaux qui sont plus coûteux de façon permanente à cause d'une mauvaise gestion. Notre méthode permet d'identifier la composante de la variabilité des coûts due au hasard moral transitoire. Mais nous ne pouvons pas discriminer entre les composantes désirables et indésirables de l'hétérogénéité non observée. De ce fait, la tutelle doit choisir entre un paiement qui ignore toute l'hétérogénéité non observée et un paiement qui finance toute l'hétérogénéité non observée. Le choix entre ces deux paiements dépend des poids accordés à l'efficacité et à la qualité des soins dans la fonction d'utilité sociale. Soulignons enfin que les économies budgétaires obtenues résultent d'une simulation à niveau d'activité constant. Or, à la différence d'une régulation par budget global, le volume d'activité n'est pas borné dans la tarification par pathologie. L'aléa budgétaire serait ainsi en partie transféré sur la tutelle, ce qui renforce l'impératif de l'instauration d'une tarification incitant les établissements à fonctionner de façon efficace.

## 5 Bibliographie

Apparitio, S., Brocas, A.-M. et Moïsdon, J.-C. (1999), "La place du PSI dans l'allocation des ressources en île-de-France", *Agence Régionale d'Hospitalisation d'île-de-France - Rapport technique*

Antweiler, W. (2001), "Nested random effects estimation in unbalanced panel data", *Journal of Econometrics* vol 101 : pp 295-313

Auriol, E. et Laffont, J.J. (1992), "Regulation by duopoly", *Journal of*

*Economics and Management Strategy*, vol 1, n°3 : pp 507-533

Baltagi, B. H. (1981), "Simultaneous equations with error components", *Journal of Econometric* vol 17 : pp 189-200

Baltagi, B. H. (1995), "*Economic Analysis of Panel Data*", John Wiley and Sons, Chichester

Baltagi, B. H., Song, S. H. et Jung, B. C. (2001), "The unbalanced nested error component regression model", *Journal of Econometrics*, vol 101 : pp 357-381

Direction des Hôpitaux de Paris, mission PMSI. (1996), "Le PMSI, analyse médico-économique de l'activité hospitalière", *La lettre d'information hospitalières*, Numéro spécial

Dormont B,(1989) "Introduction à l'économétrie des données de panel", *Monographies d'Econométrie, ADRES, CNRS*.

Ellis, R. P. (1998), "Creaming, Dumping, skimping : Provider competition on the intensive and extensive margins", *Journal of Health Economics*, vol 17 : pp 537-555

Keeler E. B. (1990) "What proportion of hospital cost differences is justifiable?", *Journal of Health Economics* 9(3), 359-365.

Laffont J. J. et Tirole J. (1993), "*A theory of incentives in procurement and regulation*", MIT Press

Linna M. (1998), "Measuring hospital cost efficiency with panel data models", *Health Economics*, 7 : pp 415-427.

Lopez-Casasnovas, G. et Saez, M. (1999), "The impact of teaching Status on Average Costs in Spanish Hospitals", *Health Economics*, vol 8, n°7 : pp 641-651

Ma, A. C. T. (1994), "Health care payment systems : cost and quality incentives", *Journal of Economics and Management Strategy*, vol 3, n°1 : pp 93-112

Ma, A. C. T. (1998), "Health care payment systems : cost and quality incentives- Reply", *Journal of Economics and Management Strategy*, vol 7, n°1 : pp 139-142

McClellan M. (1997) "Hospital reimbursement incentives : an empirical analysis", *Journal of Economics and Management Strategy*, 6(1) : 91-128.

Milcent C. (2001) "Tarification par pathologie, hétérogénéité des hôpitaux et innovations techniques", Thèse de doctorat en Sciences Economiques, Université de Paris X.

Moisson J.C et Tonneau D (1997) "Gestion externe et gestion interne du système hospitalier français. Tendances actuelles", Dix ans d'avancées en

économie de la santé, Actes des XIXes Journées des Economistes de la Santé Français. Ed : John Libbey.

Mougeot M.(1999<sub>a</sub>) “Régulation du système de santé”, *Conseil d’Analyse Economique*. La documentation Française, Paris.

Mougeot M.(1999<sub>b</sub>) ” La fonction de préférence de l’Etat. Le cas de l’assurance maladie en France ”, *Revue Economique*, 361-382.

Mougeot M. et Naegelen F. (1997), “La réglementation hospitalière : tarification par pathologie ou achat de soins ? ”, *Economie et prévision*, n°129-130

Newhouse J. P. (1996) ” Reimbursing health plans and health providers : efficiency in production versus selection ”, *Journal of Economic Literature*, Vol. XXXIV : 1236-1263.

Pope,G. (1990), “Using hospital-specific costs to improve the fairness of prospective reimbursement”, *Journal of Health Economics*, vol 9, n°3 : pp 237-251

Shleifer, A. (1985), “A theory of Yardstick Competition”, *Rand Journal of Economics*, vol 16 : pp 319-327

Wagstaff, A. (1989), “Estimating efficiency in the hospital sector : A comparaison on three statistical cost frontier models”, *Applied Economics*, vol 21 : pp 659-672

## 0.1 Annexe : estimation d'un modèle à erreurs composées sur des données de panel stratifiées à trois dimensions

Considérons des données stratifiées à trois dimensions. Les observations sont de la forme  $y_{i,h,t}$ , où  $i$  désigne le séjour hospitalier (ou l'individu, ou encore l'entreprise) observé dans l'hôpital (ou le secteur)  $h$  lors de l'année  $t$ . On peut modéliser le phénomène étudié à l'aide d'une spécification à erreurs composées<sup>1</sup> :

$$y_{i,h,t} = X'_{i,h,t} b + \underbrace{\eta_h + \varepsilon_{h,t} + u_{i,h,t}}_{v_{i,h,t}}$$

Les composantes de la perturbation  $v_{i,h,t}$  sont des perturbations aléatoires  $\eta_h$ ,  $\varepsilon_{h,t}$  et  $u_{i,h,t}$ , non corrélées entre elles, dont les variances respectives sont  $\sigma_\eta^2$ ,  $\sigma_\varepsilon^2$  et  $\sigma_u^2$ . On a :  $E(v_{i,h,t}v_{i^0,h^0,t^0}) = \delta_{hh^0}\sigma_\eta^2 + \delta_{hh^0}\delta_{tt^0}\sigma_\varepsilon^2 + \delta_{hh^0}\delta_{tt^0}\delta_{ii^0}\sigma_u^2$ , avec  $\delta_{hh^0} = 1$  si  $h = h^0$  et  $\delta_{hh^0} = 0$  si  $h \neq h^0$ .  $\delta_{tt^0}$  et  $\delta_{ii^0}$  sont définis de façon analogue.  $X'_{i,h,t}$  est le vecteur des observations de  $k$  variables explicatives supposées indépendantes de  $v_{i,h,t}$ , pour tout  $i, h$  ou  $t$ . On note  $\mathbb{P}$  le nombre total d'observations disponibles.

### 1. Données cylindrées dans chaque dimension

Dans ce cas, on a :

$$\begin{aligned} i &= 1, \dots, N \\ t &= 1, \dots, T \\ h &= 1, \dots, H \\ \mathbb{P} &= HNT. \end{aligned}$$

On observe le même nombre de séjours  $N$  chaque année dans chaque hôpital et chaque hôpital est observé pendant le même nombre d'années  $T$ . L'application des moindres carrés généralisés dans ce cas est décrite dans Baltagi (1995), dont nous nous inspirons.

Soit  $y$  le vecteur de format  $(HNT, 1)$  comprenant les observations  $y_{i,h,t}$  empilées dans l'ordre hôpital, année, individus et  $v$  le vecteur correspondant

---

<sup>1</sup>Sous une forme plus générale, ce modèle pourrait comporter un effet temporel aléatoire. On suppose que les observations dans la dimension temporelle sont suffisamment nombreuses pour que les propriétés statistiques des estimations ne soient pas affectées par une spécification de l'effet temporel en termes d'effets fixes.

des perturbations, de format  $(HNT, 1)$ . Le vecteur  $b$  des paramètres à estimer est de format  $(k, 1)$ . Le modèle à estimer est de la forme  $y = Xb + v$ , avec  $v = (v'_1 v'_2 v'_3 \dots v'_h \dots v'_H)'$ . On note  $\otimes$  le produit de kronecker,  $I_N$  désigne la matrice identité de format  $N$ ,  $J_N$  la matrice carrée de format  $N$  dont tous les éléments sont des 1,  $\bar{J}_N$  la matrice  $\frac{J_N}{N}$  et  $E_N$  la matrice  $I_N - \bar{J}_N$ .

Soit  $\Omega = E(vv')$  la matrice de variance-covariance des perturbations du modèle. On a  $\underset{(HNT, HNT)}{\Omega} = I_H \otimes \Sigma$ , avec :

$$\underset{(NT, NT)}{\Sigma} = E(v_h v'_h) = \sigma_\eta^2 J_{NT} + \sigma_\varepsilon^2 (I_T \otimes J_N) + \sigma_u^2 I_{NT}.$$

La décomposition spectrale de  $\Sigma$  permet de calculer  $\Omega^{-\frac{1}{2}}$ . On a :

$$\underset{(NT, NT)}{\Sigma} = \lambda_1 Q_1 + \lambda_2 Q_2 + \lambda_3 Q_3$$

avec 
$$\begin{cases} \lambda_1 = \sigma_u^2 & \text{et } Q_1 = I_T \otimes E_N \\ \lambda_2 = T\sigma_\varepsilon^2 + \sigma_u^2 & \text{et } Q_2 = E_T \otimes \bar{J}_N \\ \lambda_3 = NT\sigma_\eta^2 + T\sigma_\varepsilon^2 + \sigma_u^2 & \text{et } Q_3 = \bar{J}_T \otimes \bar{J}_N \end{cases}$$

On en déduit :

$$\sigma_u \Omega^{-\frac{1}{2}} = I_H \otimes \Sigma^{-\frac{1}{2}} = I_{HNT} - \left(1 - \frac{\sigma_u}{\sqrt{\lambda_2}}\right) Q_2 - \left(\frac{\sigma_u}{\sqrt{\lambda_2}} - \frac{\sigma_u}{\sqrt{\lambda_3}}\right) Q_3.$$

Appliquer les MCG revient alors à appliquer les MCO au modèle transformé  $y_{i,h,t}^* = X_{i,h,t}^* b + v_{i,h,t}^*$  avec :

$$y_{i,h,t}^* = y_{i,h,t} - \left(1 - \frac{\sigma_u}{\sqrt{\lambda_2}}\right) y_{h,t} - \left(\frac{\sigma_u}{\sqrt{\lambda_2}} - \frac{\sigma_u}{\sqrt{\lambda_3}}\right) y_{..h}$$

## 2. Données cylindriques sur une dimension seulement

Reprenons l'exemple des données hospitalières. Les données sont cylindriques sur une dimension lorsque, pour un hôpital donné, le nombre de séjours observés est identique quelle que soit l'année. En revanche, le nombre de séjours observés diffère suivant l'hôpital et les hôpitaux sont observés sur des périodes de longueurs différentes. Nous reprenons ici l'article de Baltagi, Song et Jung (2001). Les données sont triplement indicées, avec :

$$\begin{aligned}
i &= 1, \dots, N_h \\
t &= 1, \dots, T_h \\
h &= 1, \dots, H \\
\phi_h &= N_h T_h \\
\mathbb{P} &= \sum_{h=1}^H \phi_h
\end{aligned}$$

$\Omega = E(vv') = \text{diag} [\Sigma_1, \dots, \Sigma_h, \dots, \Sigma_H]$ , où  $\Sigma_h = E(v_h v_h')$ . Par la suite, nous noterons pour simplifier  $\Omega = \text{diag} [\Sigma_h]$ . Puisque  $\Sigma_h$  est de format  $(\phi_h, \phi_h)$ ,  $\Omega$  est une matrice bloc-diagonale dont les blocs ont des formats différents. De ce fait, on ne peut plus utiliser le produit de Kronecker pour formaliser simplement l'expression de  $\Omega$ . Cependant, l'approche adoptée dans le paragraphe 1 peut s'appliquer à chaque bloc de  $\Omega$ .

On a en effet :

$$\begin{aligned}
\Sigma_h &= \sigma_\eta^2 (J_{T_h} \otimes J_{N_h}) + \sigma_\varepsilon^2 (I_{T_h} \otimes J_{N_h}) + \sigma_u^2 (I_{T_h} \otimes I_{N_h}) \\
&= N_h T_h \sigma_\eta^2 (\overline{J_{T_h}} \otimes \overline{J_{N_h}}) + N_h \sigma_\varepsilon^2 (I_{T_h} \otimes \overline{J_{N_h}}) + \sigma_u^2 (I_{T_h} \otimes I_{N_h})
\end{aligned}$$

En remplaçant comme précédemment,  $I_{T_h}$  par  $E_{T_h} + \overline{J_{T_h}}$  et  $I_{N_h}$  par  $E_{N_h} + \overline{J_{N_h}}$ , on obtient la décomposition spectrale de  $\Sigma_h$  :

$$\begin{aligned}
\Sigma_h &= \lambda_{1h} Q_{1h} + \lambda_{2h} Q_{2h} + \lambda_{3h} Q_{3h} \\
\text{avec } \begin{cases} \lambda_{1h} = \sigma_u^2 & \text{et } Q_{1h} = I_{T_h} \otimes E_{N_h} \\ \lambda_{2h} = T_h \sigma_\varepsilon^2 + \sigma_u^2 & \text{et } Q_{2h} = E_{T_h} \otimes \overline{J_{N_h}} \\ \lambda_{3h} = N_h T_h \sigma_\eta^2 + T_h \sigma_\varepsilon^2 + \sigma_u^2 & \text{et } Q_{3h} = \overline{J_{T_h}} \otimes \overline{J_{N_h}} \end{cases} .
\end{aligned}$$

Chaque matrice  $Q_{ph}$ ,  $p = 1, 2, 3$ , est symétrique et idempotente. Les matrices  $Q_{ph}$  sont orthogonales entre elles et leur somme est égale à la matrice identité. La décomposition spectrale permet d'écrire :

$$\sigma_u \Omega^{-\frac{1}{2}} = \sigma_u \text{diag} \left[ \Sigma_h^{-\frac{1}{2}} \right] = \text{diag} \left[ \frac{\sigma_u}{\sqrt{\lambda_{1h}}} Q_{1h} + \frac{\sigma_u}{\sqrt{\lambda_{2h}}} Q_{2h} + \frac{\sigma_u}{\sqrt{\lambda_{3h}}} Q_{3h} \right].$$

On en déduit :

$$\sigma_u \Omega^{-\frac{1}{2}} = \text{diag} [I_{T_h} \otimes I_{N_h}] - \text{diag} \left( 1 - \frac{\sigma_u}{\sqrt{\lambda_{2h}}} \right) [I_{T_h} \otimes \overline{J_{N_h}}] - \left( \frac{\sigma_u}{\sqrt{\lambda_{2h}}} - \frac{\sigma_u}{\sqrt{\lambda_{3h}}} \right) [\overline{J_{T_h}} \otimes \overline{J_{N_h}}].$$

On retombe ainsi sur un résultat très proche de celui établi précédemment : estimer le modèle par les moindres carrés généralisés revient à appliquer les MCO au modèle transformé  $y_{i,h,t}^* = X_{i,h,t}^* b + v_{i,h,t}^*$ .

$y_{i,h,t}^*$  est défini par :

$$y_{i,h,t}^* = y_{i,h,t} - \left(1 - \frac{\sigma_u}{\sqrt{\lambda_{2h}}}\right)y_{.h,t} - \left(\frac{\sigma_u}{\sqrt{\lambda_{2h}}} - \frac{\sigma_u}{\sqrt{\lambda_{3h}}}\right)y_{..h}.$$

La même transformation est appliquée pour obtenir  $X^*$ .

### 3. Données non cylindrées sur chaque dimension

On suppose maintenant que les données sont non cylindrées dans chacune des dimensions. Dans l'exemple des données hospitalières, cela correspond au cas où le nombre de séjours diffère suivant l'hôpital et l'année d'observation et où les hôpitaux sont observés sur des périodes différentes. Ce cas a été considéré par Antweiler (2001), dont nous reprenons ici les principaux développements. Les données sont triplement indicées, avec :

$$\begin{aligned} i &= 1, \dots, N_{h,t} \\ t &= 1, \dots, T_h \\ h &= 1, \dots, H \\ \phi_h &= \sum_{t=1}^{t=T_h} N_{h,t} \\ \mathbf{P} &= \sum_{h=1}^H \phi_h \end{aligned}$$

Comme précédemment, on a  $\Omega = E(vv') = \text{diag}[\Sigma_1, \dots, \Sigma_h, \dots, \Sigma_H]$ , où  $\Sigma_h = E(v_h v_h')$  est de format  $(\phi_h, \phi_h)$ . Mais, à la différence du cas précédent,  $\Sigma_h$  ne peut être formalisée simplement à l'aide de produits de Kronecker. En effet,  $\Sigma_h = E(v_h v_h')$  s'écrit :

$$\Sigma_h = \begin{pmatrix} \boxed{\begin{matrix} \Gamma_{h1} \\ (N_{h1}, N_{h1}) \end{matrix}} & \sigma_\eta^2 & \dots & \dots & \sigma_\eta^2 \\ \sigma_\eta^2 & \boxed{\Gamma_{h2}} & & & \\ \vdots & & & & \\ \cdot & & & & \\ \cdot & & & & \\ \cdot & & & & \\ \sigma_\eta^2 & \sigma_\eta^2 & \dots & \dots & \boxed{\begin{matrix} \Gamma_{hT_h} \\ (N_{hT_h}, N_{hT_h}) \end{matrix}} \end{pmatrix}$$

$$\text{avec : } \Gamma_{ht} = (\sigma_\eta^2 + \sigma_\varepsilon^2) J_{N_{h,t}} + \sigma_u^2 I_{N_{h,t}}$$

Dans ce cas, Antweiler (2001) montre qu'il n'est pas possible de définir une transformation simple des données, qui permette d'obtenir l'estimateur des MCG via l'application des MCO à des données transformées. Il définit alors un estimateur du maximum de vraisemblance basé sur l'hypothèse de normalité des perturbations. Si cette hypothèse est justifiée, l'estimateur du maximum de vraisemblance permet d'obtenir une estimation convergente et asymptotiquement efficace. Un autre intérêt de cet estimateur est qu'il permet d'obtenir une estimation directe des variances des composantes de la perturbation, avec de bonnes performances même lorsque le degré de "non-cylindrage" des données est très fort (Antweiler, 2001).

L'expression de la log-vraisemblance est la suivante :

$$L = -\frac{1}{2} \left[ \ln(2\pi\sigma_u^2) + \sum_{h=1}^H \left\{ \ln \theta_h + \sum_{t=1}^{T_h} \left\{ \ln \theta_{h,t} + \frac{\sum_{i=1}^{N_{ht}} v_{i,h,t}^2}{\sigma_u^2} - \frac{\rho_\varepsilon}{\theta_{h,t}} \frac{\left( \sum_{i=1}^{N_{ht}} v_{i,h,t} \right)^2}{\sigma_u^2} \right\} \right\} - \frac{\rho_\eta}{\theta_h} \frac{\left( \sum_{t=1}^{T_h} \sum_{i=1}^{N_{ht}} v_{i,h,t} \right)^2}{\sigma_u^2} \right] \quad (1)$$

$$\text{avec } \rho_\eta = \frac{\sigma_\eta^2}{\sigma_u^2}, \rho_\varepsilon = \frac{\sigma_\varepsilon^2}{\sigma_u^2}, \theta_h = 1 + \rho_\eta \phi_h \text{ et } \theta_{h,t} = 1 + \rho_\varepsilon N_{ht}.$$

Sans démontrer ces résultats, nous donnons dans ce qui suit des éléments permettant de les comprendre intuitivement.

Antweiler (2001) définit un type de matrice appelé “group membership”. Dans notre cas, on note  $G_\varepsilon$  et  $G_\eta$ , des matrices blocs-diagonales (avec des blocs de formats variables), qui ont pour éléments des 1 et des 0 et sont de format  $((\mathbb{P}, \mathbb{P}))$  identique à celui de  $\Omega$ .  $G_\varepsilon$  est définie de façon à permettre de localiser les “places” de  $\Omega$  où figurent les termes  $\sigma_\varepsilon^2$ , et  $G_\eta$  les “places” de  $\Omega$  où figurent les termes  $\sigma_\eta^2$ .

On a ainsi :

$$G_{\varepsilon} = \begin{pmatrix} \boxed{J_{N_{11}}} & 0 & \dots & \dots & 0 \\ 0 & \boxed{J_{N_{12}}} & & & \\ \vdots & & \ddots & & \\ \vdots & & & \ddots & \\ 0 & & & & \boxed{J_{N_{HTH}}} \end{pmatrix}$$

et :

$$G_{\eta} = \begin{pmatrix} \boxed{J_{\phi_1}} & 0 & \dots & \dots & \\ 0 & \boxed{J_{\phi_2}} & & & \\ \vdots & & \ddots & & \\ \vdots & & & \ddots & \\ \vdots & & & & \boxed{J_{\phi_H}} \end{pmatrix}$$

Grâce à la définition des matrices  $G_\eta$  et  $G_\varepsilon$ , on peut écrire  $\Omega$  sous la forme :

$$\Omega = \sigma_u^2 \left[ I_{\mathbb{P}} + \frac{\sigma_\varepsilon^2}{\sigma_u^2} G_\varepsilon + \frac{\sigma_\eta^2}{\sigma_u^2} G_\eta \right]$$

Pour trouver l'inverse de  $\Omega$ , on peut définir deux nouvelles matrices :  $Z_\varepsilon = I_{\mathbb{P}} + \rho_\varepsilon G_\varepsilon$  et  $Z_\eta = I_{\mathbb{P}} - (Z_\varepsilon)^{-1} \rho_\eta G_\eta$ , telles que  $\Omega = \sigma_u^2 Z_\varepsilon Z_\eta$ .  $Z_\varepsilon$  et  $Z_\eta$  étant inversibles, on a :

$$\Omega^{-1} = \frac{Z_\eta^{-1} Z_\varepsilon^{-1}}{\sigma_u^2}$$

Par ailleurs, on a :  $|\Omega| = (\sigma_u^2)^{\mathbb{P}} |Z_\varepsilon| |Z_\eta|$ , d'où l'on peut déduire :

$$|\Omega| = (\sigma_u^2)^{\mathbb{P}} \prod_{h=1}^H \theta_h \prod_{t=1}^{T_h} \theta_{h,t}$$

Quelques arrangements permettent alors d'aboutir à l'écriture de la log-vraisemblance  $L = -\frac{1}{2} [\mathbb{P} \ln(2\pi) + \ln |\Omega| + v' \Omega^{-1} v]$  sous la forme de l'équation (1).