# Capitolo Primo

## L'evoluzione del concetto e della misura dell'efficienza produttiva

### 1.1 I concetti di produttività ed efficienza

Nella letteratura economica e statistica, i termini "produttività" ed "efficienza", che riflettono i due concetti comunemente utilizzati per caratterizzare l'abilità di una impresa nell'utilizzazione delle risorse, sono spesso impiegati come sinonimi. In realtà, essi denotano concetti in parte diversi, la cui misurazione può condurre ad indicatori cui sono associate diverse interpretazioni.

Un *indicatore di produttività* può essere definito mediante il rapporto tra il risultato dell'attività produttiva e i fattori impiegati per ottenerlo (Nisticò e Prosperetti, 1991). Facendo riferimento alla terminologia anglosassone, ormai invalsa nell'uso e di cui si farà particolare ricorso anche in seguito, il risultato dell'attività produttiva o prodotto viene definito *output*, mentre i fattori utilizzati nel processo per il suo ottenimento sono denominati *input*.

Come rileva Kuznets (1990), il concetto di produttività si compone di tre elementi: gli output, gli input e, infine, il processo tecnologico attraverso cui i primi due elementi sono connessi tra loro. La traduzione di questo concetto in strumenti di misurazione analitici implica che gli aggregati coinvolti, gli input e gli output, debbano essere noti e misurabili e, quindi, per la loro determinazione è necessario conoscere il processo di conversione dei primi nei secondi.

Il richiamo ad una teoria della produzione diventa, perciò, essenziale per la determinazione dei criteri in base ai quali confrontare i singoli output e i singoli input. La valutazione della produttività offre, perciò, la possibilità di impostare diverse analisi della struttura e del funzionamento di un'organizzazione economica, per valutare gli obiettivi raggiunti in rapporto ai mezzi utilizzati (Guarini e Tassinari, 1990).

Laureti Tiziana, L'efficienza rispetto alla frontiera delle possibilità produttive
ISBN-10: 88-8453-426-7 ISBN-13: 978-88-8453-426-2 

In termini molto generali, la *misura dell'efficienza* di un'unità produttiva può essere definita per confronto tra il processo di produzione effettivamente realizzato e un altro processo, opportunatamente scelto, corrispondente a uno standard di ottimalità, che può avere valenza nel tempo e nello spazio (Petretto, 1986).

È chiaro, quindi, che anche il concetto di efficienza è legato a una teoria della produzione. Come sottolineato dal suddetto Autore, la misurazione dell'efficienza deve avvenire tramite l'esame dell'evoluzione di indicatori specifici in una prospettiva *time series*, oppure attraverso il confronto degli indicatori con valori standard degli stessi, secondo una prospettiva *cross-section.*

Al fine di evidenziare gli elementi di diversità tra il concetto di efficienza e quello di produttività, si possono considerare come standard di riferimento i processi efficienti espressi dalla funzione di produzione, che indica, appunto, il massimo prodotto ottenibile da un dato livello di fattori produttivi, considerando la tecnologia esistente.

In una prospettiva *cross section*, ossia in riferimento a dati relativi a diverse unità produttive rilevate nello stesso momento, riferendosi al caso semplificato di due processi produttivi, *A* e *B*, concernenti due unità produttive, in cui viene utilizzato un unico input, *x*, che assume, rispettivamente, le specificazioni $x_A$ e $x_B$, per produrre un solo output, *y*, la cui quantità è espressa, rispettivamente, da $y_A$ e $y_B$, si può, innanzitutto, misurare la produttività di *A* e di *B* mediante il rapporto $P_A = y_A / x_A$ e $P_B = y_B / x_B$.

Se $P_A > P_B$ si conclude che il processo *A è più produttivo del processo B,* senza dover ricorrere a informazioni sulla tecnologia di produzione. Un indice di produttività di *A* relativo a *B* può essere facilmente costruito mediante il rapporto:

$$\frac{P_A}{P_B} = \frac{y_A / x_A}{y_B / x_B}.$$

Ipotizzando, invece, che la tecnologia sia descritta da una funzione di produzione, $y^* = f(x)$, si può individuare il massimo output producibile a partire dall'input $x_A$, ossia $y_A^* = f(x_A)$ e quello ottenibile utilizzando il livello $x_B$, ossia $y_B^* = f(x_B)$. In tal modo è possibile definire l'*Efficienza Tecnica (ET)* di ciascuna unità produttiva, confrontando l'output effettivamente prodotto con la quantità massima di output producibile a partire dalla quantità osservata dell'input, ottenendo misure di efficienza orientate nel senso degli output, definite, utilizzando la terminologia anglosassone, *output-oriented.*

Si può affermare che mentre la produttività è una misura assoluta, l'efficienza tecnica è una misura relativa. Nell'esempio considerato, si ottiene, per l'unità *A*, la misura di efficienza tecnica:

$$ET_A = \frac{y_A}{y_A^*} \leq 1$$

mentre per l'unità *B* l'efficienza tecnica è espressa dal rapporto:

$$ET_B = \frac{y_B}{y_B^*} \leq 1$$

Va osservato che, se un'unità produttiva, ad esempio *A,* producesse il massimo output ottenibile a partire dall'input utilizzato, $x_A$, la sua produttività sarebbe espressa da $P_A^* = y_A^* / x_A$ . La misura di efficienza tecnica dell'unità *A*, può, quindi, essere espressa anche come indice di produttività relativo ad una unità ipotetica che produca il massimo output ottenibile data la tecnologia descritta dalla funzione di produzione $y^* = f(x)$, utilizzando il medesimo livello di input $x_A$, ossia mediante:

$$ET_A = \frac{y_A}{y_A^*} = \frac{P_A}{P_A^*} = \frac{y_A / x_A}{y_A^* / x_A}$$

In modo del tutto analogo si può definire la misura di efficienza di *B* attraverso l'espressione:

$$ET_B = \frac{y_B}{y_B^*} = \frac{P_B}{P_B^*} = \frac{y_B / x_B}{y_B^* / x_B}$$

La Figura 1.1 illustra graficamente i concetti precedenti: la produttività di *A* è uguale alla pendenza della retta *OA*, mentre quella di *B* è espressa dalla pendenza del retta *OB*. La determinazione dell'efficienza delle due unità considerate si basa sulla conoscenza dei punti *A** e *B**, che esprimono il massimo output producibile, descritto dalla funzione di produzione $y^* = f(x)$, a partire, rispettivamente, dalle quantità di input, $x_A$ e $x_B$, effettivamente impiegate dalle due unità produttive.

Va osservato che, nel confrontare più processi di produzione attraverso un'analisi *cross-section,* un differenziale di produttività non comporta necessariamente un differenziale di efficienza e, d'altra parte, un differenziale di efficienza può non associarsi a un differenziale di produttività, salvo che la funzione di produzione sia caratterizzata da rendimenti di scala costanti, caratteristici di una tecnologia nella quale ad un incremento *t* dell'input corrisponde un pari aumento *t* dell'output[1]. In tal caso i due concetti sono equivalenti (Nisticò e Prosperetti, 1991).

Continuando l'esempio precedente, nella Figura 1.2a, i due processi rappresentati sono entrambi pienamente efficienti poiché giacciono ambedue sulla funzione di produzione, che presenta rendimenti di scala decrescenti non uniformemente, per cui ad un incremento dell'input non corrisponde un incremento proporzio-

[1] Un'analisi formale dei rendimenti di scala è effettuata nel paragrafo 2.7.

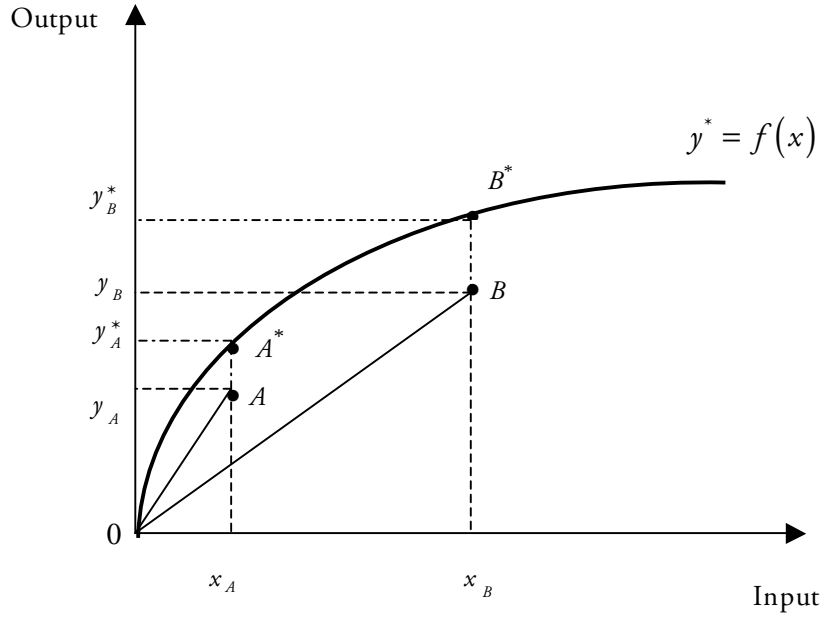


Figura 1.1 – Produttività ed efficienza.

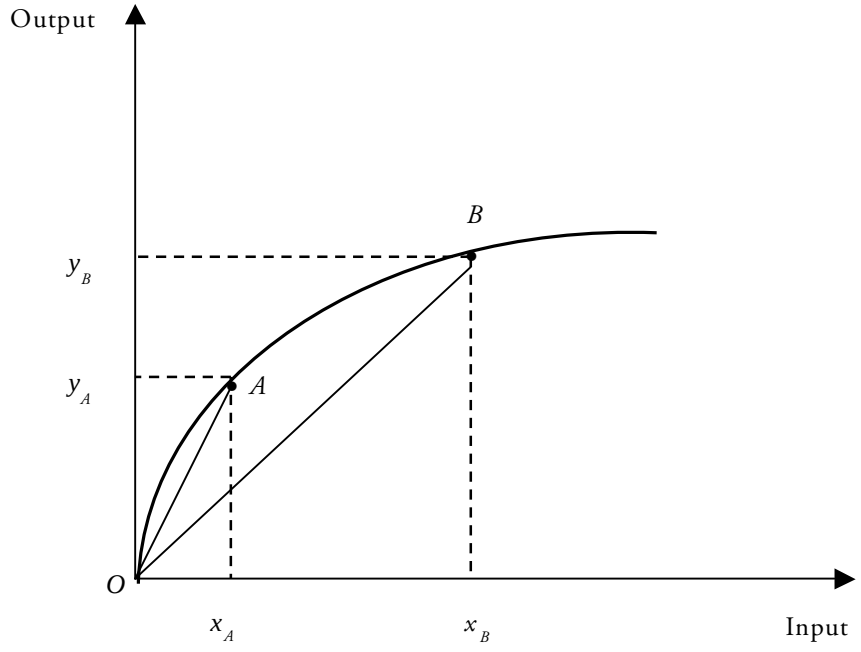


Figura 1.2a – Produttività ed efficienza: *identica efficienza diversa produttività.*

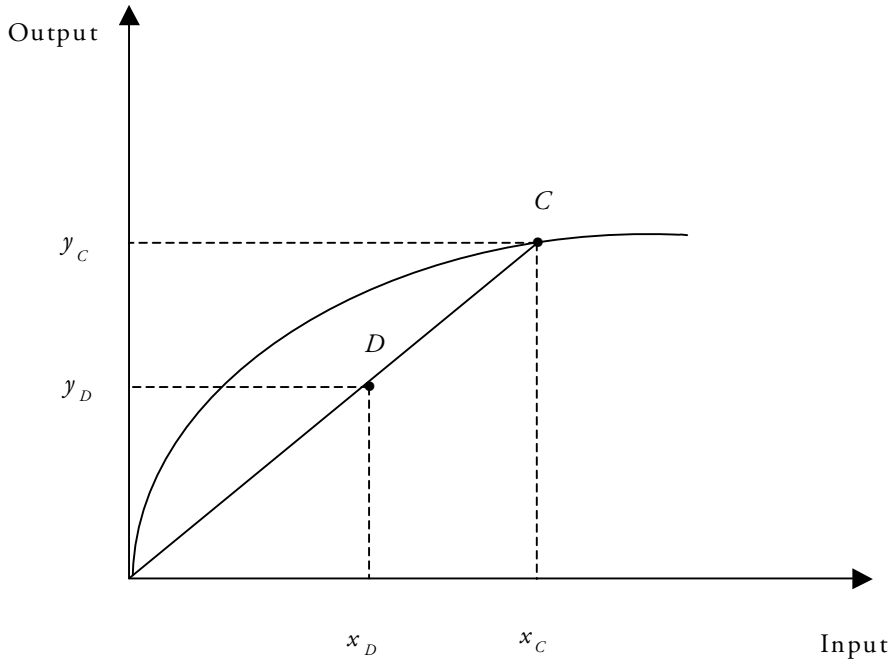


Figura 1.2b – Produttività ed efficienza: *identica produttività diversa efficienza.*

nale dell'output, ma la produttività dell'unità *A* risulta superiore a quella di *B*. Quindi, in questo caso semplificato, che coinvolge il confronto di due soli processi, ad un identico livello di efficienza, corrispondono due livelli di produttività.

La situazione opposta, caratterizzata da una situazione in cui le due unità produttive valutate presentano identici livelli di produttività ma diversa efficienza, è rappresentata nella figura 1.2b.

In quest'ultima situazione, la produttività delle due unità *C* e *D* è identica ed è misurata, rispettivamente, dai rapporti $y_C/x_C$ e $y_D/x_D$, dove $y_C/x_C = y_D/x_D$. Ma l'unità *D*, posizionandosi al di sotto della funzione di produzione, presenta una minore efficienza rispetto a *C*, che invece giace sulla funzione di produzione.

L'impiego di un indicatore di produttività come *proxy* di una misura di efficienza può condurre, quindi, a risultati fuorvianti nei casi in cui i rendimenti di scala non possono considerarsi costanti. Nonostante ciò, misure di produttività media dei fattori, come, ad esempio, il lavoro, sono state ampiamente utilizzate nelle applicazioni empiriche in qualità di indici di efficienza.

Come osservano Gazzei, Lemmi e Viviani (1997), diversi Autori, soprattutto nell'ambito di studi di carattere microeconomico, esprimono il concetto di

produttività in forma strettamente dipendente da quello di efficienza, tanto da utilizzarlo, talvolta, nella definizione stessa.

Solo per fornire qualche esempio, si possono considerare le interpretazioni di Schmookler (1952), secondo cui la produttività è "un indice di efficienza", e quella di Kendrick (1956), per il quale la produttività è un "cambiamento nell'efficienza produttiva".

Zappa (1957) sottolinea che la "produttività è spesso nota come efficienza economica e che l'efficienza è non di rado conosciuta come produttività tecnica", mentre Hatry (1982) sostiene che "l'efficienza riguarda il rapporto tra risorse utilizzate e quantità di beni o servizi prodotta. Le misure di efficienza si definiscono come la relazione tra la quantità di risorse impiegate ed il prodotto che ne deriva. Il rapporto tra prodotto e risorse impiegate viene chiamato produttività. Il suo inverso, il rapporto risorse/produzione, è definito efficienza o costo unitario. Le due forme sono equivalenti".

L'analisi del concetto di efficienza e il conseguente sviluppo di strumenti metodologici adeguati alla misurazione, che costituiscono l'oggetto del presente lavoro, prendono avvio da Farrell (1957), che, come si vedrà più avanti, pose l'accento sul problema consistente nell'individuare la tecnologia di riferimento rispetto alla quale definire, poi, un'appropriata misura di efficienza[2].

La funzione di produzione appare così come una *frontiera* tra i processi tecnicamente possibili, che giacciono al di sotto di essa, e quelli impossibili (Thiry e Tulkens, 1987).

L'efficienza di un'impresa può essere, quindi, valutata considerando la posizione del processo produttivo posto in essere rispetto alla frontiera delle possibilità di produzione.

## 1.2 Il concetto di efficienza nella teoria microeconomica

L'efficienza, considerata in un'accezione generale che include anche il concetto di efficienza tecnica appena discusso, rappresenta una nozione estremamente importante nella teoria economica e in particolare nella teoria della produzione.

---

[2] I primi approcci alla misurazione della produttività si basano sulla costruzione di indici parziali, definiti come rapporto tra quantità prodotta ed uno solo dei fattori produttivi utilizzati. Solow (1957) pose le basi per lo sviluppo di indici di produttività globali, ottenuti rapportando il prodotto all'insieme dei fattori impiegati.

Le moderne teorie della produzione si rifanno agli approcci tradizionali, ossia a quello marginalista e a quello del sovrappiù. Dopo la pubblicazione del saggio di Sraffa (1960), è stato ripreso il concetto di produzione inteso come fenomeno circolare, ossia come *riproduzione*, che ha stimolato nuovo interesse per l'analisi della produzione tipica degli economisti classici e di Marx.

Nel modello fondi-flussi di Georgescu-Roegen (1973), si propone un'analisi della produzione che concentra l'attenzione sulla durata dei processi produttivi, mettendo in evidenza la possibilità che un dato procedimento di fabbricazione possa essere realizzato con differenti forme di attivazione, ognuna delle quali connessa con uno specifico problema di utilizzazione efficiente di elementi fondo della produzione, al fine di eliminare, o attenuare, eventuali sottoutilizzazioni che si riflettono nell'efficienza.

Nel contesto dell'impostazione neoclassica, nasce una teoria della produzione della scuola di Losanna, grazie ai contributi di vari Autori, quali Walras (1874) e Pareto (1906), i quali focalizzano la loro attenzione sull'intero sistema economico e sul suo equilibrio generale piuttosto che sulla singola unità di produzione, il cui processo decisionale e organizzativo viene perciò relegato in un ruolo sussidiario e considerato come una sorta di scatola nera il cui contenuto rimane inesplorato. Allo sviluppo della scuola paretiana contribuiscono diversi Autori, tra i quali Samuelson (1947) e Frisch (1965) ed in particolare, per l'uso della teoria della dualità, Shephard (1953) e McFadden (1966).

Il modello di *analisi delle attività,* o nella terminologia anglosassone *activity analysis*, rappresenta un modo alternativo di esaminare i processi di produzione e costituisce una formulazione particolare all'interno del più generale modello degli insiemi di produzione.

In questo contesto, il problema della scelta tra possibilità alternative è direttamente espresso come problema di utilizzazione ottimale di risorse date; inoltre, si fa esplicito riferimento allo strumento della programmazione lineare, introdotto precedentemente da von Neumann (1938, 1945).

L'associazione di ciascuna situazione efficiente con un vettore di prezzi trova una formulazione particolarmente interessante, essendo definita una procedura di determinazione di tali valori. Lo sviluppo del modello di analisi delle attività si deve soprattutto ai contributi di Koopmans (1951) e Debreu (1959), i quali rappresentano due esponenti della cosiddetta teoria della produzione *neo-walrasiana*, sviluppatasi dopo la seconda guerra mondiale.

Nella formulazione neoclassica della teoria della produzione, si utilizza prevalentemente una rappresentazione analitica del processo produttivo, definito come un processo di trasformazione regolato da uomini (Frish, 1965) e descritto mediante una $n$-pla di numeri reali, $\mathbf{z}$, gli elementi negativi della quale misu-

rano gli input del processo, **x**, mentre gli elementi positivi ne misurano gli output, **y**, oppure mediante una coppia di vettori input-output (**x**, **y**).

La costruzione di un modello di scelta del processo "ottimo" è basata sul criterio di massimo risultato netto, o in via subordinata, di minimo costo (Tani, 1986). Pertanto, ad ogni processo deve essere associata una valutazione del risultato netto, oppure del costo del vettore **x** degli input.

Nel caso solitamente considerato, ossia nel cosiddetto modello di produzione classico, si suppone dato un vettore di prezzi dei fattori, **w**, e dei prodotti, **p**, ai quali è possibile acquistare gli input e vendere gli output, in qualunque quantità.

Il costo di produzione relativo al processo (**x**, **y**) è definito, quindi, dal prodotto **wx,** mentre il risultato netto del processo è ottenuto mediante l'espressione: **py – wx**.

Il modello utilizzato per la rappresentazione delle possibilità di scelta che la tecnica offre è costituito dalla funzione di produzione, che identifica per ogni possibile vettore di input la quantità massima di output producibile.

L'obiettivo del massimo profitto consente, dati i prezzi esistenti sul mercato, data la funzione di produzione, di determinare quale delle tecniche possibili possa essere adottata dall'imprenditore e questa scelta determina, simultaneamente, la quantità di output da produrre e le quantità dei fattori da impiegare.

In tal modo, risulta individuato il processo produttivo ottimo per ciascuna impresa, che consente ad ogni unità di operare in una situazione di *efficienza produttiva*. La medesima situazione si verifica per il sistema della produzione nel suo complesso, purché si consideri un'economia walrasiana perfettamente concorrenziale, in cui tutte le unità produttive risultano *price-takers.*

Un elemento estremamente importante da sottolineare è che la teoria della produzione neoclassica, in condizioni di perfetta concorrenza, studia il comportamento razionale delle unità produttive *prima* del reale svolgimento del processo produttivo.

La funzione di produzione è, infatti, parte di una teoria che richiede la determinazione contemporanea di tutte le incognite, e quindi dei prezzi dei fattori e delle quantità prodotte da ogni impresa, prima che ogni atto produttivo abbia luogo. Anche nel caso del lungo periodo è necessario supporre che tutte le modificazioni, che intervengono in relazione alla tecnologia e alle dotazioni presenti e future dei fattori, siano note al momento in cui le imprese decidono le modalità con cui allocare le loro risorse, all'inizio del processo produttivo (Vaggi, 1987).

Il grande vantaggio di questa impostazione, che ne giustifica anche il largo impiego nelle applicazioni a livello econometrico, risiede proprio nelle buone qualità analitiche che possono essere attribuite alle funzioni di produzione, le

quali consentono l'impiego di strumenti matematici potenti e di comode specificazioni econometriche.

Il limite maggiore di un tale impianto teorico consiste nel fatto che la tradizionale funzione di produzione non descrive tutte le possibili relazioni tecnologiche tra gli input e l'output, ma solo quelle tecnicamente output efficienti (Walters, 1963).

Va osservato comunque che, sebbene la funzione di produzione individui le tecniche efficienti, ciò non implica che vi sia un unico metodo efficiente di produzione che lega l'output ai fattori produttivi.

Le imprese possono, infatti, scegliere con quali proporzioni dei fattori produrre ogni data quantità di output; la funzione di produzione indica, per ogni *y*, i metodi produttivi che comportano una sostituzione fra i fattori, per cui, in generale, se si riduce un input, il livello del prodotto può restare inalterato solo se aumenta la quantità impiegata di almeno un altro fattore.

La scelta della particolare combinazione di fattori impiegata, che massimizza il risultato netto, dipende dai prezzi relativi dei fattori e dalle relazioni di sostituibilità tecnica, descritte dalla funzione di produzione.

La concezione neoclassica della produzione si fonda sull'idea di allocazione efficiente delle risorse a prezzi dati. L'ipotesi di continuità e differenziabilità della funzione di produzione consentono di determinare *quanto* produrre uguagliando la produttività marginale del fattore *j*-esimo al prezzo del fattore stesso in termini del prodotto, ossia:

$$\frac{\partial f(\mathbf{x})}{\partial x_j} = \frac{w_j}{p}$$

e *come* produrre scegliendo una particolare combinazione di input *j* e *s*, ossia:

$$\frac{x_j}{x_s}.$$

L'allocazione efficiente delle risorse implica anche che ogni fattore riceva parte del prodotto, in relazione al proprio contributo produttivo.

Spostando l'attenzione dall'analisi della singola impresa al sistema complessivo, l'allocazione delle risorse esistenti in una situazione di concorrenza perfetta viene effettuata attraverso le scelte degli imprenditori che determinano, non solo la massimizzazione dei profitti e la minimizzazione dei costi, ma anche il livello dei prezzi relativi dei fattori.

In una tale situazione si realizza, quindi, la situazione di *efficienza paretiana di tipo allocativo*, in relazione alla quale non può esistere alcun modo di aumentare la soddisfazione di un individuo senza ridurre la soddisfazione di qualcun altro.

Contro l'impostazione della teoria neoclassica, che identifica il concetto di efficienza con quello di efficienza allocativa derivante dalle condizioni concorrenziali che caratterizzano il mercato, si muove il lavoro di Leibestein (1966), che introduce un nuovo concetto di efficienza connesso alle decisioni interne all'impresa in relazione alla scelta dei processi produttivi.

Il tentativo è diretto al superamento delle limitazioni insite nell'utilizzazione della funzione di produzione che incorpora, da quanto emerso in precedenza, non solo le caratteristiche tecniche dei processi di produzione, ma anche un'ipotesi di comportamento dell'agente, le cui decisioni producono effetti nella gestione dell'attività produttiva dell'unità considerata.

Leibestein denomina questo concetto di efficienza "*efficienza x*" (*x-efficiency)* per indicare la capacità non dei mercati, ma dell'organizzazione aziendale di allocare le risorse in modo efficiente, ossia adottando un comportamento tale da rendere il saggio marginale di sostituzione tecnica uguale al rapporto tra i prezzi dei fattori, nonché di scegliere programmi di produzione tecnicamente efficienti.

Come già sottolineato in precedenza, nella teoria economica neoclassica una simile capacità viene normalmente presupposta come corollario della massimizzazione dei profitti, senza indagare sulle questioni organizzative dell'impresa.

Tuttavia, alcuni economisti erano intervenuti precisando che questo presupposto non è sempre valido nella realtà delle imprese non sottoposte alla pressione della concorrenza. L'allontanamento da una situazione di concorrenza perfetta e, quindi, l'esistenza sul mercato, ad esempio, di situazioni monopolistiche, può causare un'allocazione delle risorse non efficiente e, conseguentemente, la presenza di inefficienza allocativa.

Ad esempio Hicks (1935) con la sua "*quiet life*" del monopolista aveva affrontato la problematica, asserendo che l'assenza di pressione concorrenziale può indurre i produttori a non seguire pienamente un comportamento di massimizzazione del profitto ma, piuttosto, a posizionarsi in una situazione vicina a quella di massimizzazione.

La teoria della *efficienza x* di Leibestein fu criticata da Stigler (1976), il quale ricondusse il concetto di *x-efficiency* alla nozione classica di efficienza allocativa.

Secondo il suddetto Autore, l'esistenza di eventuali differenze nell'output prodotto da diverse imprese, che impiegano una medesima quantità di input, può essere dovuta al fatto che gli imprenditori considerati sono dotati di una diversa conoscenza tecnologica e ciò si riflette nell'utilizzazione di specifici input. L'effetto di tali variazioni viene attribuito interamente ai fattori produttivi, per cui il problema si traduce in una situazione di allocazione delle risorse non efficiente.

Al di là delle critiche, il merito della teoria della *x-efficiency* è stato quello di aver messo in luce l'esistenza di tipologie di comportamento del produttore,

definibili come razionali ma compatibili con situazioni anche tecnicamente inefficienti, conducendo gli studiosi a prendere sempre più in considerazione, anche nell'analisi empirica ed econometrica, la possibilità di osservare situazioni produttive che non si posizionano sulla funzione di produzione, e stimolando in tal modo l'analisi di situazioni non efficienti, la cui presenza era sostanzialmente negata nel contesto della teoria neoclassica.

Un ruolo fondamentale nello sviluppo dell'analisi dell'efficienza viene esercitato dall'*analisi delle attività*, attraverso la quale vengono introdotti strumenti alternativi alla funzione di produzione per affrontare i problemi di efficiente allocazione di risorse date fra produzioni alternative.

Infatti, la funzione di produzione consente di trattare solo i casi in cui, per un dato insieme di prezzi, la scelta dell'unità produttiva è unica, mentre non possono essere considerate situazioni per cui la ricerca del massimo profitto porta a più di una possibile scelta produttiva, vale a dire a un insieme di scelte produttive.

Koopmans (1951) introduce una definizione formale di *efficienza tecnica* sviluppando e adattando il concetto di efficienza paretiana, che si traduce nell'affermazione per cui in caso di efficienza perfetta in riferimento ai beni finali, nessun bene può essere incrementato senza con ciò provocare una diminuzione o uno spreco di altri output.

Segue la definizione di processo produttivo efficiente come di un processo rispetto al quale qualsiasi incremento in qualunque output richiede una riduzione in almeno un altro output o un incremento in almeno un input, e nel quale una riduzione in qualsiasi input richiede un incremento in almeno un altro input o una riduzione in almeno un output.

Pur costituendo uno strumento per distinguere i processi efficienti da quelli non efficienti, la definizione di Koopmans non consente di sviluppare una soluzione operativa per la determinazione del grado di efficienza di un processo non efficiente, o per l'identificazione di un processo efficiente o di una combinazione di processi efficienti, rispetto alla quale confrontare gli altri processi.

Questa problematica è affrontata da Debreu (1951), il quale si pone il problema di trovare una misura della perdita connessa ad una situazione non ottima, che esprima "la distanza" dalla situazione effettiva del sistema produttivo a quella di ottimo paretiano, introducendo la prima misura di efficienza produttiva, ossia il noto coefficiente di utilizzazione delle risorse (*coefficient of resource utilization*).

Farrell (1957) estende il lavoro iniziato da Koopmans e Debreu evidenziando che l'efficienza produttiva si compone di un altro elemento importante che riflette l'abilità dei produttori di selezionare la giusta combinazione di input per la produzione di una determinata quantità di output alla luce dei prezzi prevalenti sul mercato.

In un tale contesto, Farrell nel suo "*The Measurement of Productive Efficiency*" definisce l'efficienza produttiva complessiva come il prodotto dell'efficienza tecnica e dell'efficienza allocativa, denominata *price efficiency*.

La determinazione di queste misure di efficienza viene effettuata confrontando la *performance* osservata di una certa unità produttiva con uno standard di perfetta efficienza, definito secondo specifici criteri. Va osservato, comunque, che nella definizione di Farrell di efficienza allocativa si ritrova implicitamente il concetto di massimizzazione del profitto o di minimizzazione del costo da parte del produttore in mercati competitivi.

## 1.3 La funzione di produzione, la frontiera delle possibilità produttive e la misura di efficienza

Il modello della funzione di produzione, oltre a rappresentare il primo modello in senso storico ad essere utilizzato dalla teoria economica, è anche quello che consente con maggiore semplicità formale di fornire, almeno in prima istanza, risposte a questioni economiche inerenti il processo produttivo, come ad esempio al fatto se esistano diverse tecniche o modi di produrre gli stessi prodotti e di quali relazioni sussistano fra le proporzioni di impiego dei fattori produttivi utilizzati (Zamagni, 1994).

I primi studi empirici sulla funzione di produzione si basano su serie storiche di dati relativi agli input impiegati e agli output prodotti da ciascuna unità produttiva.

La funzione di produzione Cobb-Douglas (1928) è applicata per la prima volta su dati rilevati in diversi periodi di tempo (serie temporali o *time series*), per verificare empiricamente la teoria della produttività marginale per la distribuzione del prodotto totale.

Poiché con il trascorre del tempo si assiste in genere ad una crescita della popolazione, cui si accompagna un certo sviluppo tecnologico, la critica principale mossa dagli studiosi verso una tale procedura consiste nell'affermare che le eventuali relazioni stimate tra il prodotto e i fattori produttivi, capitale e lavoro, potrebbero essere esclusivamente il risultato dell'agire del tempo.

Bronfenbrenner e Douglas (1939) utilizzano, per la prima volta, dati *cross section* per la stima della funzione di produzione mediante il metodo dei minimi quadrati ordinari (*Ordinary Least Squares, OLS*), ipotizzando che tutte le deviazioni dalla funzione stimata siano dovute ad errori casuali di misurazione della variabile dipendente, o ad operazioni aleatorie di vario genere non incluse nel modello.

Secondo i suddetti Autori, i risultati ottenuti confermano i precedenti studi basati su dati temporali.

Nei lavori empirici concernenti la stima dei parametri della funzione di produzione da dati *cross-section*, la forma deterministica del modello di produzione viene, quindi, "modificata" attraverso l'introduzione di disturbi stocastici, al fine di considerare eventuali errori casuali non sistematici, dovuti, in parte, all'agire dei produttori ed effettuati nel tentativo di adattare gli input per soddisfare le condizioni necessarie per la massimizzazione del profitto.

D'altra parte, l'interpretazione di questi disturbi stocastici, che si assume abbiamo media nulla e varianza costante, non risulta chiaramente delineata in letteratura sino allo studio di Marschak e Andrews (1944), che affrontano il trattamento esplicito degli aspetti probabilistici del modello della funzione di produzione.

I suddetti Autori definiscono il termine stocastico come una componente che riflette *l'efficienza tecnica* dell'unità produttiva, e che dipende dalle conoscenze tecnologiche, dagli sforzi, dai desideri e dalla fortuna di un dato imprenditore.

In una tale situazione, la misura dell'efficienza tecnica di una certa unità produttiva *i* può essere descritta da uno o più parametri, che entrano nell'espressione generale della funzione di produzione specificandola per l'unità *i*-esima.

In altre parole, ciascuna unità produttiva è caratterizzata da una propria funzione di produzione, che differisce da quella delle altre unità per il termine di efficienza, ma che è identica in tutti gli altri aspetti.

Marschak e Andrews spiegano, inoltre, la variabilità dei risultati tra le imprese, riconoscendo l'esistenza di differenze nell'*efficienza economica*, ossia nell'abilità del produttore di scegliere la combinazione degli input più redditizia, in risposta ad eventuali modificazioni dei prezzi sul mercato dei fattori e dei prodotti. Un'ulteriore fonte di variabilità nell'output delle imprese può derivare, in una situazione di non concorrenza perfetta, dalla presenza di differenze nei prezzi pagati o ricevuti dalle diverse unità produttive.

È necessario, a questo punto, evidenziare il problema fondamentale connesso alla stima della funzione di produzione da dati *cross-section* effettuata utilizzando tecniche "classiche" come gli OLS.

La funzione ottenuta da una tale procedura, non rappresenta, in effetti, una "vera e propria" funzione di produzione dal punto di vista della teoria economica, secondo la quale la funzione identifica una frontiera, nel senso che esprime il massimo output ottenibile da una data combinazione di input, massimo che non può essere superato da nessuna unità produttiva.

A tale proposito si instaura un dibattito teorico, a cui partecipano numerosi economisti, concernente la differenza concettuale tra la funzione di produzione, utilizzata nella teoria microeconomica neoclassica, e la funzione sti-

mata da dati *cross-section,* che può definirsi, come si vedrà più avanti, una funzione "media".

Reder (1943) mette in risalto il fatto che solamente un punto sulla funzione di produzione si riferisce ad una situazione reale, ossia alla combinazione di fattori utilizzati e al livello di output prodotto, corrispondente alla situazione in cui l'unità produttiva massimizza il profitto. Tutti gli altri punti sono puramente teorici o ipotetici, ossia esprimono l'output che l'impresa avrebbe ottenuto se avesse trovato redditizio utilizzare altre quantità di fattori. Da qui la diversità concettuale con la funzione di produzione Cobb-Douglas stimata a partire dai dati osservati, che rappresenta combinazioni di fattori e livelli di output scelti da diverse unità produttive, come conseguenza del comportamento massimizzante.

Il suddetto Autore distingue la funzione di produzione teorica per ciascuna unità produttiva, che definisce "*intrafirm",* dalla funzione stimata mediante l'utilizzazione di dati provenienti da processi posti in essere da diverse unità di produzione, denominata "*interfirm"*.

Bronfenbrenner (1944) sottolinea che la determinazione della funzione di produzione "*interfirm"* è logicamente e temporalmente posteriore alla determinazione dell'equilibrio, essendo il luogo dei punti che indicano per ogni impresa le quantità dei fattori utilizzati e il livello di output raggiunto attraverso la massimizzazione del profitto, dati i prezzi dei fattori produttivi sul mercato e l'offerta dei fattori stessi.

La denominazione "*interfirm"* deriva dal fatto che tale funzione costituisce l'unione dei punti di equilibrio per diverse unità produttive e rappresenta uno strumento teorico legittimo, che può essere utilizzato per la verifica empirica della teoria della distribuzione.

È chiaro, quindi, che nei primi studi che riguardano la stima della funzione di produzione si assume che l'output ottenuto dall'impresa possa essere maggiore, o minore, di quello indicato dalla funzione di produzione dell'intera industria, ossia per l'insieme delle unità considerate.

Si ipotizza, in altre parole, che la funzione da stimare sia una sorta di *funzione media* per l'industria, per cui è ragionevole ammettere che alcune unità siano in grado di produrre più della media, altre meno. Sul significato di questa "media" si sono alternate diverse interpretazioni.

Alcuni economisti si riferiscono alla funzione media come alla funzione di una unità produttiva di "dimensioni medie", mentre altri considerano la funzione media come un legame matematico riflettente una sorta di "tecnologia media". Un'ulteriore interpretazione vede la funzione di produzione media come la funzione esprimente l'output sostenibile, ottenuta mediante l'eliminazione di fluttuazioni casuali dovute all'agire di coincidenze più o meno fortunate.

Al di là delle critiche mosse alle diverse interpretazioni del concetto di "media", è necessario sottolineare che la funzione di produzione stimata da dati *cross-section*[3] mediante gli OLS può essere validamente utilizzata solo per determinati scopi. In particolare, la funzione media potrebbe essere impiegata adeguatamente come approssimazione della funzione aggregata, oppure se l'interesse fosse sulla stima di quanto output, in media, può essere ottenuto per una unità nell'industria, con un dato insieme di input.

Per riprodurre i concetti teorici della funzione di produzione, e quindi, per ottenere un'adeguata determinazione della misura di efficienza, il termine di errore del modello econometrico dovrebbe essere caratterizzato da una forma distributiva unilaterale, o quanto meno asimmetrica. Questo tipo di specificazione identifica le cosiddette *funzioni frontiera di produzione*.

A tale riguardo, si può affermare che il lavoro di Farrell (1957) getta le basi per lo sviluppo di un nuovo approccio nello studio dell'efficienza a livello di singola impresa, in quanto focalizza l'attenzione su due elementi fondamentali: come definire il concetto di efficienza e come calcolare la tecnologia di riferimento rispetto alla quale definire misure di efficienza.

Il contributo determinante di Farrell è rappresentato dal riferimento non ad una funzione media, ma ad una *funzione empirica costruita a partire dai risultati migliori osservati nella pratica*, che si possa identificare con il concetto di frontiera di produzione. Attraverso l'ipotesi di rendimenti costanti di scala e di convessità[4], secondo la quale anche combinazione lineari di due vettori di input possono produrre la medesima quantità di output, Farrell identifica la frontiera di produzione come la "*most pessimistic or conservative estimate of the isoquant*".

Per quanto riguarda la stima della frontiera da un punto di vista statistico, il suddetto Autore osserva, dapprima, che esistono alcune funzioni efficienti, rispetto alle quali tutti i punti osservati deviano casualmente ma in un'unica direzione, e si riferisce poi alle analogie esistenti con la stima dei parametri di distribuzioni estreme.

L'isoquanto unitario utilizzato da Farrell può, quindi, considerasi il precursore delle frontiere di produzione, sia nel contesto parametrico, dove riproduce il concetto economico di funzione di produzione, sia in quello non parametrico, dove descrive la superficie dell'insieme delle possibilità produttive.

---

[3] Hoch (1955), riferendosi alla "*average firm*", di cui modella il comportamento economico, stima la funzione di produzione Cobb-Douglas combinando dati *time-series* e *cross section* utilizzando l'analisi della covarianza.

[4] Le proprietà della tecnologia di produzione saranno oggetto di un'analisi formale nel paragrafo 2.3.

Un primo approccio, sviluppato in letteratura per la costruzione delle frontiere di tipo parametrico, rappresenta un'evoluzione del metodo econometrico tradizionale per la stima delle funzioni di produzione ed ha origine da un suggerimento dello stesso Farrell, quando afferma: "*There exists some efficient function, from which all the observed points deviate randomly but in the same direction.*"

Il suggerimento contribuisce dapprima allo sviluppo del cosiddetto **approccio parametrico deterministico** (Aigner e Chu, 1968), che successivamente matura originando il concetto di frontiere stocastiche, attraverso le quali viene introdotto nel modello oltre al termine di errore unilaterale anche un termine bilaterale. Farrell riconosce, infatti, che: "*Errors of observation will introduce an optimistic bias, which can only be eliminated if the distributions of both errors and efficiencies are known. This is an interesting problem for any theoretical statistician.*"

In particolare, la presenza di differenze nell'*efficienza tecnica* di diverse imprese appartenenti ad una medesima industria, intesa come settore produttivo, viene ricondotta da Aigner e Chu alla diversità esistente nelle dimensioni delle imprese e alla conseguente diversa disponibilità e al diverso uso dei fattori produttivi.

Nonostante sia possibile riconoscere l'esistenza di una funzione di produzione dell'industria (*industry production function),* la cui forma è identica per un insieme di unità produttive omogenee, i parametri tecnici della funzione stessa possono variare tra le unità, in conseguenza della presenza di differenze nell'efficienza tecnica. Agli effetti esercitati da fattori puramente casuali, quali condizioni ambientali sfavorevoli e shock esterni, si aggiungono le diversità esistenti tra le imprese nei livelli di *efficienza economica* raggiunti, che determinano il livello di output ottenuto e, quindi, specificano una particolare funzione di produzione per ciascuna unità produttiva.

L'efficienza economica di un'impresa viene ricollegata all'abilità dell'imprenditore nell'effettuare aggiustamenti delle quantità dell'output e degli input, in conseguenza di mutamenti nei prezzi di mercato, al fine di raggiungere la massimizzazione del profitto.

L'agire dei fattori precedenti può determinare il raggiungimento da parte dell'impresa di un livello inferiore di output rispetto a quello espresso dalla funzione di produzione, ma mai di un livello superiore. In quest'ultimo caso, infatti, si sarebbe in presenza di un cambiamento della tecnologia, di cui si dovrebbe tener conto nella valutazione dell'efficienza.

Aigner e Chu, mantenendo una specificazione parametrica della frontiera di produzione, descritta, in particolare, dalla funzione Cobb-Douglas, propongono di utilizzare metodi di programmazione matematica (lineare e quadratica) per la stima della frontiera, in modo tale che i punti osservati giacciano al di sotto della frontiera stessa.

Il metodo introdotto dai suddetti Autori intende caratterizzare le differenze che possono manifestarsi nell'output tra unità produttive che impiegano identici vettori di input, e spiegare perché l'output di una data impresa possa essere inferiore a quello espresso dalla frontiera, attraverso l'introduzione esplicita di un termine di errore che esprime l'efficienza tecnica delle unità produttive.

La specificazione di una particolare forma distributiva del termine di errore, che descrive l'efficienza tecnica, si deve ad Afriat (1972), che ipotizza, in particolare, una distribuzione beta.

L'introduzione di particolari metodi di stima statistici (Førsund, Lovell e Schmidt, 1980) denominati dei minimi quadrati corretti (*Corrected Ordinary Least Squares, COLS)* e dei minimi quadrati modificati *(Modified Ordinary Least Squares, MOLS)* si deve, rispettivamente, a Richmond (1974) e a Greene (1980).

Per superare il limite principale dell'approccio parametrico deterministico, ossia l'estrema sensibilità ai valori anomali (*outliers),* Timmer (1971) propone di stimare la frontiera utilizzando le medesime tecniche di programmazione lineare di Aigner e Chu, ma consentendo ad una proporzione specificata delle osservazioni di posizionarsi al di sopra della frontiera. Anche se l'obiettivo di introdurre nel modello una certa variabilità casuale nei dati si può considerare in parte raggiunto, va osservato che, essendo la specificazione della proporzione essenzialmente arbitraria, il metodo risulta privo di esplicite giustificazioni economiche o statistiche.

Il tentativo di superare l'ulteriore limitazione del metodo della programmazione lineare, ossia l'incapacità di ottenere stime dotate delle usuali proprietà statistiche, si deve a Schmidt (1976) il quale afferma che gli stimatori ottenuti rappresentano, in realtà, stimatori di massima verosimiglianza, nel caso in cui le variabili che descrivono l'efficienza delle unità produttive seguano una distribuzione esponenziale o una distribuzione metà normale. Questa osservazione non si dimostra però sufficiente a migliorare le proprietà statistiche della frontiera di produzione, stimata con il metodo della programmazione lineare, in quanto è facilmente verificabile che le usuali condizioni di regolarità per l'applicazione della massima verosimiglianza non vengono soddisfatte.

Si può, quindi, ragionevolmente affermare che l'approccio econometrico riesce a fornire una stima della funzione di produzione che rappresenti il concetto economico di funzione frontiera e che, al contempo, consenta di ottenere stime dei parametri con le usuali proprietà statistiche, solo in conseguenza dei contributi di Aigner, Lovell e Schmidt (1977) e di Meeusen e van den Broeck (1977).

L'innovazione di tale approccio, che sarà denominato **parametrico stocastico**, consiste nell'ipotizzare che il processo di produzione sia sottoposto a due distinti disturbi casuali, economicamente distinguibili e con differenti caratteristiche.

La prima componente rappresentata il termine di inefficienza *u,* con distribuzione unilaterale, mentre la seconda componente *v*, viene descritta da una distribuzione normale e rappresenta una componente puramente aleatoria.

Per comprendere la portata e l'intensità dell'innovazione generata dal nuovo approccio nel contesto delle frontiere di produzione, è necessario ricordare che nella letteratura economica vi erano stati, in precedenza, diversi tentativi diretti a trasformare il modello di produzione in un modello stocastico, come il già citato lavoro di Marschak e Andrews (1941).

Lo studio che più si "avvicina" al concetto di frontiera di produzione stocastica, è rappresentato dal lavoro di Zellner, Kmenta e Drezé (1966), nel quale alla funzione di produzione si aggiunge un termine stocastico descrivente l'agire di fattori legati ad eventi puramente aleatori, rappresentati, ad esempio, da variazioni imprevedibili nelle *performance* dei macchinari utilizzati o da eventi meteorologici.

Le più recenti versioni della teoria neoclassica non ricorrono alla funzione di produzione, e quindi allo strumento del calcolo differenziale, ma alla teoria degli insiemi. Ciò non modifica, comunque, la concezione della teoria della produzione come allocazione delle risorse a mezzo di decisioni decentrate, dirette alla massimizzazione dell'utile netto, a un dato sistema di prezzi (Koopmans, 1951).

L'esistenza di diverse alternative di produzione, ossia di processi diversi che permettono di ottenere lo stesso prodotto, ovvero di utilizzazioni diverse degli stessi input, viene rappresentata mediante l'*insieme di produzione* $\widetilde{Z}$ , (o $Z$) costituito da tutti i vettori **z** (o da tutte le coppie di vettori (**x**, **y**)), che rappresentano, secondo la convenzione prescelta, processi di produzione possibili in una data situazione.

Le proprietà formali che caratterizzano l'insieme $Z$ rispetto ai legami esistenti tra i diversi processi possibili, sono dirette alla semplificazione dell'analisi delle scelte di processi ottimali e alla rappresentazione analitica di caratteristiche che possono trovarsi nelle alternative tecnologiche reali (Tani, 1986).

L'insieme delle produzioni possibili, che descrive le caratteristiche della tecnologia, ossia le relazioni fra input e output, prende il posto della funzione di produzione.

Nei casi di produzione singola questa funzione può essere considerata come la rappresentazione della frontiera efficiente dell'insieme $Z$ (Koopmans, 1951).

Le imprese devono scegliere una delle attività comprese in $Z$, determinando le quantità prodotte e la domanda di fattori, e fissando, in tal modo, i rapporti relativi tra le quantità impiegate dei vari input e il metodo di produzione.

Nella scelta del processo di produzione da attivare, le unità produttive sono guidate sempre dal criterio della massimizzazione del profitto, il cui raggiungimento è garantito dalla condizione di convessità di Z, attraverso la quale si assicura che, per ogni vettore di prezzi, si trovi la giusta combinazione di fattori e di prodotti.

In un tale contesto, un processo di produzione appartenente all'insieme $Z$ è detto tecnicamente efficiente se è massimale in Z rispetto alla relazione di ordine $\leq$ , definita tra i vettori, per cui non è possibile ridurre la quantità di alcun input senza aumentare la quantità di un altro input e/o ridurre la quantità di uno o più output; né è possibile aumentare la quantità di qualche output senza ridurre la quantità di qualche altro output o aumentare la quantità di qualche input (Tani, 1986).

Tale definizione si inserisce chiaramente nell'ottica paretiana, ripresa successivamente da Koopmans. Anche in questa impostazione, al concetto di efficienza tecnica, collegato al buon uso degli input disponibili a livello di singola unità produttiva, si accompagna il concetto di efficienza economica, connesso ad una opportuna allocazione delle risorse disponibili tra le diverse produzioni, il cui raggiungimento viene garantito dall'esistenza di un sistema di prezzi che agisce a livello aggregato.

Per lo sviluppo di metodologie che definiranno l'**approccio non parametrico** all'analisi dell'efficienza è di nuovo fondamentale il lavoro di Farrell (1957), che estende il concetto utilizzato da Pareto e Koopmans in riferimento all'intero sistema economico, agli input così come agli output di qualsiasi organizzazione produttiva ed, esplicitamente, evita qualunque uso dei prezzi e dei relativi meccanismi di scambio.

Ancora più sostanziale appare la decisione di utilizzare le *performance* delle altre unità produttive per valutare il comportamento di ogni unità in relazione all'output prodotto e agli input utilizzati.

In tal modo è possibile procedere empiricamente nel determinare l'efficienza relativa delle unità di produzione sviluppando il concetto di frontiera sia da un punto di vista econometrico, come già visto, che da un punto di vista non parametrico.

Ispirato da Debreu (1951), Farrell introduce, a livello "micro", misure di efficienza basate sulla contrazione radiale che collega i punti non efficienti osservati per le unità di produzione con i punti di riferimento (non osservati) sulla frontiera di produzione, stimata attraverso l'impiego di sistemi di equazioni lineari.

Farrell e Fieldhouse (1962), nel tentativo di generalizzare il metodo precedente al caso di rendimenti crescenti di scala, suggeriscono l'utilizzazione di problemi di programmazione lineare. Boles (1967) fornisce, successivamente, un'indicazione per la formulazione dei problemi di programmazione lineare al caso di output multipli.

Ma è il lavoro di Charnes, Cooper e Rhodes (1978) che, ispirandosi agli studi di Farrell, contribuisce in modo fondamentale allo sviluppo di strumenti

non parametrici per la determinazione della frontiera di produzione, attraverso l'introduzione di una metodologia che sarà successivamente denominata *Data Envelopment Analysis* (DEA).

I suddetti Autori introducono, per ciascun'unità produttiva, una misura di efficienza definendola come il massimo del rapporto tra una media ponderata degli output e una media ponderata degli input, che caratterizzano il processo dell'unità analizzata, con il vincolo che gli altri rapporti definibili per le restanti unità produttive siano minori o uguali ad uno. La determinazione dell'efficienza relativa di ciascuna unità viene anche formulata mediante un ordinario problema di programmazione lineare.

Nel fornire un'interpretazione della funzione di produzione che deriva dall'applicazione del metodo proposto, Charnes, Cooper e Rhodes affermano di aver introdotto un nuovo tipo di funzione di produzione che vagamente si ricollega al concetto di impresa rappresentativa, utilizzato tempo prima da Marshall.

Quest'ultimo Autore proponeva di caratterizzare le possibilità produttive attraverso la media degli output ottenuti da dati input e di definire l'utilizzatore di questa relazione media come l'unità rappresentativa. Eventuali variazioni tra i costi sostenuti dalle imprese per l'ottenimento di una data quantità di output erano riconducibili all'età dell'impresa e ad eventuali differenze esistenti nelle capacità imprenditoriali.

Nell'approccio suggerito da Charnes, Cooper e Rhodes, l'ottimo rispetto al quale determinare l'efficienza di ciascuna unità produttiva è costituito, piuttosto che da imprese rappresentative "medie", da imprese rappresentative efficienti.

Questo insieme di metodologie non parametriche presenta un carattere puramente deterministico, perciò, in genere, i disturbi casuali sono inclusi nella misura di efficienza. D'altra parte, nell'approccio econometrico, dove si impone una struttura parametrica sia alla tecnologia, descritta attraverso una funzione di produzione, sia alla distribuzione del termine di efficienza, si può verificare una commistione tra errori di specificazione ed efficienza (Ferrier e Lovell, 1990).

Se da una parte, quindi, l'approccio non parametrico, assicura una maggiore flessibilità nella descrizione della tecnologia e nella misurazione dell'efficienza tecnica, dall'altra, il non tener conto di una certa variabilità stocastica, determina un'eccessiva dipendenza dei risultati dai dati osservati, che si traduce in una forte sensibilità ai valori anomali, come nel caso dell'approccio deterministico, sebbene recenti sviluppi suggeriscano l'introduzione di un approccio stocastico alla DEA (Simar e Wilson, 2000, Cazals, Florens e Simar, 2002).