1 Introduzione

Nel corso dell'ultimo decennio, i sistemi sanitari di molti paesi sono stati attraversati da ampi processi di riforma che hanno mirato ad aumentare il grado di competizione tra fornitori, come pure tra fornitori ed acquirenti, così da creare condizioni più vicine possibili a quelle che si riscontrano entro i mercati di concorrenza perfetta. Per incrementare la competizione entro sistemi generalmente caratterizzati dalla presenza di monopolisti pubblici verticalmente integrati, sono state individuate diverse strade. In alcune circostanze si è tentato di "spezzare" il monopolista pubblico creando diverse unità (sempre pubbliche) di domanda e di offerta, così da generare gli incentivi adatti a riprodurre il funzionamento di un mercato; in altre circostanze, ma talora congiuntamente, il mercato è stato aperto a diversi fornitori, così che le unità pubbliche di offerta sono state messe in competizione tra loro e con imprese caratterizzate da una differente forma proprietaria (a fine di lucro o nonprofit).

In questo contesto di competizione crescente, un ruolo rilevante è stato svolto proprio dalle imprese senza fine di lucro. La presenza delle organizzazioni nonprofit entro il settore sanitario viene solitamente spiegata a partire dalla loro capacità di ovviare ad alcuni classici "fallimenti del mercato" che si presentano in un contesto caratterizzato da incertezza ed asimmetria informativa. Il vincolo di non distribuzione dei profitti rappresenterebbe infatti uno strumento valido per superare tanto i problemi delle imprese a fine di lucro, che soffrono i fallimenti del mercato, che quelli delle stesse imprese pubbliche, soggette al *categorical constraint* (Douglas, 1983) che le obbliga a fornire trattamenti identici a tutti i cittadini. Ai vantaggi delle organizzazioni nonprofit si sommano però alcuni costi. Prima di tutto quello di una minore efficienza prodotta dalla inappropriabilità dei residui; a causa di ciò, i manager delle organizzazioni nonprofit avrebbero infatti scarsi incentivi a minimizzare i costi di produzione. I risultati dei lavori empirici (soprattutto statunitensi) che hanno mirato a stimare l'efficienza delle diverse forme proprietarie sono tuttavia controversi, come si mostra sia nel quarto paragrafo di questo lavoro che in altre rassegne (Marmor, Schlesinger e Smithey, 1987).

In questa sede ci proponiamo di affrontare proprio il problema della efficienza tecnica di imprese operanti entro il settore sanitario e caratterizzate da diverse strutture proprietarie; non si intende discutere se l'operatore pubblico

debba cessare, oppure no, di finanziare la fornitura di servizi sanitari, ma solo se sia opportuno (sulla base della diversa efficienza tecnica delle imprese) che esso produca in proprio questi servizi.

L'obiettivo del lavoro è duplice. Da una parte esso vuole offrire una occasione di riflessione sulla metodologia che può essere utilizzata per stimare l'efficienza tecnica delle imprese. In particolare, ci concentriamo sul confronto di indicatori di efficienza ottenuti con metodologie alternative utilizzando dati sulla sanità ospedaliera lombarda. Dall'altra, il lavoro mira ad analizzare le cause che spiegano i diversi livelli di efficienza delle imprese, in riferimento soprattutto alla loro struttura proprietaria.

A tal fine, il secondo paragrafo introduce la nozione di efficienza tecnica, differenziandola da quella di efficienza allocativa, che male si presta per stimare l'efficienza di strutture proprietarie che perseguono obiettivi diversi dalla minimizzazione dei costi; il terzo paragrafo presenta diverse tecniche utilizzate per stimare l'efficienza, concentrandosi in particolare sulla Data Envelopment Analysis e sulla stima di frontiere stocastiche; il quarto paragrafo presenta una rassegna della letteratura empirica sulla efficienza nel settore sanitario; il quinto paragrafo illustra la metodologia ed i risultati delle stime relative al caso lombardo, mentre il sesto paragrafo presenta le conclusioni del lavoro.

2 Le definizioni economiche di inefficienza

2.1 L'inefficienza tecnica in processi produttivi semplificati

Al fine di fornire una definizione teorica dei diversi aspetti dell'inefficienza nella produzione, consideriamo dapprima un semplice processo produttivo che utilizza un solo input x per ottenere un solo output y. Sia:

$$y = p(x) \Leftrightarrow x = l(y) \tag{1}$$

la funzione di produzione che definisce, dato un certo livello della tecnologia, il massimo output y ottenibile per dato input x (oppure, inversamente,

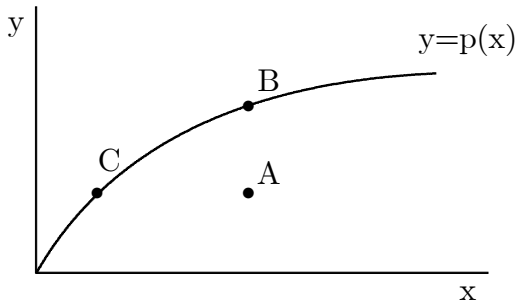

Figure 1: Efficienza tecnica

il minimo input x necessario per ottenere un dato output y). Assumiamo inoltre $y' > 0$ e $y'' < 0$. Il processo produttivo che abbiamo descritto é rappresentato graficamente in figura 1.

E' abbastanza intuitivo considerare come inefficiente qualsiasi unità produttiva che, dato un certo livello di utilizzo dell'unico input x, non produce il massimo output ottenibile y come descritto dalla funzione di produzione (oppure, parallelamente, dato un certo livello di produzione y non minimizza l'utilizzo dell'input x). Parliamo in questo caso di *inefficienza tecnica* o *inefficienza nel senso di Debreu-Farrell*.[1]

Una prima intuitiva misura dell'inefficienza tecnica può essere derivata da un confronto fra il prodotto medio di differenti unità. Considerando la figura 1, ad esempio, possiamo scrivere che:

$$P.Me.^A = \frac{y^A}{x^A} < P.Me.^C = \frac{y^C}{x^C} \tag{2}$$

dato che sia A che C producono lo stesso livello di output ma C utilizza l'unico input x in misura minore. In questo modo possiamo identificare nell'unità produttiva C la più efficiente fra le due.

Una misura di inefficienza tecnica *in termini di input* (derivata dalla precedente) considerando C come termine di confronto è data dal semplice rapporto:

[1] Si veda ad esempio Lovell (1993), pp. 10 ss.

$$T.E._i(y^A, x^A) = \frac{x^A}{x^C} \tag{3}$$

Si noti che l'indice è determinato *per dato livello di output*. Valori dell'indice superiori all'unità implicano inefficienza tecnica dell'unità produttiva considerata. E' infatti possibile ottenere lo stesso livello di output y utilizzando una quantità inferiore dell'input x. Ovviamente, un indice pari ad 1 è sinonimo di efficienza. Sulla scorta di queste considerazioni possiamo scrivere:

$$\left(\frac{x^A}{x^C}\right)\lambda = 1 \tag{4}$$

dove $\lambda \leq 1$ identifica la riduzione nell'utilizzo dell'input x necessaria per raggiungere l'efficienza. In questo caso λ rappresenta l'indicatore di inefficienza tecnica introdotto nella letteratura da Debreu e Farrell.[2] In termini più formalizzati possiamo scrivere[3]:

$$D.F._i(y^A, x^A) = \min\left\{\lambda : x^A\lambda \in x = l(y)\right\} \tag{5}$$

Indicatore del tutto analogo può essere ottenuto considerando la funzione di distanza introdotta da Shephard, che formalizza la misura di inefficienza tecnica $T.E._i$ che abbiamo visto in precedenza[4]:

$$D_i(y^A, x^A) = \max\left\{\lambda : x^A/\lambda \in x = l(y)\right\} \tag{6}$$

Chiaramente, la funzione di distanza può essere riscritta come:

$$D_i(y^A, x^A) = \frac{1}{D.F._i(y^A, x^A)} \tag{7}$$

[2]Cfr. Lovell (1993), pp. 10-11.

[3]Si noti che, nel nostro caso, $D.F._i = \frac{x^C}{x^A} = \frac{x^A\lambda}{x^A}$.

[4]In altre parole, nel nostro caso, $D_i = T.E._i = \frac{x^A}{x^C} = \left(\frac{x^A}{x^A/\lambda}\right)$.

ad indicare che la misura di inefficienza di Debreu-Farrell è semplicemente l'inversa di quella di Shephard.

Le misure di inefficienza tecnica viste sinora possono essere definite anche *in termini di output*. Il semplice rapporto:

$$T.E._o(x^A, y^A) = \frac{y^A}{y^B} \tag{8}$$

definito *per dato livello di input*, è un indicatore di inefficienza tecnica negli output. Valori dell'indice inferiori all'unità implicano inefficienza dell'unità produttiva in quanto è possibile ottenere un più elevato livello di output con lo stesso input. Come in precedenza, partendo da queste considerazioni possiamo scrivere:

$$\left(\frac{y^A}{y^B}\right)\theta = 1 \tag{9}$$

dove $\theta \geq 1$ indica l'espansione nella produzione dell'output y - ottenibile a parità di input - necessaria per raggiungere l'efficienza. In modo più formale, possiamo scrivere la misura di inefficienza di Debreu-Farrell negli output come[5]:

$$D.F._o(x^A, y^A) = \max\left\{\theta : y^A\theta \in y = p(x)\right\} \tag{10}$$

Indicatore analogo può essere derivato considerando la funzione di distanza di Shephard[6]:

$$D_o(x^A, y^A) = \frac{1}{D.F._o(x^A, y^A)} = \min\left\{\theta : y^A/\theta \in y = p(x)\right\} \tag{11}$$

Le osservazioni sulle misure di inefficienza sono raccolte nella tavola 1.

[5] Come in precedenza, nel nostro caso $D.F._o = \frac{y^B}{y^A} = \frac{\theta y^A}{y^A}$.

[6] Analogamente al caso precedente, $D_o = T.E._o = \frac{y^A}{y^B} = \left(\frac{y^A}{y^A/\theta}\right)$.

Tavola 1. Misure di inefficienza

orientamento	**Debreu-Farrell** $D.F.$	**Distanza di Shephard** D
	$oss.^{eff.}/oss.^{i}$	$oss.^{i}/oss.^{eff.}$
input	$\lambda \leq 1$	$1/\lambda \geq 1$
output	$\theta \geq 1$	$1/\theta \leq 1$

La misura di inefficienza di Debreu-Farrell gode delle seguenti proprietà:[7]
1) $D.F._i(y,x)$ è omogenea di grado -1 negli input:
$D.F._i(y,\alpha x) = D.F._i(y,x)/\alpha$;
2) $D.F._i(y,x)$ è monotona decrescente negli input: $\partial D.F._i(y,x)/\partial x \leq 0$;
3) $D.F._i(y,x)$ è invariante rispetto a modifiche nell'unità di misura.
Inoltre, se - e solo se - la tecnologia presenta rendimenti di scala costanti vale la proprietà:
4) $D.F._i(y,x)$ è l'inversa della misura di Debreu-Farrell negli output: $D.F._i(y,x) = 1/D.F._o(x,y)$.

2.2 L'inefficienza tecnica in processi produttivi multi-input e multi-output

Gli indicatori di inefficienza tecnica che abbiamo individuato nel caso di tecnologie semplificate possono essere estesi al caso di processi produttivi più complessi come nel caso di un ospedale, dove si utilizzano generalmente una molteplicità di input per ottenere una molteplicità di output. Supponiamo che il processo produttivo utilizzi n input $\mathbf{x} \in \mathbf{R}_+^N$ per ottenere m output $\mathbf{y} \in \mathbf{R}_+^M$. Il massimo livello di output $\mathbf{y}$ ottenibili dall'utilizzo di dati input $\mathbf{x}$ e, per converso, il minimo livello di input $\mathbf{x}$ necessario per ottenere dati output $\mathbf{y}$, è descritto dalle corrispondenze:

$$P : \mathbf{R}_+^N \longrightarrow \mathbf{y} = P(\mathbf{x}) : \mathbf{R}_+^M \Leftrightarrow L : \mathbf{R}_+^M \longrightarrow \mathbf{x} = L(\mathbf{y}) : \mathbf{R}_+^N \qquad (12)$$

Vengono mantenute anche in questo caso le ipotesi sulla concavità del processo produttivo avanzate in precedenza. Un problema nella determinazione delle misure di inefficienza in processi produttivi complessi è rappresentato

[7] Vedi Lovell (1993), p. 13.

dalla necessità di ridurre a scalari i vettori di input e di output. Generalmente il problema è risolto seguendo due strategie alternative. Una prima via è quella di omogeneizzare i differenti input e output attraverso l'utilizzo di pesi opportuni. Ad esempio, il prodotto medio dell'unità produttiva A può essere definito in questo contesto come:

$$P.Me.^{A} = \frac{\sum_{i=1}^{m} p_i y_i^A}{\sum_{j=1}^{n} w_j x_j^A} \tag{13}$$

dove p e w rappresentano i pesi, la cui scelta è soggettiva, per omogeneizzare e poter così confrontare input ed output.[8]

Una seconda via è quella di considerare la natura di $\mathbf{y}$ e $\mathbf{x}$ di vettori in spazi m- ed n-dimensionali. In questo caso, la norma euclidea *"can be interpreted as a scalar measure of the "size" of the output [input] vector or its distance from the origin"*.[9] Il prodotto medio può quindi essere riscritto come[10]:

$$P.Me.^{A} = \frac{\left\|\mathbf{y}^A\right\|}{\left\|\mathbf{x}^A\right\|} = \frac{\left(\sum_{i=1}^{m} y_i^2\right)^{1/2}}{\left(\sum_{j=1}^{n} x_j^2\right)^{1/2}} \tag{14}$$

In questo quadro, la funzione di distanza *negli input* può essere definita come:

$$D_i(\mathbf{y}, \mathbf{x}) = \max\left\{\lambda : \mathbf{x}/\lambda \in \mathbf{x} = L(\mathbf{y})\right\} \tag{15}$$

Ancora una volta λ rappresenta un indicatore di inefficienza tecnica, questa volta esteso a processi produttivi multi-input e multi-output. Nel caso, ad esempio, della generica unità produttiva A, percorrendo la seconda via proposta avremo:

$$D_i(\mathbf{y}^A, \mathbf{x}^A) = \frac{\left\|\mathbf{x}^A\right\|}{\left\|\mathbf{x}^A/D_i(\mathbf{y}^A, \mathbf{x}^A)\right\|} = \frac{\left\|\mathbf{x}^A\right\|}{\left\|\mathbf{x}^A/\lambda\right\|} \tag{16}$$

[8] Generalmente i pesi p e w rappresentano i prezzi degli input e degli output.

[9] Cfr. Grosskopf e al. (1997), p. 118.

[10] Sul punto si veda ad esempio Gerdtham et al. (1999), p. 154.

In modo del tutto analogo possiamo poi definire la funzione di distanza *negli output* come:

$$D_o(\mathbf{x}, \mathbf{y}) = \min \{\theta : \mathbf{y}/\theta \in \mathbf{y} = P(\mathbf{x})\} \tag{17}$$

dove, θ rappresenta l'indice di inefficienza tecnica. Nel caso della generica unità produttiva A avremo quindi:

$$D_o(\mathbf{x}^A, \mathbf{y}^A) = \frac{\|\mathbf{y}^A\|}{\|\mathbf{y}^A/D_o(\mathbf{y}^A, \mathbf{x}^A)\|} = \frac{\|\mathbf{y}^A\|}{\|\mathbf{y}^A/\theta\|} \tag{18}$$

2.3 L'inefficienza allocativa

Una definizione di inefficienza differente dall'inefficienza tecnica è quella di *inefficienza allocativa* o *inefficienza nel senso di Koopmans.*[11] Consideriamo ancora il semplice processo produttivo che utilizza un input per la produzione di un output. Sia w il costo dell'input x e p il prezzo dell'output y. E' noto che, *se l'unità produttiva vuole minimizzare i propri costi* sceglierà di impiegare una quantità di input x tale per cui, *al margine*, il prodotto di x è uguale al rapporto tra il costo dell'input x ed il prezzo dell'output y. Diremo che un'unità produttiva è efficiente in senso allocativo se sceglie la quantità di input x che minimizza i suoi costi. Un'unità produttiva che è efficiente in senso tecnico potrebbe quindi essere inefficiente in senso allocativo, ma non è vero il contrario. In altre parole, l'efficienza nel senso di Debreu-Farrell è condizione necessaria ma non sufficiente per l'efficienza nel senso di Koopmans.[12]

Sono almeno due le ragioni per le quali ci concentreremo esclusivamente sulla definizione di inefficienza tecnica in questo lavoro. In primo luogo perchè l'efficienza allocativa implica assunzioni sulle regole comportamentali delle unità produttive e solo alcune delle organizzazioni operanti nei settori che andremo ad esaminare sembrano essere unità che minimizzano i propri

[11] Sebbene Koopmans non parli - nella sua definizione di efficienza - di prezzi degli input e degli output, sembra possibile interpretarla come inefficienza in senso allocativo se vale l'ipotesi di stretta concavità della funzione di produzione come nel nostro caso. Sulla definizione di Koopmans si veda ad esempio Lovell (1993), pp. 10 ss.

[12] Si veda Lovell (1993), p. 13.

costi. Il settore della sanità mostra la compresenza di imprese lucrative, di organizzazioni nonprofit e di unità di produzione pubbliche. Analizzare l'influenza della forma organizzativa sull'efficienza utilizzando una misura di efficienza che implica la minimizzazione dei costi è, quantomeno, fuorviante.[13] In secondo luogo considereremo solo il concetto di efficienza tecnica perchè l'efficienza allocativa implica la conoscenza dei dati di prezzo degli input e degli output e, nel caso dei servizi sanitari, uno dei problemi che si incontrano è proprio la mancanza di questo tipo di informazione.

3 Efficienza e frontiere dell'insieme di produzione

Come abbiamo visto nel paragrafo precedente, la misurazione dell'inefficienza è basata, in generale, sulla *distanza* dalla frontiera dell'insieme di produzione del vettore input-output che sintetizza il processo produttivo di una certa organizzazione. Il problema della misurazione dell'inefficienza è quindi riconducibile alla definizione della frontiera, chiaramente non nota, partendo da un campione di unità produttive con i relativi vettori input-output. Una volta ottenuta una *stima* della frontiera dell'insieme di produzione come *"best practice" frontier*, è poi agevole definire l'inefficienza associata alle unità produttive del campione attraverso gli indicatori analizzati. Si noti che la procedura di stima non conduce alla *vera* frontiera, dal momento che normalmente viene considerato solo un campione - per quanto rappresentativo - di unità produttive.[14] Ciò implica che alcune unità di produzione potrebbero essere ritenute efficienti sulla base della *"best practice" frontier* pur essendo inefficienti sulla base della *vera* - ma ignota - frontiera dell'insieme di produzione.

Esistono metodi differenti per la stima della frontiera dell'insieme di produzione di una certa industria. Si distinguono innanzitutto metodi *deterministici* e metodi *stocastici*. Per i primi le deviazioni dalla frontiera produttiva dipendono esclusivamente dall'inefficienza dell'unità di produzione. Per i secondi, invece, le deviazioni dalla frontiera dipendono sia dall'inefficienza

[13] L'argomento è utilizzato spesso nella letteratura sull'efficienza di imprese pubbliche. Si veda Lovell (1993), p. 26.

[14] Come ha felicemente notato Lovell (1993), p. 5, "it is as difficult for the analyst to determine empirically the potential of a production unit as it is for the producer to achieve that potential".

dell'unità produttiva che da variabili aleatorie che potenzialmente potrebbero influenzare il processo di produzione. Un'altra distinzione è fra metodi *parametrici* e metodi *non parametrici*: i primi specificano una ben definita forma funzionale per $\mathbf{y} = p(\mathbf{x})$ [15] mentre i secondi non specificano alcuna forma funzionale. Nel seguito di questo paragrafo ci concentreremo su due tra i metodi più utilizzati dalla letteratura empirica: la *Data Envelopment Analysis* (D.E.A.), che rappresenta essenzialmente un metodo deterministico ma non parametrico[16], ed il metodo delle frontiere stocastiche, che al contrario rappresenta un metodo stocastico di tipo parametrico.

3.1 La Data Envelopment Analysis

Si consideri, per semplicità espositiva, un semplice processo produttivo che utilizza un solo input x per ottenere un solo output y. La metodologia della Data Envelopment Analysis definisce la *"best practice" frontier* come l'inviluppo superiore dei vettori input-output del campione di unità produttive. La figura 2 è utile nell'evidenziare il processo di costruzione della frontiera efficiente. Si considerino, a titolo esemplificativo, le unità produttive A e B che producono livelli di output y differenti, pur utilizzando la stessa quantità di input x. Il metodo della D.E.A. considera B come appartenente alla frontiera, mentre A come unità produttiva inefficiente. In altre parole, per ogni livello di input x, si considera come appartenente alla *"best practice" frontier* l'unità produttiva che raggiunge il più elevato livello di produzione. In termini analitici, considerando un modello a rendimenti di scala costanti, il problema può essere scritto come[17]:

$$\begin{aligned} \max_{k,v} \quad P.Me. &= \frac{ky_j}{vx_j} \\ s.t. \quad \frac{ky_i}{vx_i} &\leq 1;\ i = 1, 2, ..., I \\ k, v &\geq 0 \end{aligned} \tag{19}$$

[15] Un esempio in questo senso utilizzato nella letteratura empirica è la classica funzione di produzione Cobb-Douglas.

[16] Parte della letteratura ha esteso la D.E.A. per incorporare elementi stocastici. Si vedano ad esempio Kalirajan - Shand (1999), p. 155, e Lovell (1993), pp. 34 ss.

[17] E' il cosiddetto problema in "ratio form". Si veda Coelli (1996), p. 9.

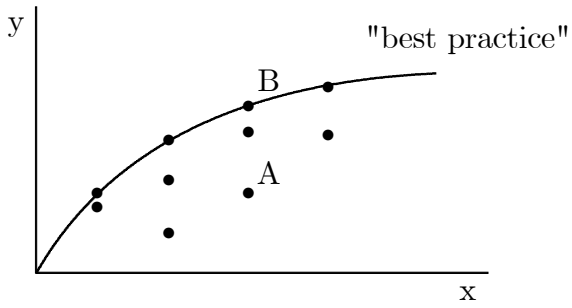

Figure 2: Costruzione della "best practice" frontier negli input seguendo il metodo della D.E.A.

dove k e v sono dei pesi riferiti rispettivamente agli output e agli input, mentre I rappresenta il numero complessivo di imprese nel campione. La presenza di una funzione obiettivo non lineare ed il numero infinito di soluzioni[18] rendono il problema **??** di scarsa applicazione pratica.

Tuttavia, linearizzando la funzione obiettivo possiamo scrivere[19]:

$$\begin{aligned} \max_{\varkappa,v} \quad y &= \varkappa y_j \\ s.t. \quad vx_j &= 1 \\ \varkappa y_i - vx_i &\leq 0;\ i = 1, 2, ..., I \\ \varkappa, v &\geq 0 \end{aligned} \tag{20}$$

dove $\varkappa$ e v rappresentano dei (nuovi) pesi riferiti rispettivamente agli output e agli input. Il problema **??** presenta 2 variabili ($\varkappa$ e v) ed $1 + I$ vincoli. Il problema duale al problema **??** avrà quindi $1 + I$ variabili e solo 2 vincoli. Normalmente viene risolto quest'ultimo, sia perchè più semplice, sia perchè ci consente di definire più chiaramente un collegamento logico con le misure di inefficienza sviluppate in precedenza. Sfruttando la dualità avremo infatti[20]:

[18] Si noti che se k^* e v^* sono soluzioni del problema, allora anche αk^* e αv^* sono soluzioni.

[19] E' il cosiddetto problema in "multiplier form". Si veda Coelli (1996), p. 10.

[20] E' il cosiddetto problema in "envelopment form". Si veda Coelli (1996), p. 10.

$$\begin{aligned} \min_{\lambda,\pi} \quad D.F._i &= \lambda \\ s.t. \quad -y_i + \mathbf{Y}\boldsymbol{\pi} &\geq 0 \\ \lambda x_i - \mathbf{X}\boldsymbol{\pi} &\geq 0 \\ \boldsymbol{\pi} &\geq 0 \end{aligned} \tag{21}$$

dove λ rappresenta la misura di Debreu-Farrell, Y ed X rispettivamente i vettori degli output e degli input, π il vettore dei pesi che definiscono la frontiera efficiente. E' facile notare, sfruttando la definizione della misura di Debreu-Farrell nell'equazione **??**, come in questo caso gli indicatori siano *input-orientated:* valori degli indicatori minori di 1 segnalano situazioni di potenziali riduzioni degli input. Chiaramente, il problema deve essere risolto I volte, una per ogni impresa appartenente al campione.

L'assenza di restrizioni sulla forma funzionale della relazione che lega gli input agli output $y = p(x)$ rappresenta chiaramente uno dei vantaggi della D.E.A. Un ulteriore vantaggio della D.E.A. deriva poi dal fatto che, essendo un metodo deterministico, non richiede l'imposizione di alcuna ipotesi sulla funzione di distribuzione degli *score* di inefficienza.[21] La D.E.A. può essere inoltre facilmente estesa al caso di tecnologie che utilizzano più di un input per ottenere più di un output. Nel caso, ad esempio, di indicatori definiti *negli output* possiamo scrivere:

$$\begin{aligned} \max_{\theta,\pi} \quad D.F._o &= \theta \\ s.t. \quad \theta\mathbf{y}_i - \mathbf{Y}\boldsymbol{\pi} &\leq 0 \\ -\mathbf{x}_i + \mathbf{X}\boldsymbol{\pi} &\leq 0 \\ \boldsymbol{\pi} &\geq 0 \end{aligned} \tag{22}$$

semplicemente sostituendo tecnologie multi-input e multi-output a tecnologie che prevedono un solo input e un solo output, mantenendo inalterata la struttura del problema.

[21] In assenza di una specifica distribuzione degli *score* di efficienza, è chiaro tuttavia come non risulti nemmeno possibile l'implementazione di test di significatività degli stimatori ottenuti. Tuttavia, esistono in letteratura dei lavori miranti a definire le distribuzioni statistiche degli *score* ottenuti con la D.E.A. Per una rassegna delle tecniche di inferenza in modelli non parametrici si veda Simar e Wilson (1999).

L'approccio della D.E.A. nella stima della *"best practice" frontier* produce, per converso, indicatori estremamente sensibili alla scelta delle variabili di input e di output e, conseguentemente, facilmente influenzabili da errori di misurazione. Nel caso poi di piccoli campioni e di processi produttivi complessi, gli *score* di inefficienza possono risultare sensibili alla differenza tra il numero di unità produttive considerate e la somma di input e di output.[22]

3.2 Le frontiere stocastiche

Se il metodo della D.E.A. rappresenta un metodo deterministico non-parametrico, quello delle frontiere stocastiche rappresenta un metodo generalmente parametrico ma stocastico. Partiamo, ancora una volta, dal semplice processo produttivo che utilizza un singolo input x per produrre un singolo output y. La relazione deterministica viene modificata per tenere conto sia dell'inefficienza, che del potenziale effetto di variabili aleatorie. Il livello di produzione della *i-esima* unità è definito da $y_i = p(x_i) + \varepsilon_i$, dove $\varepsilon_i = v_i - u_i$ rappresenta le deviazioni dalla frontiera di produzione dovute all'azione combinata dell'inefficienza (u_i) e di altri elementi aleatori (v_i).[23]

Appare subito evidente che uno dei grossi problemi dell'approccio delle frontiere stocastiche è la necessità di specificare la funzione di distribuzione dell'errore composto ε_i. Generalmente si assume che il termine di errore v_i sia distribuito come una variabile casuale normale $v_i \sim \mathbf{NID}(0, \sigma_v^2)$, mentre per il termine legato all'inefficienza u_i si possono avere differenti specificazioni, anche se la più utilizzata nella letteratura empirica sembra essere la *half-normal distribution.*[24] In questo particolare caso $u_i \sim \mathbf{NID}(\mu, \sigma_u^2)$. Si assume inoltre che u_i e v_i siano indipendenti: $Cov(u_i, v_i) = 0, \forall i$.

Una volta definite le funzioni di distribuzione degli errori, si utilizzano tecniche econometriche per stimare i parametri della *"best practice" frontier* $y_i = p\left(x_i, \beta\right) + \varepsilon_i$. Se specifichiamo una forma funzionale per la relazione, ad esempio la classica Cobb-Douglas, possiamo scrivere:

$$y_i = \alpha x_i^{\beta} \varepsilon_i \tag{23}$$

[22] Sul punto cfr. Kalirajan - Shand (1999), pp. 167.

[23] Si pensi per esempio al caso di uno sciopero. E' bene ricordare che, nel caso di una funzione di costo, si assume che $\varepsilon_i = v_i + u_i$.

[24] Sul punto si veda Greene (1993), p. 80.

dove α rappresenta il livello (esogeno) della tecnologia. Passando ai logaritmi abbiamo:

$$\ln y_i = \ln \alpha + \beta \ln x_i + \ln \varepsilon_i \tag{24}$$

che può essere riscritta per semplicità come:

$$y_i^* = \alpha^* + \beta x_i^* + \varepsilon_i^* \tag{25}$$

dove (*) vicino ad una variabile indica i logaritmi. L'equazione precedente può essere stimata seguendo almeno due procedure differenziate: la tecnica dei *Minimi Quadrati Ordinari Modificati* (M.O.L.S.) o la più tradizionale *Massima Verosimiglianza* (M.L.E.).

La tecnica dei M.O.L.S. cerca di ripristinare una delle ipotesi di base dei Minimi Quadrati Ordinari per ottenere stimatori consistenti.[25] Infatti, se assumiamo che $v_i \sim \mathbf{NID}(0, \sigma_v^2)$, $u_i \sim \mathbf{NID}(\mu, \sigma_u^2)$ e $Cov(u_i, v_i) = 0, \forall i$, considerando $\varepsilon_i^* = v_i - u_i$ abbiamo che $\mathbf{E}(\varepsilon_i^*) = |\mu| \neq 0$. Perchè il metodo dei Minimi Quadrati Ordinari applicato all'equazione **??** restituisca stimatori consistenti è necessario ripristinare l'ipotesi $\mathbf{E}(\varepsilon_i^*) = 0$. Un modo semplice ed intuitivo è chiaramente quello di sottrarre dall'errore composto ε_i^* la media stimata dai residui $\mathbf{E}(u_i)$. L'equazione **??** può essere riscritta come:

$$y_i^* = (\alpha^* + \mu) + \beta x_i^* + (\varepsilon_i^* - \mu) \tag{26}$$

oppure:

$$y_i^* = a + \beta x_i^* + e_i \tag{27}$$

dove $a = \alpha^* + \mu$ ed $e_i = \varepsilon_i^* - \mu$. Gli O.L.S. applicati all'equazione così modificata consentono di ottenere stimatori consistenti per il coefficiente β ma non per la costante α. Risulta tuttavia agevole ottenere uno stimatore consistente anche della costante considerando la relazione:

[25] Gli stimatori rimangono comunque inefficienti rispetto a quelli ottenuti con la M.L.E. Sul punto cfr. Greene (1993), p. 77.

$$\widehat{\alpha}^* = a - \mu \tag{28}$$

Nel caso particolare di una *half-normal distribution* è possibile mostrare che $\mathbf{E}(u_i) = (2/\pi)^{1/2}\sigma_u$, quindi avremo[26]:

$$\widehat{\alpha}^* = a - \widehat{\sigma}_u \left(\frac{2}{\pi}\right)^{\frac{1}{2}} \tag{29}$$

Una particolarità nell'utilizzo della *half-normal distribution* è la possibilità di una diagnostica sulla specificazione del modello basata sulla *skewness* dei residui. Normalmente la distribuzione del termine d'errore dovrebbe presentare una *skewness* a sinistra, ma se il modello non è ben specificato la distribuzione potrebbe presentare una *skewness* a destra.[27]

La *"best practice" frontier* può essere stimata utilizzando la tecnica alternativa della M.L.E. Se assumiamo, come in precedenza, una *half-normal distribution* per il termine legato all'inefficienza u_i è possibile mostrare che la funzione di verosimiglianza può essere scritta come[28]:

$$l(\alpha, \beta, \sigma, \lambda) = -I \ln \sigma - \psi + \sum_{i=1}^{I} \left[\ln \Phi \left(\frac{-\varepsilon_i \lambda}{\sigma} \right) - \frac{1}{2} \left(\frac{\varepsilon_i}{\sigma} \right)^2 \right] \tag{30}$$

dove I è il numero di unità di produzione, ψ è una costante, $\lambda = (\sigma_u/\sigma_v)$, $\sigma^2 = \sigma_u^2 + \sigma_v^2$ e $\Phi(.)$ è la funzione di ripartizione di una variabile casuale normale standard. La massimizzazione della funzione di verosimiglianza porta alla definizione di stimatori non solo consistenti, ma anche efficienti rispetto agli stimatori ottenuti con il metodo dei Minimi Quadrati Ordinari.

Si noti che, una volta stimata la frontiera efficiente con uno dei due metodi appena descritti (M.O.L.S. o M.L.E.), si ottengono i residui $\varepsilon_i^* = v_i - u_i$ e non la sola componente legata all'inefficienza u_i. Quest'ultima deve essere quindi

[26] Si veda Greene (1993), p. 77.

[27] Sul punto vedi Greene (1993), p. 78. I principali *package* econometrici utilizzano come base di partenza per le iterazioni della M.L.E. proprio le stime ottenute con gli O.L.S., effettuando il test di errata specificazione del modello basato sulla *skewness* dei residui prima di proseguire nella stima.

[28] Si veda Greene (1993), p. 76.

osservata indirettamente. Nel caso della *half-normal distribution* possiamo scrivere[29]:

$$\mathbf{E}\left[u_i \mid \varepsilon_i\right] = \frac{\sigma\lambda}{(1+\lambda^2)}\left[\frac{\phi(\varepsilon_i\lambda/\sigma)}{\Phi(-\varepsilon_i\lambda/\sigma)} - \frac{\varepsilon_i\lambda}{\sigma}\right] \tag{31}$$

dove $\phi(.)$ è la funzione di densità di una variabile casuale normale standard, mentre σ, λ e $\Phi(.)$ restano definiti come in precedenza. Ciò consente di ottenere stime non distorte e tuttavia non consistenti della componente di inefficienza u_i.[30]

Sembra chiaro che l'imposizione di una certa forma funzionale alla frontiera efficiente, nonchè di una certa funzione di distribuzione alla componente di errore composto ε_i^* rappresentano i principali limiti dell'approccio delle frontiere stocastiche nella misurazione dell'efficienza produttiva. D'altro canto, l'utilizzo di tecniche econometriche consente di testare statisticamente sia le varie ipotesi relative alla tecnologia, come quelle legate alle misure di efficienza della singola unità di produzione.[31] L'approccio delle frontiere stocastiche consente inoltre di distinguere tra l'inefficienza e l'operare di altre variabili aleatorie che potenzialmente potrebbero influenzare il risultato del processo di produzione.

4 Una rassegna della letteratura empirica

Esiste ormai una vasta letteratura empirica sull'analisi dell'efficienza produttiva attraverso la stima di frontiere di produzione o frontiere di costo. Le applicazioni spaziano nei più disparati settori produttivi e comprendono anche il settore della sanità. La presente rassegna non ha ovviamente alcuna pretesa di essere esaustiva. In questo paragrafo analizziamo la più recente letteratura sul tema per evidenziare le tecniche per misurare l'inefficienza e le metodologie di stima utilizzate, le variabili individuate ed i principali risultati.

[29] Si veda Greene (1993), pp. 80-81.

[30] Come osservato da Greene (1993), p. 81, le stime "are inconsistent because regardless of N [the number of observations], the variance of the estimates remains nonzero, not because they converge to some other quantity".

[31] Sul punto si veda Kalirajan - Shand (1999), p. 168.

In generale, due sembrano essere le principali indicazioni derivabili dall'analisi. Innanzitutto, nonostante alcuni degli studi analizzati comparino i risultati ottenuti perseguendo strategie differenti e sembrino confermare l'idea che le indicazioni desumibili dagli *score* di inefficienza non mutano al mutare della metodologia di stima adottata[32], non sembra emergere con chiarezza un'unica metodologia per misurare l'inefficienza. In secondo luogo, gli studi relativi al nostro paese sono pressochè inesistenti sia per l'obiettiva difficoltà nel reperimento di dati, che per l'interesse relativamente recente ad utilizzare le *"best practice" frontier* nello studio dell'efficienza produttiva. Tentativi di misurare la produttività, l'efficienza e l'efficacia nel settore dei servizi pubblici sono stati infatti proposti in Italia sul finire degli anni '80, ma paiono essere tutti basati su indicatori più o meno complessi e non sulla nozione di distanza da una frontiera efficiente.[33]

Gli studi empirici relativi ai settori della sanità e dell'assistenza mostrano, come detto, un'ampia variabilità per quanto riguarda le tecniche utilizzate per la misurazione dell'efficienza. E' stata indifferentemente usata come *"best practice" frontier* sia la funzione di costo che la funzione di produzione, pur considerando le critiche avanzate a proposito della stima di una funzione di costo in settori caratterizzati dalla compresenza di unità di produzione con differenti funzioni obiettivo. La stessa variabilità si nota nelle procedure di stima. Esistono studi che utilizzano versioni più o meno sofisticate della D.E.A. e studi che, per converso, sfruttano la metodologia delle frontiere stocastiche. In quest'ultimo caso la tecnica di stima più usata sembra essere quella della M.L.E.[34]

Un aspetto centrale della letteratura empirica sull'efficienza nei settori della sanità e dell'assistenza è chiaramente l'individuazione degli input, degli output e dei relativi prezzi. Tutti gli studi analizzati sono concordi nell'affermare che il processo produttivo dei servizi sanitari e dei servizi di assistenza ha come obiettivo finale il miglioramento delle condizioni di salute dei pazienti.[35] Tuttavia il risultato finale della produzione è di difficile valutazione e

[32]Sul punto cfr. in particolare Linna (1998). In senso contrario, cfr. invece Giuffrida e Gravelle (1999).

[33]Si veda Tronti (1991) per una rassegna delle principali esperienze italiane in tema di indicatori di produttività utilizzati nei settori della sanità e dell'istruzione.

[34]L'unico studio che utilizza la tecnica dei M.O.L.S. è quello di Filippini (1998).

[35]Nel modello di Grossman (1972) infatti - la base interpretativa dell'analisi economica del settore sanità - la domanda di servizi sanitari è una domanda derivata dalla domanda di salute.

risulta essere praticamente impossibile da misurare. Si ricorre così a variabili *proxy* del risultato finale del processo produttivo, rappresentate normalmente dagli output intermedi. Queste misure includono il numero dei pazienti, delle visite, delle giornate complessive di degenza, dei trattamenti medici di vario tipo prodotti dall'ospedale o dalla casa di cura al fine di migliorare la salute dei propri pazienti. Se l'ospedale è anche un centro universitario di ricerca, si considerano fra gli output anche le pubblicazioni scientifiche dei ricercatori afferenti all'istituto stesso.

Un problema importante per l'identificazione degli output nei settori della sanità e dell'assistenza è rappresentato dalla dimensione qualitativa del prodotto. La mancata considerazione della qualità dei servizi nella misurazione dell'efficienza produttiva potrebbe portare a confondere la produzione di servizi di migliore qualità con l'inefficienza dell'unità di produzione. Tentativi di misurare la qualità del servizio sono presenti nella letteratura empirica. Un esempio in questo senso è quello offerto da Vitaliano e Toren (1994) che considerano la somma di tutte le deficienze nel servizio segnalate al *Department of Health* statunitense dai pazienti o dai loro parenti. Un ulteriore esempio è quello fornito da Filippini (1998) che misura la qualità del servizio considerando il rapporto tra personale medico effettivo delle case di riposo e personale medico teorico, sulla base di quanto stabilito dal Dipartimento Opere Sociali svizzero.

Un ulteriore problema, evidente soprattutto nella produzione di servizi sanitari è la disomogeneità dei trattamenti ospedalieri.Valutazioni di efficienza produttiva o di efficienza di costo condotte senza tener conto di tale aspetto della produzione possono portare a conclusioni errate. Ad esempio, i costi di un ospedale possono essere più elevati rispetto a quelli di un altro semplicemente perchè i casi trattati sono più complessi. Una possibile soluzione per ottenere una misura scalare di output è il ricorso ai D.R.G. (*Diagnosis Related Groups*) e a figure di "prezzi" degli output. Un'altra soluzione proposta dalla letteratura è invece la valutazione dell'efficienza a livello del singolo paziente e, solo indirettamente, a livello di singola unità produttiva.[36] Valutazioni a livello di singoli pazienti consentono di concentrare l'attenzione solo su pazienti che presentano casistiche omogenee. Si evita, in questo modo, il problema della disomogeneità degli output.

Nel caso della stima di una funzione di produzione, gli input utilizzati sono rappresentati normalmente dal numero dei medici, dal numero degli

[36]Esempi in questo senso sono Zuckerman e al. (1994) e Puig-Junoy (1998).

infermieri nonchè dal numero dei posti letto a disposizione. Quest'ultima variabile figura come una *proxy* del capitale impiegato nel processo produttivo; raramente infatti si ricorre alla valorizzazione degli *asset* patrimoniali per misurare il capitale. Nel caso della stima di una funzione di costo i prezzi degli input utilizzati sfruttano in genere dati contabili, opportunamente manipolati, per definire un "prezzo" per il fattore lavoro e per il fattore capitale.

I risultati dei lavori sul settore della sanità possono essere classificati in due grosse classi. Una prima classe comprende risultati metodologici, relativi alle procedure di stima ed all'utilizzo dei dati. In questo senso, le stime di inefficienza sembrano non essere influenzate nè dalla scelta di dati a livello di singolo paziente piuttosto che a livello di singola unità di produzione[37], nè dalla particolare metodologia di stima - D.E.A. o frontiera stocastica - adottata.[38] Nel caso poi della stima di una funzione di produzione, la funzione Cobb-Douglas sembra essere la migliore per descrivere il processo produttivo di servizi sanitari.[39]

Una seconda classe di risultati si concentra invece sulla descrizione delle caratteristiche del processo produttivo e sulle potenziali cause dell'inefficienza. Per quanto attiene alle caratteristiche del processo produttivo si osserva la possibile presenza di economie di scala per le case di riposo per anziani[40], mentre non può essere rigettata l'ipotesi contraria di rendimenti di scala costanti nel caso degli ospedali.[41] Economie di diversificazione, in particolare per le specialità di ginecologia e pediatria, sono rilevate ancora per gli ospedali.[42] Una relazione negativa tra efficienza e qualità del servizio è rilevata poi per le case di riposo.[43] Per quanto riguarda invece le potenziali cause dell'inefficienza, la letteratura si è concentrata in particolare sull'effetto di allocazioni alternative dei diritti di proprietà. I risultati sono contraddittori. Alcuni studi non riscontrano significative differenze nell'efficienza di differenti forme proprietarie[44]; altri trovano invece che unità di produzione pubbliche e nonprofit presentano un più alto livello di efficienza tecnica.[45]

[37] Si veda Zuckerman e al. (1994), p. 267.
[38] Si veda Linna (1998), p. 424.
[39] Si veda Gerdtham e al. (1999), p. 159.
[40] Cfr. Filippini (1998), p. 19.
[41] Cfr. Gerdtham e al. (1999), p. 160.
[42] Cfr. Prior (1996), p. 1300.
[43] Cfr. Kooreman (1994), p. 313.
[44] Cfr. Vitaliano - Toren (1994), p. 291.
[45] Cfr. in questo senso Zuckerman e al. (1994), p. 273 e Puig-Junoy (1998), p. 273. Filippini (1998) studia invece le differenze tra fondazioni, case di riposo consortili e comunali,

Una critica a questi risultati, generalmente ottenuti da regressioni degli *score* di inefficienza su variabili esplicative come la forma proprietaria, deriva dalla possibilità di ottenere stime distorte e non consistenti nel secondo *stage* di analisi.[46]

5 Un'applicazione al caso della Lombardia

La rassegna della letteratura empirica ha mostrato da un lato come non emerga un'unica metodologia per la misurazione dell'efficienza tecnica, dall'altro il ridotto numero di studi basati su dati italiani. In questo lavoro abbiamo utilizzato dati sul settore della sanità in Lombardia.[47] Gli obiettivi della ricerca sono almeno due. Innanzitutto, compiere un primo passo verso la comparazione di indicatori di efficienza ottenuti con metodologie alternative basandoci su dati italiani. In secondo luogo, studiare l'effetto sull'efficienza tecnica della struttura proprietaria delle imprese utilizzando tecniche econometriche.

5.1 La descrizione dei dati

La Regione Lombardia pubblica annualmente rapporti sul settore della sanità e sulle imprese che offrono servizi sanitari nella regione; tali rapporti consentono di ottenere dati riguardanti la singola unità produttiva.[48] Abbiamo ricostruito per il 1998 una *cross-section* di 107 imprese, sulle circa 200 censite nella regione, non considerando quelle unità produttive che presentavano dati mancanti o incompleti. A partire dai dati di queste imprese abbiamo stimato una *funzione di produzione di breve periodo* per i servizi sanitari.

Le variabili di input si riferiscono sostanzialmente al personale sanitario e misurano il numero di lavoratori coinvolti nel processo produttivo. Viene

trovando un minor livello di inefficienza nelle fondazioni.

[46] Sul punto si veda per esempio Gerdtham e al. (1999), p. 152, con i relativi riferimenti bibliografici.

[47] Altri lavori che utilizzano dati sulla sanità in Lombardia sono Giuffrida et. altri (1999a), (1999b) e (1999c).

[48] Cfr. Regione Lombardia (1999a) e Regione Lombardia (1999b).

considerato il numero di medici (MEDICI), di altro personale laureato (ALTRI), del personale impiegato in un ruolo didattico (DIDATT), di infermieri di prima (INF1) e di seconda categoria (INF2), del personale tecnico sanitario (TECN), del personale di riabilitazione (RIAB), di altro personale impiegato in un ruolo professionale di tipo sanitario (PROF). Tutte queste variabili sono state anche aggregate in un'unica variabile di input (PERSM), che misura il personale sanitario complessivo impiegato presso l'unità di produzione.

Le variabili di output considerano innanzitutto i giorni di degenza ospedaliera (DO), ponderati per un indice di complessità dei casi trattati determinato sulla base dei pesi assegnati ai D.R.G. prodotti da ogni struttura. Come precisato nel rapporto regionale, la scala dei pesi utilizzata è quella originale del programma americano "Medicare". Ne consegue che i pesi non tengono conto delle tariffe corrisposte dalla Regione Lombardia ma sono invece basati sulla complessità del paziente medio americano.[49] Altre variabili di output considerano poi i giorni di *Day Hospital* (DH), il numero delle prestazioni di pronto soccorso (PSOCC) ed il numero di prestazioni "equivalenti" di servizi diagnostici (DIAGNOS) prodotti dalle singole strutture. Anche nel caso degli output è stata costruita una variabile di sintesi (NORMA), che rappresenta la norma euclidea delle variabili di output.

La tabella 1 descrive le variabili utilizzate nelle elaborazioni. La figura di ospedale *medio* che ne emerge è quella di un'unità di produzione che impiega 116 medici, circa 300 infermieri, e altro personale per un numero complessivo di lavoratori pari a 485 unità, per produrre 80 mila giornate di degenza ospedaliera, più di 7 mila giorni di *Day Hospital*, più di 34 mila prestazioni di pronto soccorso e più di 1 milione di prestazioni "equivalenti" di servizi diagnostici. Come si può notare, la variabilità fra imprese è molto ampia per tutte le variabili considerate.

[49] Cfr. Regione Lombardia (1999a), p. 9.

Tabella 1. Statistiche descrittive delle variabili utilizzate

Variabile	**Media**	**Dev. Std.**	**Minimo**	**Massimo**	**Nr. oss.**
MEDICI	116	131.8	9	666	107
ALTRI	8.6	11.4	0	60	107
DIDATT	1.2	1.4	0	7	107
INF1	259.8	276.6	16	1488	107
INF2	42.3	48.6	1	356	107
TECN	41.2	47.1	0	265	107
RIAB	15.3	14.6	0	64	107
PROF	1.2	1.8	0	8	107
PERSM	485.7	509	32	2462	107
DO	80023.8	97706.2	7317.8	540813.7	107
DH	7487.5	11107	0	59841	107
PSOCC	34535.9	41266.5	0	187647	107
DIAGNOS	1005301.1	978051.8	18239	4409530	107
NORMA	1010691.5	982198.2	22755.2	4444406.7	107

I coefficienti di correlazione fra le variabili di input e di output sono raccolti nella tabella 2. Il numero dei medici risulta fortemente correlato per quanto riguarda gli input al numero degli altri laureati, al numero degli infermieri e al numero di altro personale tecnico sanitario; per quanto riguarda gli output alle giornate di degenza ospedaliera e di *Day Hospital*, nonchè al numero di prestazioni "equivalenti" di servizi diagnostici.

Tabella 2a. Matrice di correlazione delle variabili di input

	MED	ALTRI	DID	INF1	INF2	TECN	RIAB	PROF
MEDICI	1	.86	.49	.94	.87	.96	.53	.71
ALTRI		1	.63	.78	.79	.82	.51	.68
DIDATT			1	.48	.41	.49	.46	.47
INF1				1	.76	.98	.53	.66
INF2					1	.77	.41	.60
TECN						1	.53	.67
RIAB							1	.55
PROF								1

Tabella 2b. Matrice di correlazione delle variabili di output

	DO	DH	PSOCC	DIAGNOS
DO	1	.89	.51	.86
DH		1	.51	.86
PSOCC			1	.67
DIAGNOS				1

Tabella 2c. Matrice di correlazione tra le variabili di input e di output

	DO	DH	PSOCC	DIAGNOS
MEDICI	.93	.91	.62	.93
ALTRI	.78	.76	.47	.82
DIDATT	.45	.42	.29	.49
INF1	.95	.89	.58	.91
INF2	.76	.71	.55	.78
TECN	.94	.91	.56	.93
RIAB	.51	.51	.13	.47
PROF	.65	.64	.42	.65

Tabella 2d. Matrice di correlazione tra le variabili aggregate

	PERSM	NORMA
PERSM	1	.94
NORMA		1

5.2 I risultati con il metodo della D.E.A.

Abbiamo calcolato le misure di Debreu-Farrell (negli output) $D.F._o$ nel caso della sanità lombarda, utilizzando innanzitutto il metodo della D.E.A.

Per verificare la sensibilità dei risultati alla specificazione del modello, abbiamo testato tre differenti forme della frontiera dell'insieme di produzione, mantenendo costante l'ipotesi di *rendimenti di scala variabili*. Questa specificazione del modello consente di cogliere la natura dei rendimenti di scala, distinguendo tra rendimenti di scala crescenti e decrescenti. Il modello A considera separatamente sia gli output (DO, DH, PSOCC, DIAGNOS) che gli input (MEDICI, ALTRI, DIDATT, INF1, INF2, TECN, RIAB, PROF);

il modello B considera in modo aggregato sia gli output (NORMA) che gli input (PERSM); infine il modello C considera un solo output (NORMA) tenendo disaggregati gli input.[50]

Il modello di riferimento generale può essere rappresentato come:

$$\begin{aligned} \max_{\theta,\pi} \theta \qquad s.t. \quad \theta y_i - \mathbf{Y}\pi &\leq 0 \\ -x_i + \mathbf{X}\pi &\leq 0 \\ \mathbf{N}'\pi &= 1 \\ \pi &\geq 0 \end{aligned} \tag{32}$$

dove θ rappresenta la misura di Debreu-Farrell, π il vettore dei pesi che definiscono la frontiera efficiente, y_i e x_i gli output e gli input; infine $\mathbf{Y}$, $\mathbf{X}$ ed $\mathbf{N}$ rappresentano rispettivamente il vettore degli output, il vettore degli input ed un vettore unità per garantire la convessità della frontiera.[51]

I risultati appaiono sensibili alle differenti specificazioni del modello, confermando lo svantaggio della D.E.A. di produrre stime dell'inefficienza sensibili alla scelta delle variabili di input e di output.[52] Considerando la tabella 4, il valore medio del livello di efficienza (misurato attraverso l'indicatore di Debreu-Farrell, *orientato agli output*, $D.F._o$) è pari a 1,095 se consideriamo il modello A, a 1,467 se consideriamo il modello C, mentre risulta estremamente basso se guardiamo ai risultati del modello B (2,705). In quest'ultimo caso, le imprese operanti nella sanità lombarda potrebbero aumentare in media 2,705 volte l'output, senza alcuna variazione negli input.

5.3 I risultati con il metodo delle frontiere stocastiche

La seconda metodologia per il calcolo delle misure di Debreu-Farrell $D.F._o$ che abbiamo applicato al caso della sanità lombarda, è quella delle frontiere stocastiche. Considerata l'elevata collinearità tra le variabili di input, abbiamo testato unicamente il modello

[50] E' evidente che i livelli assoluti degli *score* di inefficienza cresceranno all'aumentare del grado di aggregazione delle variabili. Tuttavia, ciò non inficia i risultati della nostra analisi perchè siamo interessati allo studio della correlazione tra gli *score* di inefficienza e non ai loro livelli assoluti.

[51] Cfr. Coelli (1996), pp. 17-18.

[52] Cfr. Kalirajan - Shand (1999), p. 167.

$$\ln NORMA_i = \beta_0 + \beta_1 \ln PERSM_i + \ln \varepsilon_i \tag{33}$$

dove il termine d'errore composto $\ln \varepsilon_i = v_i - u_i$. La parte di errore non legata all'inefficienza è, per ipotesi, distribuita come una variabile casuale normale $v_i \sim \mathbf{N}(0, \sigma_u^2)$, mentre il termine legato all'inefficienza segue una distribuzione *half-normal* $u_i \sim \mathbf{N}(\mu, \sigma_v^2)$. L'utilizzo del *package* statistico LIMDEP 7.0 consente di ottenere stime della componente di inefficienza u_i attraverso l'applicazione della procedura descritta nel par. 3.2. E' attraverso queste stime che possiamo poi definire la versione econometrica della misura di Shephard[53]:

$$D_o = \frac{f(x_i, \beta) \exp\{\varepsilon_i\}}{f(x_i, \beta) \exp\{v_i\}} = \exp\{-u_i\} \tag{34}$$

da cui possiamo ricavare la misura di Debreu-Farrell:

$$D.F._o = \frac{1}{\exp\{-u_i\}} \tag{35}$$

I risultati delle stime sono contenuti nella tabella 3, dove vengono riportati per completezza anche i valori di partenza delle iterazioni per la M.L.E., ottenuti con il metodo dei Minimi Quadrati Ordinari. I coefficienti stimati sono significativi. La varianza della parte di errore legata all'inefficienza u_i è pari all'82% della varianza complessiva.[54] Seguendo Linna (1998), p. 422-423, abbiamo voluto testare l'ipotesi $H_0 : \lambda = 0$ per verificare se la distribuzione dell'errore u_i fosse coerente con le nostre ipotesi. La statistica LR è risultata pari a 4,92. Poichè il valore critico della distribuzione $\chi^2_{(1)}$ è uguale a 3,84, possiamo rigettare H_0 ad un livello di significatività del 5%. Non troviamo quindi evidenza per rifiutare l'ipotesi di *half-normal distribution* del termine di errore u_i.

[53] Cfr. Lovell (1993), p. 20.

[54] E' facile vedere che $\gamma = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_v^2} = 0.82$. Come messo in evidenza da Linna (1998), p. 423, "the efficiency estimates are more precise and confidence intervals narrower when ... σ_u^2 is large compared to σ_v^2".

Tabella 3. Risultati delle stime con il metodo delle frontiere stocastiche

Var. dipendente	$\ln NORMA$	$\ln NORMA$
costante	8.05 (29.841) ***	8.79 (29.636) ***
$\ln PERSM$	0.93 (20.082) ***	0.88 (17.195) ***
R^2	0.79	-
$\overline{R}^2$	0.79	-
$F - test$	403.30 $F_{(1,105)}$	-
$\lambda = \sigma_u/\sigma_v$	-	2.132 (3.546) ***
$\sigma = \sqrt{\sigma_u^2 + \sigma_v^2}$	-	0.633 (10.918) ***

(O.L.S. e M.L.E.; t-test in parentesi; liv. di signific.: * 10%, ** 5%, *** 1%)

Dalla tabella 4, il livello di efficienza medio stimato risulta sostanzialmente modesto. In accordo con la misura di Debreu-Farrell, le imprese operanti nella sanità lombarda potrebbero aumentare, in media, di 1,652 volte gli output senza modificare il livello di input.

La tabella 5 presenta i coefficienti di correlazione tra gli indicatori di efficienza ottenuti con il metodo della D.E.A., sulla base delle tre specificazioni del modello (DF_{DEAA}, DF_{DEAB} e DF_{DEAC} ad indicare rispettivamente gli indicatori determinati con il modello A, B e C), e con il metodo delle frontiere stocastiche (DF_{SF}). Il livello di correlazione si presenta contenuto all'interno delle stime ottenute con la D.E.A. Se consideriamo invece la correlazione fra gli indicatori ottenuti con la D.E.A. e gli indicatori ottenuti con la metodologia econometrica delle frontiere stocastiche, possiamo notare una correlazione positiva elevata per gli indicatori derivati dal modello B (0,81). L'utilizzo di metodologie diverse *utilizzando gli stessi input e gli stessi output* porta quindi a misure dell'inefficienza sostanzialmente coincidenti.

La stima dell'inefficienza sembra quindi essere sensibile a specificazioni alternative dei modelli (quindi alla scelta delle variabili di input e di output). La scelta della metodologia per la definizione della *"best practice" frontier* sembra invece condurre a risultati coerenti, *a parità di variabili di input e di output utilizzate.*

I risultati da noi ottenuti sembrano essere in linea con quelli di Linna (1998), che riscontrava una forte correlazione positiva tra le stime di inefficienza ottenute perseguendo approcci metodologici alternativi come la D.E.A. e le frontiere stocastiche.

Tabella 5. Matrice di correlazione fra le misure di Debreu-Farrell ottenuti con il metodo della D.E.A. e con il metodo delle frontiere stocastiche

	DF_{SF}	DF_{DEAA}	DF_{DEAB}	DF_{DEAC}
DF_{SF}	1	.13	.81	.49
DF_{DEAA}		1	.21	.46
DF_{DEAB}			1	.59
DF_{DEAC}				1

5.4 Le determinanti dell'inefficienza

La stima di *score* di inefficienza rappresenta il primo necessario passo per indagare sulle cause dell'inefficienza. In particolare, la composizione del nostro campione di imprese operanti nella sanità lombarda sembra prestarsi molto bene per testare l'effetto della forma proprietaria nonprofit sull'efficienza tecnica delle imprese. I dati raccolti dalla Regione Lombardia consentono di distinguere sei differenti forme organizzative. In questo lavoro abbiamo considerato gli I.R.C.C.S. pubblici, le Aziende Ospedaliere e gli ospedali a gestione A.S.L. come *imprese pubbliche*; gli I.R.C.C.S. privati e gli ospedali classificati come *imprese private nonprofit*; infine le case di cura come *imprese private forprofit*. Il nostro campione di 107 unità di produzione risulta così costituito principalmente da imprese pubbliche, da 6 imprese nonprofit e da 2 case di cura forprofit.

Abbiamo stimato con il metodo dei Minimi Quadrati Ordinari il modello generale:

$$\begin{aligned} D.F._{ji} &= \beta_0 + \beta_1 LETTI_i + \beta_2 DUMNPO_i + \\ &\quad + \beta_3 DUMDIDATT_i + \varepsilon_i; \\ j &= SF, DEAA, DEAB, DEAC \end{aligned} \tag{36}$$

dove la variabile $LETTI$ misura il numero di posti letto utilizzati dall'ospedale nel processo produttivo e può essere considerata una *proxy* del capitale impiegato [55]; $DUMNPO$ rappresenta una variabile *dummy* che assume valore 1 quando l'unità di produzione presenta la struttura proprietaria

[55] Poichè il nostro periodo di osservazione è un solo anno, seguendo Lovell (1993), p. 53, abbiamo incluso l'input capitale esclusivamente nell'analisi di secondo stadio.

dell'impresa nonprofit; $DUMDIDATT$ rappresenta una variabile *dummy* che assume valore 1 quando parte del personale medico svolge anche compiti didattici; infine ε_i è il termine di errore che rispetta le ipotesi standard dei Minimi Quadrati Ordinari.

I risultati delle stime sono raccolti nella tabella 6. Le relazioni stimate utilizzando gli *score* di efficienza ottenuti attraverso differenti metodologie e differenti aggregazioni delle variabili di input e di output sembrano produrre risultati coerenti tra loro, sia in termini di significatività, che per quel che riguarda i coefficienti associati ai regressori. L'inefficienza sembra ridursi al crescere del numero di posti letto impiegati dalle singole unità di produzione. Il risultato lascia intendere l'esistenza di "economie di scala" nella produzione di servizi ospedalieri; un incremento del numero di prestazioni erogate tenderebbe dunque a ridurre l'inefficienza tecnica dell'unità produttiva.

La struttura proprietaria dell'unità di produzione sembrerebbe non esercitare alcuna influenza sulle *performance* di efficienza. Nell'unico caso in cui il coefficiente associato alla *dummy* relativa alla forma proprietaria risulta significativo, le organizzazioni con forma proprietaria senza fini di lucro parrebbero tuttavia relativamente più efficienti rispetto alle altre unità produttive. Il risultato contrasta, ad esempio, con quello di Valdmanis (1992) che trova invece un livello di efficienza tecnica più elevato per gli ospedali pubblici. Il numero limitato di imprese nonprofit considerato nel nostro campione e, per converso, il numero limitato di imprese pubbliche considerate in Valdmanis (1992)[56] consiglia di leggere con cautela le conclusioni dell'analisi di efficienza per quanto riguarda gli effetti della struttura proprietaria.

Infine, la presenza di personale dedicato allo svolgimento di attività didattica sembrerebbe influenzare negativamente la performance delle organizzazioni.

6 Conclusioni

Il lavoro porta a risultati che riteniamo di un certo interesse da due diversi punti di vista: quello metodologico - relativo alla consistenza ed alla validità delle diverse procedure di stima - e quello di merito, che fa invece riferimento alle possibili cause dell'inefficienza tecnica.

[56] Il campione considerato in Valdmanis (1992) è composto da 33 ospedali privati nonprofit e da soli 8 ospedali pubblici.

Dal primo punto di vista abbiamo innanzitutto osservato, come evidenzia bene la tabella 4, che i valori medi dei livelli di efficienza misurati attraverso gli indicatori di Debreu-Farrell ottenuti con il metodo della D.E.A. divergono sensibilmente nel caso in cui le variabili di input e di output vengano specificate diversamente nel modello. Coerentemente con la letteratura, questo risultato sembra dunque confermare una debolezza specifica della metodologia di stima dell'efficienza basata sulla D.E.A., rappresentata dalla sua sensibilità alle modalità con cui vengono specificati gli input e gli output del modello.

In secondo luogo, si osserva come le stime di efficienza che si ottengono facendo uso di diverse metodologie di calcolo (D.E.A. e frontiere stocastiche) risultino tra loro fortemente correlate qualora le variabili di input e di output dei modelli siano specificate nello stesso modo (come nel caso degli indicatori ottenuti con il metodo delle frontiere stocastiche e con il modello B della D.E.A. in tabella 5). Questo risultato è confortante poiché consente di fondare su basi piuttosto solide le successive analisi delle cause di inefficienza.

Per quello che riguarda queste ultime, paiono emergere alcuni risultati interessanti. In primo luogo, il segno dei coefficienti stimati sembrerebbe individuare una relazione diretta tra l'efficienza tecnica e la dimensione delle imprese, misurata attraverso il numero di posti letto impiegati nella produzione, una *proxy* del fattore capitale. Pare dunque rilevarsi una sorta di effetto di scala che accresce l'efficienza tecnica delle organizzazioni al crescere del numero di prestazioni erogate nelle diverse aree di attività.

La struttura proprietaria sembrerebbe non avere alcun effetto sull'efficienza tecnica delle imprese. Tuttavia, nell'unico caso in cui il coefficiente stimato risulta statisticamente significativo, la forma dell'impresa nonprofit parrebbe esercitare un'influenza positiva sui livelli di efficienza.

Infine, la presenza di personale dedicato all'attività didattica sembrerebbe influenzare negativamente l'efficienza tecnica.

Tabella 4. Statistiche descrittive delle misure di Debreu-Farrell ottenute con metodologie alter

Misure	**Media**	**Dev. std.**	**Min.**	**25° perc.**	**Mediana**	**75° perc.**	**Mass.**	**Nr. c**
DF_{SF}	1.652	.687	1.084	1.298	1.477	1.677	5.687	107
DF_{DEAA}	1.095	.228	1	1	1	1.066	2.519	107
DF_{DEAB}	2.705	1.785	1	1.65	2.188	3.279	13.514	107
DF_{DEAC}	1.467	.932	1	1	1.112	1.577	7.692	107

Tabella 6. Le determinanti dell'inefficienza

Var. dipendente	$\mathbf{DF}_{SF}$	$\mathbf{DF}_{DEAA}$	$\mathbf{DF}_{DEAB}$
costante	1.63 (15.251) ***	1.037 (81.495) ***	2.618 (15.907)
$LETTI$	-0.00221 (-2.178) **	-0.00106 (-1.999) **	-0.0149 (-4.772
$DUMNPO$	0.805 (1.220)	-0.143 (-3.008) ***	2.459 (1.394)
$DUMDIDATT$	0.746 (0.632)	0.155 (3.158) ***	0.6756 (2.493)
Breusch-Pagan test $[H_0 : \sigma_i^2 = \sigma^2 \quad \forall i]$	90.3954 *** $\chi^2_{(3)}$	49.1901 *** $\chi^2_{(3)}$	136.4679 *** χ
LR test $\left[H_0 : \beta_j = 0; \quad j = 1, .., 3\right]$	10.2982 ** $\chi^2_{(3)}$	10.3112 ** $\chi^2_{(3)}$	24.3062 *** $\chi^2_{(}$
Nr. oss.	107	107	107

(O.L.S.; S.E. corretti per eteroschedasticità con la procedura di White; t-statistici in parentesi; livello di s

References

[1] Angeloni L., G. Fiorentini (1996), *Analisi di efficienza per organizzazioni non-profit*, in Borzaga, Fiorentini, Matacena (eds.), pp. 261-298

[2] Bosco B., L. Parisio (1996), *Efficienza nella produzione pubblica di beni e servizi*, La Nuova Italia Scientifica

[3] Borzaga C., G. Fiorentini, A. Matacena (eds.) (1996), *Non-profit e sistemi di welfare*, NIS

[4] Coelli T. (1996), A guide to DEAP Version 2.1: a Data Envelopment Analysis (Computer) Program, *Centre for Efficiency and Productivity Analysis Working Paper 96/08*, University of New England

[5] Douglas J. (1983), *Why charity? The case for a third sector*, Sage

[6] Filippini M. (1998), Efficienza di costo nell'offerta di servizi assistenziali residenziali per anziani, *Economia Pubblica*, n. 5, pp. 5-25

[7] Fried H. O., C. A. K. Lovell, S. S. Schmidt (eds.) (1993), *The measurement of productive efficiency*, Oxford University Press

[8] Gerdtham U.-G., Löthgren M., Tambour M., Rehnberg C. (1999), Internal markets and health care efficiency: a multiple-output stochastic frontier analysis, *Health Economics*, vol. 8, pp. 151-164

[9] Giuffrida A., H. Gravelle (1999), Measuring performance in primary care: econometric analysis and DEA, *University of York Discussion Paper*, n. 99/36

[10] Giuffrida A., Lapecorella F., Pignataro G. (1999a), Analisi dell'efficienza tecnica e di scala dei servizi ospedalieri, Atti del IV Workshop di Economia Sanitaria "Autonomia regionale e sistema sanitario", Dipartimento di Scienze Economiche "G. Prato", *Quaderni dell'Università degli Studi di Torino*, n. 39

[11] Giuffrida A., Lapecorella F., Pignataro G. (1999b), Incentivi all'efficienza nel sistema sanitario italiano: analisi dell'efficienza delle aziende ospedaliere e presidi ospedalieri dopo la riforma, *Economia Pubblica*, forthcoming

[12] Giuffrida A., Lapecorella F., Pignataro G. (1999c), Efficiency of health care production in different hierarchically structured hospitals, *Technical Discussion Paper Series*, Centre for Health Economics, University of York, n. 14

[13] Greene W. H. (1993), *The econometric approach to efficiency analysis*, in Fried, Lovell, Schmidt (eds.), pp. 68-119

[14] Grossman S. (1972), On the concept of health capital and the demand for health, *Journal of Political Economy*, vol. 80, n. 2, pp. 223-255

[15] Kalirajan K. P., R. T. Shand (1999), Frontier production functions and technical efficiency measures, *Journal of Economic Surveys*, vol. 13, n. 2, pp. 149-172

[16] Kooreman P. (1994), Nursing home care in The Netherlands: a non-parametric efficiency analysis, *Journal of Health Economics*, vol. 13, pp. 301-316

[17] Linna M. (1998), Measuring hospital cost efficiency with panel data models, *Health Economics*, vol. 7, pp. 415-427

[18] Lovell C. A. K. (1993), *Production frontiers and production efficiency*, in Fried, Lovell, Schmidt (eds.), pp. 3-67

[19] Marmor T., M. Schlesinger, R. Smithey (1987), Nonprofit organizations and health care, in W. Powell (ed.), *The nonprofit sector. A research handbook*, Yale University Press

[20] Puig-Junoy J. (1998), Technical efficiency in the clinical management of critically ill patients, *Health Economics*, vol. 7, pp. 263-277

[21] Prior D. (1996), Technical efficiency and scope economies in hospitals, *Applied Economics*, vol. 28, pp. 1295-1301

[22] Solimene L. (1994), *Regolamentazione dei mercati, efficienza e produttività*, Giuffrè

[23] Regione Lombardia (1999a), *Ricoveri in Lombardia 1998*

[24] Regione Lombardia (1999b), *Rilevazione delle attività gestionali delle ASL e delle Aziende Ospedaliere. Flussi informativi di cui al D.M. 23 dic. 1996 per l'anno 1998*

[25] Simar L., P. W. Wilson (1999), Statistical inference in nonparametric frontier models: the state of the art, *Institut de statistique Discussion Paper*, 99.04

[26] Tami L. M. (1996), Do nonprofit organizations offer advantages in markets characterized by asymmetric information?, *Journal of Human Resources*, vol. 31, n. 3

[27] Tronti L. (1991), Gli indicatori di produttività dei servizi pubblici. Istruzione, sanità e amministrazione locale, *Economia & Lavoro*, XXV, n. 3, pp. 89-112

[28] Valdmanis V. (1992), Sensitivity analysis for DEA models An empirical example using public vs. NFP hospitals, *Journal of Public Economics*, vol. 48, pp. 185-205

[29] Vitaliano D. F., M. Toren (1994), Cost and efficiency in nursing homes: a stochastic frontier approach, *Journal of Health Economics*, vol. 13, pp. 281-300

[30] Zuckerman S., J. Hadley, L. Iezzoni (1994), Measuring hospital efficiency with frontier cost functions, *Journal of Health Economics*, vol. 13, pp. 255-280