A list of Working Papers on the last pages

No. 241, 1989

ON THE ECONOMETRIC ANALYSIS OF PRODUCTION WHEN THERE ARE NO OUTPUT DATA

by

Erik Mellander and Bengt-Christer Ysander

Paper prepared for IUI's 50th Anniversary Symposium, November 15–17, 1989

December, 1989

ON THE ECONOMETRIC ANALYSIS OF PRODUCTION WHEN THERE ARE NO OUTPUT DATA *

Erik MELLANDER
The Industrial Institute for Economic and Social Research (IUI), S – 114 85 Stockholm, Sweden

Bengt-Christer YSANDER
University of Uppsala, S – 751 20 Uppsala, Sweden

October 1989

Abstract: A general method is described which allows a production activity to be analyzed by means of input data only. According to duality theory, the input cost shares can be completely specified without any information about output if the technology is homothetic. It is demonstrated that these cost shares can yield information about elasticities of substitution and factor demand and on productivity development. Moreover, the system of share equations can be generalized to allow for technical and allocative inefficiency and the effects of these inefficiencies on total costs and input demands can be estimated.

* Financial support from the Bank of Sweden Tercentenary Foundation is gratefully acknowledged. We have benefited from helpful comments and suggestions from Ernst Berndt. Eilev Jansen's comments at the 1989 European Meeting of the Econometric Society were most useful. We also wish to thank Rolf Färe for commenting on an earlier version of this paper, and seminar participants at IUI, Uppsala University, and Lund University for useful discussions.

MAILING ADDRESS: Erik Mellander
Industriens Utredningsinstitut
Box 5501
S – 114 85 STOCKHOLM
Sweden

1. INTRODUCTION

For many productive activities it is very difficult to define a relevant output measure – and often practically impossible to implement it, once defined. In particular, this is the case regarding the rapidly expanding service sector. Most services have several quality and quantity dimensions, some of which are largely unobservable. For instance, health care not only results in actual changes of patients' health status. It also helps prevent future, potential illnesses. Obviously, to quantify the latter effect is an almost hopeless task.

The most severe output measurement problems are probably encountered in the public sector. In the national accounts system, this has lead to the convention that the value of a public service is set equal to the value of the resources used to produce it. Volume measures are obtained by weighing the inputs by constant, rather than current, prices. Accordingly, the volume of a particular service, q say, in year t is defined as

$$q_t \equiv x_{1t} + x_{2t} + \cdots + x_{nt}$$

where x_i is the amount used of input i, valued at constant prices.[1] This accounting practice implies several strong assumptions about the productive performance of the public sector, some of which do not seem to have been generally recognized.

First, it implies that the growth in total factor productivity, defined as the difference between the growth in real output and the growth of the cost share weighted inputs according to, e.g., Jorgenson and Griliches (1967) will always be equal to zero. Apart from almost certainly yielding an incorrect measure of the productivity development in the public sector itself, this will also bias calculations of aggregate growth (e.g. GDP), as soon as the size of the public sector changes.

Secondly, it means that the production in the public sector is assumed to be

[1] In some countries, e.g. the U.S., only labor input is considered.

efficient in the sense that there is no slack in the utilization of the various factors of production – if such slack were to exist it would be possible to reduce the usage of some of the inputs without reducing output but, according to the chosen method of measurement, any such reduction would decrease the level of output. This is in contrast not only with widely held beliefs but also with theoretical considerations predicting lower efficiency in the public than in the private sector.

Thirdly, it can readily be seen that a proportionate increase in all inputs will increase q by the same proportion, implying that constant returns to scale are assumed. In view of the fact that diminishing average costs, i.e. increasing returns, is an important motivation for public production, this is somewhat unfortunate.

Finally, the additive formulation amounts to assuming that all inputs are perfect substitutes. However, given this technological property a cost-minimizing producer would of course only use the cheapest input. Hence, it must be implicitly assumed that public producers ignore the effects of relative prices on total costs. Although this seems to agree quite well with common opinion it would be preferable to regard it as an hypothesis to be tested rather than as a *maintained* hypothesis.

In this theoretical paper we show that given (time series) input data, three of these four issues, namely productivity growth, efficiency in production, and sensitivity to changes in relative input prices, are amenable to econometric analysis *even in the absence* of output measures. There is thus no need to arbitrarily determine them *a priori*. Moreover, although we cannot explicitly study properties concerning returns to scale, our approach allows for the possibility of variable returns.

Our results are completely general in the sense that they can be applied to *any* production activity, i.e. not only to those in the public sector. Within the private sector, the production of banking services is an example of an interesting object of study. In the national accounts, value measures of the output in the banking sector are obtained by adding the bank's service charges and the net proceeds from their lending operations. To construct volume measures of output, various *ad hoc*

assumptions are made. In the Swedish national accounts, e.g., it is assumed that the banking industry every year experiences a 2% increase in average labor productivity.[2] Similar procedures are employed in other countries, too. Our approach makes it possible to investigate the empirical validity of such assumptions.

In contrast to the method that we are going to propose, analyses of production activities for which there are no reliable output measures traditionally have employed proxy variables, intended to mirror the unknown output. Attempts have also been made to take several dimensions of output into account simultaneously, either by aggregating several proxies into an output indicator index or by modeling production as multiple-output processes.[3] Still, studies of this kind can always be criticized for failing to account for such basically unobservable output dimensions as the one exemplified in the first paragraph above. Since, by their very nature, such characteristics cannot be explicitly incorporated into the analysis the only way to escape this criticism is to find some method of avoiding the measurement of output altogether, as we do in this paper. To our knowledge, the only previous attempt in this vein is Hulten's (1984) study of productivity changes in the public sector.

Inspired by household production theory, Hulten models the whole economy as a "household", maximizing a utility function in an aggregate private sector good, directly available for consumption, and an aggregate public sector commodity, which is produced by the community for internal consumption.[4] The production process yielding the public sector commodity is assumed to exhibit constant returns to scale. Productivity changes are further presumed to be Hicks-neutral and are

---

[2] As far as we know, no empirical support exists for the particular choice of 2%. It is interesting to note that in an attempt to measure average labor productivity in American banks over the period 1927-1979, Rhoades and White (1984) could not find any indication of growth in average labor productivity since the mid 1950's.

[3] Concerning examples from the public sector and the banking industry, an extensive bibliography is available from the authors on request.

[4] Household production theory is of interest in this context as it makes it possible to analyze a household's internal production of non-market commodities, although the commodities produced generally are unmeasureable, just like the output of the public sector.

modeled by means of an exponential time trend. Duality theory can then be used to express the price of the public sector commodity in terms of the prices of the factors of production and a time index. Hulten shows that this result in turn makes the ratio of the "household" budget shares for the private and public sector outputs a function of the price of the private sector good, the factor prices and the time index. By means of this equation the rate of public sector productivity growth can be estimated without an explicit measure of the public sector output.

In addition to the rather restrictive assumptions about the production technology a serious problem with Hulten's approach is the maintained, and therefore untestable, hypothesis that the household/community analogue is indeed valid, which is far from obvious. Our method is based only on standard neoclassical production theory and, hence, can be applied to the public sector without any such assumption. Moreover, in contrast to Hulten, we do not have to presume the availability of any other information than input data for the particular production process studied.

Given only input data, production or profit functions are infeasible as instruments of analysis, since in studies based on these the level of output is endogenously determined. This leaves a cost function analysis, in which the output level is treated as predetermined, as the only possibly practical alternative.[5]

Output predeterminacy alone will not make it possible to analyze the production process by means of input data only. Both the cost function and the input demands which can be derived from it will always be dependent upon the level of output. However, if the production technology is homothetic, i.e if the proportions in which the factors of production are employed are unaffected by the scale of operation, then the shares of the various inputs in total cost will be independent of the output level.

---

[5] The treatment of output as a predetermined variable does not necessarily imply that the output decision is taken by someone else than the producer. It can be justified even if the output level is set by the producer himself, provided that the problem of minimizing unit costs can be separated from the problem of choosing the level of output so as to maximize profits (or, in the context of public production, some net benefits or welfare criterion). In fact, this independence condition is fulfilled by the homothetic technologies that we will consider in this paper.

The input cost shares will thus be the endogenous variables in our analysis.

The property that the input cost shares of a homothetic technology are invariant to the level of output has long been recognized in the econometrics literature. The extent to which these cost shares can yield information about the production process has not been thoroughly investigated, however. We perform such an investigation, based on a homothetic cost function formulated in general terms.

The paper unfolds as follows. In Section 2, some rather well known implications of homotheticity are briefly stated, e.g. that it allows price and substitution elasticities to be estimated by means of input data only. Estimation of the effects of non-neutral technical change on input requirements, total costs, and on total factor productivity is taken up in Section 3. In Section 4 we consider the fact that the theoretical derivation of the input cost shares assumes that production costs are minimized. We demonstrate how the dual representation can be generalized to allow for the existence of technical inefficiency (overutilization of inputs) and allocative inefficiency (inoptimal factor proportions), implying higher than minimum costs. Moreover, we show how that the thus generalized system of cost shares can be used to estimate the increases in total costs brought about by the inefficiencies, as well as their effects on input utilization. As far as we know, the fact that this is possible even when there are no data on output has not been demonstrated earlier. Formulas for comparing price and substitution elasticities, and the estimated effects of technical change, when there are inefficiencies in production with the corresponding measures under cost minimization are also given. Section 5 contains a brief summary of our findings.

2. SOME IMPLICATIONS OF HOMOTHETICITY

From now on, we assume that information is available about the quantities used

of the different factors of production and their respective prices, but that there are no data on output. To simplify the discussion, we will in this section disregard technical change and the possibility of inefficiencies in production. We will come back to these issues in Sections 3 and 4. For the time being we thus assume a static technology and cost-minimizing producers.

Let the minimum cost function be $C = C(y,p)$, where y is the unknown output ($y > 0$) and p denotes the vector of (strictly positive) input prices, $p = (p_1,...,p_n)$.[6] The cost function must fulfill certain regularity conditions which, e.g., can be formulated as in Diewert (1971, pp. 489-90). To be regular, $C(y,p)$ should be non-decreasing in both y and p, and be linearly homogeneous and concave in p. If these conditions are all satisfied, the cost function will describe all economically relevant aspects of the production technology. In addition, it will be assumed here that $C(y,p)$ is twice differentiable with respect to each of its arguments.

The producer's input demands can be derived by means of *Shephard's lemma*, according to which:

$$(1) \qquad x_i = x_i(y,p) \equiv \frac{\partial C(y,p)}{\partial p_i}, \qquad i = 1,...,n.$$

Since the cost function is linearly homogeneous in p, it further holds – by *Euler's theorem* – that:

$$(2) \qquad \sum_{k=1}^{n} p_k \cdot x_k(y,p) = \sum_{k=1}^{n} p_k \cdot \frac{\partial C(y,p)}{\partial p_k} = C(y,p) .$$

In accordance with (1) and (2), the input cost shares can be written

---

[6] As pointed out to us by Rolf Färe, the results in the following are valid not only for single output technologies, but for multiple output technologies as well. In principle, the scalar y can thus be replaced by an m vector $y = (y_1,...,y_m)$ of outputs.

$$(3) \qquad S_i \equiv \frac{p_i \cdot x_i(y,p)}{C(y,p)}, \qquad i = 1,\ldots,n .$$

In contrast to the input demands, the cost shares may be independent of the level of output, y. This will be the case if the cost function is multiplicatively separable in y and p. Shephard (1953, pp. 45-47) has shown that the cost function has this property if, and only if, the production technology is *homothetic*, implying that

$$(4) \qquad C(y,p) = f(y)\cdot g(p) ,$$

where f is a monotonically increasing function of y. By means of restrictions on $f(y)$ the homothetic technology can be specialized into a homogeneous technology. In particular, linear homogeneity, i.e. constant returns to scale, requires that f be equal to the identity function. It can be shown that given an appropriate definition of $f(y)$, the function $g(p)$ equals the cost of producing one unit of output, i.e. $C(1,p)$.

It is the function $g(p)$ that we will be interested in. However, separate identification of $f(y)$ and $g(p)$ requires some kind of normalizing restriction – without such a restriction (4) can always be alternatively expressed as $C(y,p) = \bar{f}(y)\cdot\bar{g}(p)$ where $\bar{f}(y) = k\cdot f(y)$, $\bar{g}(p) = k^{-1}\cdot g(p)$, and k is an arbitrarily chosen constant.[7]

Given (4) the system of input cost shares becomes

$$(5) \qquad S_i = S_i(p) = \frac{p_i \cdot \frac{\partial g(p)}{\partial p_i}}{g(p)}, \qquad i = 1,\ldots,n.$$

---

[7] In applied work, where the elements of the vector p are often price indices rather than (absolute) price levels, it is convenient to impose the normalizing restriction that $g(1) = 1$. This condition is most easily interpreted in the context of a producer employing a constant returns technology in a competitive environment. Under such circumstances $g(p)$ will be equal to the price of output and the constraint $g(1) = 1$ merely ensures that the base-year for the output price index will be equal to that of the input price indices. The constraint also has a natural interpretation in a more general context; it will then have the effect of constructing a unit cost index which can be consistently compared with the input price indices.

Specification of an explicit functional form for C, which has the properties (4) and (2), thus makes it possible to estimate the system (5), and hence the function $g(p)$, without having to take the level of output into account. This obviously solves our main problem, i.e. that of eliminating the unknown entity y from the analysis.

It is clear, however, that the system (5) cannot provide a full description of the production technology as it does not yield complete information about the function $f(y)$.[8] In contrast, the system (1) of input demands contains all the information available in the original cost function, since the input demands multiplied by the factor prices add up to $C(y,p)$, as shown in (2). This difference in informational content between the two systems is explained by the fact that whereas the system (1) is of full rank (i.e. n) the rank of the system (5) of cost shares is only n-1, which can easily be seen by noting that both sides of (5) sum identically to one. As a consequence, one of the share equations will be dropped in the actual estimation.[9] This, in turn, implies that if the functional form chosen for C is *flexible* symmetry has to be imposed *a priori* to ascertain identification of the function $g(p)$.

The information loss incurred by studying the system of input cost shares only concerns the scaling properties of the technology, however. Factor substitution and the price reponsiveness of input demands can still be analyzed. Using the results of Uzawa (1962), the Allen partial elasticities of substitution [Allen (1959)] can be expressed in terms of the input cost shares and the factor prices according to

$$(6) \qquad \sigma_{ij} \equiv \frac{C \cdot \frac{\partial^2 C}{\partial p_i \partial p_j}}{\frac{\partial C}{\partial p_i}\frac{\partial C}{\partial p_j}} = \left[S_i S_j + p_j \cdot \frac{\partial S_i}{\partial p_j}\right] \cdot \frac{1}{S_i S_j},$$

---

[8] Given an estimate $g^*(p)$ of $g(p)$ an estimate of $f(y)$ can be obtained by means of the ratio $p'x/g^*(p)$ where $p'x$ is observed total cost. The form of the function f and the value of y cannot be inferred, however, except in the special case when there are constant returns to scale, implying that $f(y) = y$.

[9] If the estimation method is that of maximum likelihood and the stochastic disturbance terms are additively appended to the share equations (5) the estimation results will be invariant to the choice of the left out equation, cf. Barten (1969).

while the price elasticities can be calculated as

$$\eta_{ij} \equiv \frac{\partial x_i}{\partial p_j}\frac{p_j}{x_i} = S_j \sigma_{ij} \,. \tag{7}$$

The homotheticity assumption can of course be questioned. In a static environment – i.e. in the absence of technical change – it implies that the cost-minimizing input mix is determined by relative input prices only, which is often a restrictive assumption. As a consequence, with constant relative prices the expansion path, describing the optimal factor proportions at succesively higher output levels, will be linear [cf. Färe (1974)]. This may not be consistent with the often noted tendency to increase the capital intensity at larger scales of operation.[10]

Homotheticity has also been decisively rejected in many studies of, e.g., the manufacturing sector. However, it is easier to defend this assumption in the context of service production than in the production of goods. The reason is, of course, that services are more difficult to routinize, making the scope for automatization more limited. Although this argument should be used with caution the homotheticity assumption appears to be most applicable where it is most needed, i.e. in service production where no reliable output measures are available. In the case of government services, homotheticity may, moreover, reflect centralized decision making which tends to treat establishments of different size – e.g. schools – all alike.[11]

---

[10] Changing the scale of operation generally takes some time, during which relative input prices may change, too. Thus, an observed increase in the capital intensity during an output expansion need not necessarily be inconsistent with homotheticity.

[11] It should be noted that there is no obvious conflict between centralized decision making and cost minimization. If the central decisions take the form of requirements on the input mix, conditioned upon a given set of factor prices, they may have precisely the effect of imposing a homotheticity constraint on the production possibilities facing the local producers. As long as the central decrees are optimally adjusted to changes in the relative input prices, costs will be minimized, albeit subject to a homotheticity restriction.

3. TECHNICAL CHANGE AND TOTAL FACTOR PRODUCTIVITY

In accordance with common practice, we assume that technical change (of a disembodied nature) can be modeled by means of a time index, t.[12] The cost function can then be formulated according to

$$C = C(y,p,t) = f(y)\cdot g(p,t)\,, \tag{8}$$

i.e. the price function is augmented to include t as an additional argument.[13] [14] The system of equations to be estimated, i.e. the cost shares corresponding to (8), is now

$$S_i = S_i(p,t) = \frac{p_i \cdot \frac{\partial g(p,t)}{\partial p_i}}{g(p,t)}\,, \qquad i = 1,\ldots,n\,. \tag{9}$$

We begin by considering the effect of technical change on the input requirements and the cost shares. We will then use the connection between technical change and total factor productivity to investigate what conclusions can be drawn about the rate of total factor productivity growth.

3.1. *Effects on input demands and input cost shares.* By including the time

---

[12] For examples, see, e.g., Parks (1971), Binswanger (1974), Berndt and Khaled (1979), and Nadiri and Schankerman (1980).

[13] In principle, it is conceivable that technical change might also affect the scaling properties of the technology, in which case one would think that the function $f(\cdot)$ should be dependent on t, too. However, in the context of a homothetic technology effects of technical change on returns to scale are equivalent to (Hicks-) neutral technical change. Since neutral technical change can be – and will be – considered within the framework of (8) there is thus no need to include t as an argument in the scaling function $f(\cdot)$.

[14] In the context of service production, demographic and/or socio-economic variables are sometimes included as arguments in the cost function, too; see, e.g., Hulten (1984) and Schwab and Zampelli (1987). The inclusion of such variables will not be discussed in this paper, however, since from a methodological point of view it is analogous to modeling technical change.

index, input demands are allowed to shift over time not only in response to changes in relative factor prices but also because of exogenously determined technological developments. These developments affect the input requirements over time and, hence, also the input cost shares. In the following, we will use the letter τ to denote a relative time derivative. Accordingly, the rate of change in the usage of factor i resulting from technical change can be written

$$\tau_{x_i} \equiv \frac{\partial x_i(y,p,t)}{\partial t} \frac{1}{x_i(y,p,t)}, \qquad i = 1,\ldots,n .$$

Since in our case

$$x_i(y,p,t) = f(y) \cdot \frac{\partial g(p,t)}{\partial p_i}, \qquad i = 1,\ldots,n , \tag{10}$$

the rate of change in the demand for input x_i can be expressed in terms of only the input prices and the time index according to

$$\tau_{x_i} = \frac{\partial^2 g(p,t)}{\partial p_i \partial t} \cdot \left[\frac{\partial g(p,t)}{\partial p_i}\right]^{-1}, \qquad i = 1,\ldots,n. \tag{11}$$

Further, it can be shown that the relative effects of technical change on the cost shares – i.e. the Binswanger (1974) measures of the biases in technical change – can be expressed in terms of the effects on the input demands, in the following way

$$\tau_{S_i} \equiv \frac{\partial S_i(p,t)}{\partial t} \frac{1}{S_i(p,t)} = \tau_{x_i} - \sum_{k=1}^{n} S_k \tau_{x_k}, \quad i = 1,\ldots,n. \tag{12}$$

The technically induced rate of change in the i'th cost share will thus be equal to the difference between the rate of change in the demand for the i'th input and

the corresponding cost-weighted average rate of change in demand, taken over all n inputs. This implies that in order to determine the τ_{S_i}'s we must be able to estimate the effects of technical change on each of the n inputs. However, due to the linear dependence among the input cost share equations, the system (9) can provide at most n-1 independent estimates of the input effects, i.e. τ_{x_i}'s. Fortunately, either the condition that the cost function be linearly homogeneous in the input prices, or the normalizing restriction required to separate $g(p,t)$ from $f(y)$ can be used to impose one restriction on the τ_{x_i}'s.[15] Hence, the maximum number of τ_{x_i}'s to be estimated coincides with the maximum number that the system of cost shares is capable of generating.

If $\tau_{S_i} < 0$ technical change is characterized as relatively factor i-saving and if $\tau_{S_i} > 0$ it is said to be relatively factor i-using. We define technical change to be *non-neutral* when it is either relatively factor i-saving or relatively factor i-using for at least one $i = 1,...,n$, thereby indicating that it effects the relative development of the input cost shares over time and, hence, also the factor proportions x_i/x_j, $i \neq j$. If, instead, $\tau_{x_i} = \tau_x \neq 0$ for all i, so that $\tau_{S_i} = 0$ for all i, then technical change is defined as *neutral*.[16] This can only happen if the

---

[15] Which of these two alternatives that will apply depends on the specific functional form chosen for the cost function. For instance, if the *translog* cost function of Christensen, Jorgenson, and Lau (1973) is used, $g(p,t)$ can be specified according to

$$g(p,t) = g(p) \cdot \exp[\Sigma_k \gamma_k(t \cdot \ln p_k)]$$

where the γ_k's are unknown parameters. As $g(p)$ is linearly homogeneous in p the sum $(\gamma_1 + \cdots + \gamma_n)$ must equal zero to ascertain that $g(p,t)$, too, is linearly homogeneous in p. If, instead, the technology is the *Generalized Leontief* suggested by Diewert (1971) technical change can e.g. be modeled as

$$g(p,t) = g(p) + \Sigma_k \gamma_k(t \cdot p_k)$$

in which case the normalizing restriction $g(1,t) = g(1) = 1$ (cf. footnote 7) can be used to impose the same constraint, i.e. that the γ_k's should sum to zero. In both cases there will be only n-1 independent τ_{xi}'s to estimate.

[16] Strictly, technical change is defined as neutral if the marginal rates of technical substitution between each pair of inputs are not affected by it [Hicks (1932)]. Blackorby, Lovell, and Thursby (1976) have demonstrated, however, that if the technology is homothetic Hick's definition is equivalent to the one given here.

function $g(p,t)$ is weakly separable in p and t, i.e. if $g(\cdot)$ can be expressed in terms of two other functions, h and ϕ say, according to

$$g(p,t) = h[\phi(p),t] \,. \tag{13}$$

Using the linear homogeneity of $g(p,t)$ in p, one can easily verify that given (13) the system (9) degenerates to the system (5), implying that the cost shares are invariant to neutral technical change.[17]

Accordingly, to capture any effects of technical change it is necessary to specify technical change as being non-neutral. This is no drawback as far as modeling is concerned; neutral technical change is probably a very rare phenomenon. Moreover, in the present context there is also a theoretical argument for disregarding this particular form of technical change: As shown by Sato (1980), the effects of neutral technical change and the effects of returns to scale are not independently identifiable when the technology is homothetic. Thus, even if we had had an output measure we would not have been able to separate these two effects.[18]

3.2. *Total factor productivity.* The effects of technical change on the rate of total

---

[17] This proves that (13) is a sufficient condition for technical change to be neutral. To prove necessity, notice that neutrality requires the right hand side of (11) to be equal for all i. Since we know that, in general, $[\partial g(p,t)/\partial p_i] \neq [\partial g(p,t)/\partial p_j]$ this implies that the function $g(p,t)$ must have the property that

$$\frac{\partial^2 g(p,t)}{\partial p_i \partial t} = \frac{\partial g(p,t)}{\partial p_i} \cdot \xi(p,t) \,, \qquad i = 1,\ldots,n,$$

where ξ is a (non-zero) function of p and t. To have this property the function $g(p,t)$ must, however, be weakly separable in p and t. □

[18] Sato's result has an important implication with respect to the interpretation of a statistical comparison of the systems (9) and (5) – based, e.g., on a likelihood ratio test. Obviously, such a comparison could always be formulated as a test of the null hypothesis H_o : "There has been no technical change *or* technical change has been (Hicks–)neutral." against the alternative H_a : "Technical change has been non-neutral". However, Sato's conclusion makes it possible to strengthen and simplify the null hypothesis to $\mathcal{H}_o$: "There has not been (any kind of) technical change." and to reformulate the alternative to $\mathcal{H}_a$: "Technical change has occurred." since from an operational point of view (H_o, H_a) and $(\mathcal{H}_o, \mathcal{H}_a)$ are equivalent.

factor productivity can be analyzed by means of a general duality result derived by Ohta (1974)[19]. Let $\psi(x,t)$ denote the production function to which the cost function (8) is dual. The *primal* rate of total factor productivity can then be defined according to

$$\tau_{\psi} \equiv \frac{\partial \psi(x,t)}{\partial t} \frac{1}{\psi(x,t)} .$$

What Ohta has shown is that the following dual relationship holds

$$\tau_{\psi} = (-\tau_C) \cdot (e_{C_y})^{-1} \tag{14}$$

where

$$\tau_C \equiv \frac{\partial C(y,p,t)}{\partial t} \frac{1}{C(y,p,t)} \tag{15}$$

and

$$e_{C_y} \equiv \frac{\partial C(y,p,t)}{\partial y} \frac{y}{C(y,p,t)} . \tag{16}$$

The first factor in (14), the negative of the rate of total cost diminution, is the dual representation of technical change. The second factor, the inverse of the elasticity of total cost with respect to output, is the dual form of the rate of return to scale. Returns to scale are increasing if $e_{C_y} < 1$, constant if $e_{C_y} = 1$, and decreasing if $e_{C_y} > 1$. It can be shown that for a homothetic technology e_{C_y} will always be strictly positive, see e.g. Førsund (1975), a property which will prove useful in the following.

In the present context (14) becomes[20]

---

[19] Ohta's result is not limited to homothetic technologies but can be applied to non-homothetic technologies as well, see e.g. Berndt and Khaled (1979). Our presentation in the following is closely related to the one given by Berndt and Khaled.

[20] Since it can be shown that

$$\tau_C = \frac{\partial g(p,t)}{\partial t} \frac{1}{g(p,t)} = \Sigma_{k=1}^{n} S_k \tau_{x_k}$$

$$(17) \qquad \tau_{\psi} = \left[-\frac{\partial g(p,t)}{\partial t}\frac{1}{g(p,t)}\right] \times \left[\frac{y \cdot f'(y)}{f(y)}\right]^{-1} .$$

Because of the occurrence of y in the last factor of (19), it is obvious that, in general, the system (9) of input cost shares does not provide all the information needed to calculate an estimate of the rate of total factor productivity. However, since we know that e_{C_y} will be strictly positive the sign of (17) will be equal to the sign of the first factor on the right hand side, i.e. the dual rate of technical change. Accordingly, the question of whether total factor productivity is increasing or decreasing can always be answered by means of the first factor in (17), which can be obtained from the estimation of the system of cost shares.

If, further, the technology is homogeneous then the rate of return to scale will be independent of the level of y and so the last factor in (17) will be equal to a constant, instead of being a function of y. In that case

$$\frac{\tau_{\psi}^{t_2}}{\tau_{\psi}^{t_1}} = \frac{\tau_C^{t_2}}{\tau_C^{t_1}}$$

i.e. the relations between the rates of total factor productivity at two points in time, t_1 and t_2, will be equal to the corresponding relation between the rates of cost diminution, making it possible to construct an index of total factor productivity growth. In other words, τ_{ψ} is determined up to a constant of proportionality. If, finally, the technology is linearly homogeneous, i.e. characterized by constant returns to scale, then e_{C_y} will be equal to unity and the negative of τ_C will be identical with the rate of change in total factor productivity.

The conclusions that can be drawn about productivity growth when only input

---

the computation of the dual rate of technical change, i.e. the first factor in (17), does not require any extra effort; the sum on the right hand side of the last equality will be necessary in the calculations of the Binswanger measures (12) of the biases in technical change, too.

data are available will thus depend on how restrictive assumptions we are willing to make concerning returns to scale. Thus, while the homotheticity assumption always allows us to determine the sign of the productivity growth rate we need to assume constant returns to scale to be able to obtain a complete characterization of the growth in total factor productivity.

4. DEVIATIONS FROM COST MINIMIZATION

By definition, $C(y,p,t)$ denotes the smallest total cost attainable in time period t for input vectors yielding at least the output y. If costs are not minimized, estimation of the system (9) may yield biased estimates of the price and substitution elasticities (6) and (7), and of the effects of technical change on the production process. These considerations are of particular importance concerning public sector applications, as there are theoretical arguments for questioning cost minimization as the primary objective of public producers.[21]

Deviations from minimum costs – which, of course, must always be positive – are commonly taken to arise because of inefficient producer behavior, but there may be other reasons as well.[22] We will not try to discriminate between different sources of inefficiency, however. Following the literature in this field, we will be content with

---

[21] For a summary of these arguments, see Byrnes, Grosskopf, and Hayes (1986). It is interesting to note, however, that the conclusion consistently reached in theoretical analyses, namely that privately-owned firms should be more efficient, and thus have lower costs than their public counterparts, has received rather weak support from empirical, comparative, studies. Public enterprises are often found to be no less efficient than private firms, see, e.g., Feigenbaum and Teeples (1983) and Byrnes et al. (op. cit.) for studies of water utilities, Färe, Grosskopf and Logan (1985) and Atkinson and Halvorsen (1986) on electric utilities, and Register and Bruning (1987) concerning hospital care. The last study may be critized for employing a questionable output measure but that criticism does not apply to the public utilities analyses where output measurement is relatively straightforward.

[22] Another possible cause may be the existence of regulatory constraints, see, e.g., Atkinson and Halvorsen (1980, 1984, 1986). Moreover, if the exogenously given demand is highly variable it may be impossible to avoid some slack in off-peak periods in order to be able to cope with the peaks, cf. Fuss and McFadden (1978).

merely examining in what ways the *existence* of inefficiencies can be modeled and, secondly, how their effects on total costs and input demands can be estimated. Also in line with the literature, we will henceforth sometimes speak of the *degree of efficiency* instead of inefficiency. The degree of efficiency should be interpreted here as a (truncated) fractional measure such that a degree of efficiency in the open interval]0,1[implies a certain amount of inefficiency, whereas a degree of efficiency equal to one means (fully) efficient.

We have chosen to model inefficiency parametrically, which makes it possible to implement our results with a wide variety of flexible functional forms. Two types of inefficiency are considered: technical inefficiency, concerning overutilization of inputs, and allocative or price inefficiency, referring to situations where the factor proportions are inconsistent with cost minimization, given the relative input prices.

Among the various concepts of input efficiency we are thus not considering scale inefficiency.[23] The reason is that, by construction, our approach does not permit any analysis of scaling properties as the cost shares are invariant to the scale of production. This means, however, that if production is not scale efficient this will not introduce any bias in the conclusions that we actually are able draw by means of the system of input cost shares. It should also be noted that from the perspective of an individual producer, scale efficiency is not a well defined concept in the sense that it is not always consistent with cost minimization. The potential inconsistency arises when the level of output is exogenously given to the producer and the optimal scale, y^* say, i.e. the scale for which the dual scale elasticity $(e_{Cy})^{-1}$ is equal to

---

[23] For a thorough discussion of the various concepts of productive efficiency, see Färe, Grosskopf, and Lovell (1985). In addition to technical, allocative, and scale efficiency they consider yet another input efficiency concept, namely that of (absence of) congestion. However, congestion can only arise when the technology is characterized by weak disposability of inputs (WDI), implying that an increase in the utilization of some input(s) may in some cases decrease the amount of output. Since free disposability of inputs (FDI) – increases in input can never decrease output – is one of the regularity conditions which have to be fulfilled to ascertain a dual representation of a production technology [cf., e.g., Diewert (1971)], WDI technologies, and thus congestion, are of no interest in our context.

unity, is greater than the exogeneously given level of output, y. In that case an adjustment towards scale efficiency can never decrease total costs, but may well increase them, since $C(y,p,t)$ is non-decreasing in y; cf. Section 2. In the following we will take the perspective of the individual firm. Thus, we will consider a producer which is both technically and allocatively efficient to be overall efficient and identify minimum total costs with the total costs incurred in the context of the so defined overall efficiency.

We will begin by discussing how the system (9) of cost-minimizing cost shares can be generalized to take allocative inefficiency into account, given that the production process is technically efficient. We then show that the resulting cost shares are invariant with respect to the possible existence of *radial technical inefficiency*, the technical inefficiency counterpart to neutral technical change. Next, we demonstrate that it is possible to formulate a cost share system which captures the *combined* effect of both allocative and technical inefficiency – but not their separate effects. We establish, however, that under weak conditions the information required for a decomposition of the combined effect is contained in the first specification, i.e. the one allowing explicitly for allocative and implicitly for radial technical inefficiency. Finally, we show how the computations of the price and substitution elasticities are affected by deviations from cost minimization and, similarly, how the various measures of technical change should be calculated.

4.1. *Allocative inefficiency.* As is well known, the first order conditions for cost minimization require that the inputs be chosen such that the ratio of their marginal productivity values, or shadow prices, be equal to the ratio of their (actual) prices. Since the marginal rates of technical substitution are equal to the corresponding ratios of marginal productivity values, this requirement can equivalently be expressed as requiring equality between

$$\frac{\partial\psi(\tilde{\mathbf{x}},t)/\partial x_i}{\partial\psi(\tilde{\mathbf{x}},t)/\partial x_j} \equiv \frac{w_i}{w_j} \quad \text{and} \quad \frac{p_i}{p_j}$$

for all $i \neq j$, where, as before, $\psi(\cdot)$ denotes the production function to which the cost function $C(\cdot)$ is dual, $\partial\psi(\tilde{\mathbf{x}},t)/\partial x_i$ denotes the partial derivative of ψ with respect to x_i, evaluated at the (hypothetical) point $\tilde{\mathbf{x}} = (\tilde{x}_1,\ldots,\tilde{x}_n)$, and w_i is the shadow price of input i. Using w_n and p_n to normalize, the production process can be defined as allocatively efficient if $w_i/w_n = p_i/p_n$ for $i = 1,\ldots,n\text{-}1$.

A simple, yet quite powerful, specification by means of which deviations from allocative efficiency can be studied is the following one, originally proposed by Lau and Yotopoulos (1971) and introduced in a dual, time series context by Toda (1976, 1977).[24] The shadow prices are assumed to be proportional to the factor prices actually observed, the p_i's, according to

$$(18) \qquad w_i = \lambda_i p_i\,, \qquad \lambda_i > 0\,, \qquad i = 1,\ldots,n\,,$$

where λ_i is an (unknown) input-specific proportionality constant.[25]

In this section, we are assuming that the production process is technically efficient. Thus, the *realized* cost shares – as opposed to the cost minimizing shares – can be derived as follows. First, notice that if the producer is technically efficient his/her choice of input levels can be regarded as the result of minimizing total shadow costs, $\Sigma_{k=1}^{n} w_k x_k$, with respect to the x_k's. Using (8) and (18), the *minimum total shadow costs* can be expressed in terms of the actually observed prices as

---

[24] Toda considered only the two input case. The generalization to the n input case which we use in the following is due to Atkinson and Halvorsen (1980, 1984).

[25] It is of course possible to conceive of other ways than (18) to model deviations from allocative efficiency. An additive formulation of the type $w_i = p_i + \theta_i$, where θ_i is an unknown parameter, can be found in Eakin and Kniesner (1988). The specification (18) is, however, by far, the one most commonly used.

(19) $$\tilde{C} = \tilde{C}(y,\Lambda_d p,t) = f(y) \cdot g(\Lambda_d p,t)\,,$$

where Λ_d denotes an $n \times n$ diagonal matrix with ii:th element equal to λ_i. It is easily established that $\tilde{C}(y,\Lambda_d p,t)$ fulfills the regularity conditions cited in Section 2 when the input price vector is taken to be $w = \Lambda_d p$ and, thus, that for this price vector it is a proper dual representation of some underlying production technology. Application of Shephard's lemma to (19) yields the input levels which minimize total shadow costs, the $\tilde{x}_i$'s, according to

(20) $$\tilde{x}_i \equiv \frac{\partial \tilde{C}}{\partial(\lambda_i p_i)} = f(y) \cdot \frac{\partial g(\Lambda_d p,t)}{\partial(\lambda_i p_i)}\,, \qquad i = 1,\ldots,n\,.$$

Using (20), the *realized total cost*, $\tilde{C}^r$, can be written

(21) $$\tilde{C}^r \equiv \sum_{k=1}^{n} p_k \tilde{x}_k = f(y) \cdot \sum_{k=1}^{n} p_k \cdot \frac{\partial g(\Lambda_d p,t)}{\partial(\lambda_k p_k)}\,,$$

which differs from $\tilde{C}$ in that the partial derivatives are weighted by the actual input prices, the p_k's, rather than by the shadow prices, the w_k's. Concerning the relationship between $\tilde{C}^r$ and C, their respective definitions imply that $\tilde{C}^r \geq C$. Equality holds only if $\lambda_k = 1$ for $k = 1,\ldots,n$, in which case $\tilde{C}^r = \tilde{C} = C$.

The system of cost shares to be estimated will thus be

(22) $$\tilde{S}_i^r \equiv \frac{p_i \cdot \tilde{x}_i}{\tilde{C}^r} = \frac{p_i \cdot \dfrac{\partial g(\Lambda_d p,t)}{\partial(\lambda_i p_i)}}{\sum_{k=1}^{n} p_k \cdot \dfrac{\partial g(\Lambda_d p,t)}{\partial(\lambda_k p_k)}}\,, \qquad i = 1,\ldots,n.$$

However, as $\tilde{C}$ is linearly homogeneous in the $(\lambda_i p_i)$'s, its derivatives, (20), must be homogeneous of degree zero in the same variables or, equivalently, in the λ_i's.

Accordingly, the cost shares (22) must be homogeneous of degree zero in the λ_i's, too. Therefore, the absolute values of the λ_i's cannot be obtained from estimation of the system (22). This property is consistent with the fact that the first order conditions for cost minimization only concern relative prices. The following normalization rule can thus be imposed without loss of generality

$$\lambda_n = 1 \,, \tag{23}$$

cf., e.g., Atkinson and Halvorsen (1984).

Regarding the λ_i's which are to be estimated (i.e. $\lambda_1,\ldots,\lambda_{n-1}$) the positivity constraints [cf. (18)] constitute a potential estimation problem. To ascertain that the λ_i's stay positive they can be defined in terms of a transformation function, according to $\lambda_i \equiv \varphi(\mu_i)$ where μ_i is an unrestricted parameter and φ a function whose image is equal to the set of positive real numbers. For instance, φ might be an exponential function as suggested by Lau (1978). Another example is the hyperbolic transformation proposed by Mellander and Jansson (1987), which has the attractive property that it leaves the estimation practically unaffected as long as the positivity constraints are not binding..

Testing allocative efficiency means testing the hypothesis that all the λ_i's are equal to unity, in which case (22) is identically equal to the system (9) of cost minimizing shares. Fig. 1 can be used to illustrate the test in the two input case. To be capable of illustrating both allocative and technical inefficiency the diagram has been drawn in the space of input/output-coefficients. Thus, all points lying on or to the northeast of the isoquant II' correspond to same volume of output.

Fig. 1.

The isocost shown by a solid line corresponds to the factor prices actually

observed, i.e. p_1 and p_2. Since we are here assuming that the producer is technically efficient, production must be taking place somewhere along the isoquant II'. The producer will minimize costs by operating at the point E. We assume, however, that production is actually taking place at the point M. With the input prices at the observed levels this point is obviously not allocatively efficient. However, M would have been an allocatively efficient location if the isocost had been given not by the solid but by the dotted line. The slope, α, of this latter isocost equals the ratio of the shadow prices since, given that production occurs at M, this is the relative price corresponding to cost minimization. The hypothesis to be tested is thus whether the slope of the hypothetical isocost, α, is significantly different from ν, the slope of the actual isocost. In the two input case this simply means testing if $\lambda_1 = 1$ since, in accordance with (23), $\lambda_2 = 1$ *a priori*.

Farrell (1957) has suggested a scalar measure of the degree of price efficiency. In terms of Fig. 1, Farrell's measure of allocative efficiency (AE) is defined as

$$AE \equiv \frac{OZ}{OM}. \tag{24}$$

This ratio is equal to the relation between the costs which would have resulted at the efficient point, E, (corresponding to OZ) and the total costs incurred at the actual point of production, M. Thus, AE can be computed according to

$$AE = \frac{C}{\tilde{C}^r} = \frac{g(p,t)}{\sum_{k=1}^{n} p_k \cdot \frac{\partial g(\Lambda_d p,t)}{\partial(\lambda_k p_k)}}. \tag{25}$$

The denominator in the last equality is equal to the denominator of the realized cost shares (22) and, thus, can be directly obtained from the estimation of that system. The numerator is also easy to obtain; due to the linear homogeneity of $g(\Lambda_d p,t)$ and $g(p,t)$ in $\Lambda_d p$ and p, respectively, it holds that

$$(26) \qquad g(p,t) = \sum_{k=1}^{n} (\lambda_k p_k) \cdot \frac{\partial g(\Lambda_d p, t)}{\partial(\lambda_k p_k)} \Bigg|_{\Lambda_d = I_n}$$

where I_n is the identity matrix of order n. Thus, $g(p,t)$ can be computed simply by setting all the λ_i's in the denominator of (25) equal to one. Notice that changes in the relative input prices and in the time index will cause AE to vary over time, yielding estimates of the degree of allocative efficiency for each point of observation.

From (25) it is clear that once AE has been computed we can easily estimate the relative increase in total costs caused by the misallocation of inputs, in spite of the fact that we have no measure of output. The relative increase $(\tilde{C}^r - C)/C$ is simply equal to $(1 - AE)/AE$. Finally, for later reference, we note that the cost-minimizing input demands can be expressed in terms of the $\tilde{x}_i$'s, according to

$$(27) \qquad x_i = \tilde{x}_i \cdot \frac{\dfrac{\partial g(p,t)}{\partial p_i}}{\dfrac{\partial g(\Lambda_d p, t)}{\partial(\lambda_i p_i)}}.$$

4.2. *Allocative and technical inefficiency* We now relax the assumption of technical efficiency. In general terms, a producer is defined as technically inefficient if, at a given level of production, he/she can reduce the utilization of any input and still produce the same amount of output. In Fig. 1 above, technical inefficiency is illustrated by the point B, which cannot be technically efficient as it is not on the efficient production surface II'.

Farrell (op. cit.) has proposed a simple measure of the degree of technical efficiency (TE). In terms of the diagram, it is defined as

$$(28) \qquad TE \equiv \frac{OM}{OB}.$$

A convenient interpretation of this measure is obtained by considering the difference

$1 - TE$ which, by definition, belongs to the open interval $]0,1]$. For the given level of output, $1 - TE$ shows the potential relative decrease in the input utilization, when the factor proportions are held constant, i.e. when the relative reduction is constrained to be the same for all inputs. Since TE is defined relative to the factor ray through the origin and the observed point it is a *radial* measure of technical efficiency. Like the Farrell measure of allocative efficiency *(AE)*, TE can also be expressed in terms of total costs. The ratio OM/OB is equal to the total costs associated with the technically efficient (but allocatively inefficient) point M, divided by the total costs incurred at B, the point actually observed. If we denote the actually observed total costs by C^a then

$$TE = \frac{\tilde{C}^I}{C^a} \tag{29}$$

In Farrell's original formulation TE was defined in the context of a constant returns technology. However, radial measures of technical inefficiency can be applied to homothetic technologies, too.

An appealing property of the technical efficiency measure TE is that it does not affect the Farrell measure of the degree of allocative efficiency, AE. This is easily seen in Fig. 1. Different degrees of technical efficiency correspond to different locations on the dashed ray through the origin, on or above the isoquant II'. But for all these points the degree of allocative efficiency is the same, namely OZ/OM. This means that degrees of allocative and technical efficiency can be independently computed. The degree of overall efficiency *(OE)* is simply given by

$$OE = TE \times AE = \frac{C}{C^a}, \tag{30}$$

where the last equality follows from (25) and (29).

Unfortunately, it is not possible to estimate TE directly by means of the system of cost shares. The reason is that the radial specification of technical inefficiency is *input neutral* and, hence, like neutral technical change, has no effect on the input cost shares. This is easy to show formally, as follows. Let x_i^a denote the actually observed usage of input i. We may then define x_i^a in terms of the technically efficient (but allocatively inefficient) input demands given by (20). In order to be consistent with (28) the definition must have the following form

$$x_i^a \equiv (1 + \zeta) \cdot \tilde{x}_i , \qquad \zeta \geq 0, \qquad i = 1,...,n. \tag{31}$$

where ζ represents the common degree of overutilization, implying that

$$TE = (1 + \zeta)^{-1} . \tag{32}$$

However, in analogy with (21), the total costs corresponding to (31) are

$$C^a \equiv \sum_{k=1}^{n} p_k x_k^a = (1 + \zeta) \cdot \tilde{C}^r . \tag{33}$$

Together with (31), (33) implies that the actual cost shares, defined according to $S_i^a \equiv p_i \cdot x_i^a / C^a$, are equal to the cost shares prevailing in the context of allocative inefficiency only, i.e $S_i^a = \tilde{S}_i^r$ for $i = 1,...,n$. It should be noticed that this result implies that, in addition to allocative inefficiency, the system (22) also implicitly allows for possible radial technical inefficiency.[26]

However, to be able to take technical inefficiency *explicitly* into account we have to let it affect the input usage in a *non-radial* fashion, i.e. allow the degree of

---

[26] Since the system (9) of input cost shares is a special case of the system (22), this invariance property implies that in the presence of radial technical inefficiency estimation of the system (9) is still valid, and will yield unbiased estimates, in spite of the fact that the assumption of cost minimization is violated.

overutilization to vary among the inputs.[27] To this end we will derive a system of input cost shares which takes the combined effects of technical and allocative inefficiency, i.e. overall inefficiency, into account and which includes the system (22) as a special case.

Unfortunately, in the system of cost shares allowing for overall inefficiency it is not possible to separate allocative from technical inefficiency in an unambiguous way. The reason is that the introduction of input-specific degrees of overutilization removes the independence between the measures of technical and allocative efficiency, which is characteristic of the Farrell scheme, cf. Kopp (1981).[28] Provided, however, that we make the assumption that the production technology satisfies *strong free disposability of inputs* (SFDI) the system allowing for overall inefficiency can be combined with the system (22) to yield a Farrell decomposition of the overall inefficiency in accordance with (25) and (29). SFDI implies that when production is taking place at a technically efficient point an increase in the utilization of some input(s) will always result in some, however small, increase in output. As noted by Kopp (op. cit.), most of the functional forms employed in econometric production studies satisfy SFDI. Among them are the CES and the translog; cf. Färe and Lovell (1978) and Kopp and Diewert (1982), respectively.[29] The condition of SFDI ascertains that a given degree of overall efficiency can always be equivalently decomposed into *either* non-radial technical inefficiency and allocative inefficiency *or* radial technical inefficiency and allocative inefficiency (although the measures of

---

[27] Non-radial specifications of technical inefficiency have been considered by Färe (1975) and by Färe and Lovell (1978).

[28] This property does not seem to have been generally recognized in the literature. For instance, in the empirical application of a model allowing for both allocative and non-radial technical inefficiency, Lovell and Sickles (1983) use an estimation method which treats these two types of inefficiency as if they were independent.

[29] Notice that the condition of SFDI is slightly more restrictive than that of free disposability of inputs (FDI), which is fulfilled by all technologies which have a dual representation (cf. footnote 23). An example of a flexible functional form which does not satisfy SFDI globally is the Generalized Leontief. In particular, its special case the (ordinary) Leontief technology fails SFDI everywhere.

allocative inefficiency will differ in the two cases).

In our context there is an additional complication: while a non-radial specification of technical inefficiency implies that there are n different degrees of overutilization to estimate – one for each input – the system of input cost shares can only provide us with n-1 independent estimates, as the share system is of rank n-1. Due to the interdependence between the measures of non-radial technical inefficiency and allocative inefficiency we can always set the overutilization of an arbitrarily chosen input equal to zero, however. To see this, consider Fig. 2.

Fig. 2.

The isoquant and the points B, M, and E have been reproduced from Fig. 1. We now make the thought experiment that the producer operating at the point B moves to the efficient point, E. This movement, illustrated by the solid arrow, can be considered as the sum of two vectors, representing movements towards technical and allocative efficiency, respectively. In principle, the sum can be decomposed in an infinite number of ways. The vector corresponding to the adjustment towards technical efficiency must, however, result in a point on the boldly drawn part of the isoquant whose endpoints coincide with the points M' and M''. This is so because, by definition, technical inefficiency corresponds to *over*utilization of inputs. The movement to a technically efficient point thus cannot involve an increase in the use of any input. Of the infinitely many admissible decompositions of the sum, two are shown in the figure.

The dashed vectors illustrate the special case in which the adjustment towards technical efficiency is radial, i.e. when (31) holds. The adjustment yielding technical efficiency is represented by the vector from B to M, whereas the other vector is equivalent to the movement from M to E, i.e. the movement for allocative efficiency. Of the dotted vectors the one pointing due south, to M',

corresponds to a non-radial adjustment towards technical efficiency where the amount of x_2 is held constant while decreasing the use of x_1. By elimination, the other vector must then show the movement yielding allocative efficiency.

Another possible alternative would be to use M'' as the reference point for technical efficiency. Due to the interdependence between non-radial technical inefficiency and price inefficiency it is always possible to decompose the overall inefficiency (i.e. the solid arrow) in such a way that the adjustment towards technical efficiency results in either of the points M' and M'', corresponding to zero overutilization of x_2 and x_1, respectively. The effect of choosing M'' instead of M' is just that it yields another decomposition of overall efficiency into allocative and technical components.

Obviously, if the overutilization of one of the inputs can be set equal to zero in the two input case then, *a fortiori*, it must be possible to impose this constraint in the context of n inputs, too. We now proceed to derive a system of input cost shares allowing for both allocative inefficiency and non-radial technical inefficiency. We begin by assuming that the actual input demands (31) can be equivalently represented according to

$$x_i^a = f(y)\cdot\frac{\partial g(\check{\Lambda}_d p,t)}{\partial(\check{\lambda}_i p_i)} + f(y)\cdot\delta_i\,, \quad \delta_i \geq 0 \quad i = 1,\ldots,n. \tag{34}$$

The last term on the right hand side (RHS) represents the excessive usage of input i. For simplicity, the δ_i's are here taken to be parametrical constants.[30] The excessive input usage is thus assumed to vary between the inputs and to change with the scale of operation; in the context of a constant returns to scale technology

[30] As parameters, the δ_i's may not be identified for all kinds of functional forms. However, regarding, e.g., the CES and translog functional forms, which we know satisfy SFDI, identification is always possible.

the (input-specific) overutilization is proportional to the level of output.[31] In accordance with the above discussion one of the δ_i's can be set equal to zero, e.g.

$$\delta_n = 0 . \tag{35}$$

Notice that, in general, the first term on the RHS of (34) is not equal to $\tilde{x}_i$, given by (20). The "˘" on the matrix $\breve{A}_d$ indicates that the δ_i's will affect the estimated shadow prices and, hence, also the estimated degree of allocated efficiency, as shown above. Only if all the δ_i's are equal to zero will the matrices $\breve{A}_d$ and A_d be equal. The partial derivative $\partial g(\breve{A}_d p,t)/\partial(\breve{\lambda}_i p_i)$ may thus be either greater or smaller than the partial derivative $\partial g(A_d p,t)/\partial(\lambda_i p_i)$ which, together with $f(y)$, determines $\tilde{x}_i$, according to (20). This means that in addition to being non-negative the δ_i's must also fulfill the condition

$$\delta_i \geq \frac{\partial g(A_d p,t)}{\partial(\lambda_i p_i)} - \frac{\partial g(\breve{A}_d p,t)}{\partial(\breve{\lambda}_i p_i)}, \qquad i = 1,\ldots,n, \tag{36}$$

for all price vectors and and all values on t, in order to ensure that the inequalities $x_i^a \geq \tilde{x}_i$, $i = 1,\ldots,n$, always hold. As (35) will have the consequence that the RHS of the inequality (36) will always be negative for $i = n$ (cf. Figure 2) these constraints actually are of concern only with respect to the n-1 first inputs.

The specification (34) is just one among several possible ways to account for non-neutral technical inefficiency. We have chosen this particular specification for two reasons: it is simple and, in contrast to related specifications used by, e.g., Lovell and Sickles (1983), it leads to input cost shares which are independent of y.

Up to a constant of integration, the cost function corresponding to (34) is

---

[31] It should be noticed that since the overutilization is measured in the same units as the the left hand side variables in (36) technical inefficiency is not radial even if all the δ_i's are equal in magnitude.

$$(37) \qquad \check{C} = \check{C}(y,\check{A}_d p,t,\delta) \equiv f(y)\cdot[g(\check{A}_d p,t) + \delta'\check{A}_d p] ,$$

where the prime in the last term denotes transposition. It is straightforward to show that for the price vector $\check{w} = \check{A}_d p$ the cost function $\check{C}$ is regular. Thus, $\check{C}$ can safely be considered to be the dual representation of some production technology and the application of Shephard's lemma to (37), yielding (34), is justified. The total costs actually observed are not given by (37), however. In accordance with the definition given in (33) the total costs actually observed can be expressed as

$$(38) \qquad C^a = f(y) \cdot \sum_{k=1}^{n} p_k \cdot \left[\frac{\partial g(\check{A}_d p,t)}{\partial(\check{\lambda}_k p_k)} + \delta_k \right] .$$

If all the δ_k's are equal to zero (and, hence, $\check{\lambda}_k = \lambda_k$ for $k = 1,...,n$) then (38) reduces to (21), the total cost realized in the context of allocative inefficiency only. Given (34) and (38) the observed input costs shares can be written

$$(39) \qquad S_i^a \equiv \frac{p_i \cdot x_i^a}{C^a} = \frac{p_i \cdot \dfrac{\partial g(\check{A}_d p,t)}{\partial(\check{\lambda}_i p_i)} + p_i\delta_i}{\sum_{k=1}^{n} p_k \cdot \left[\dfrac{\partial g(\check{A}_d p,t)}{\partial(\check{\lambda}_k p_k)} + \delta_k \right]} , \qquad i = 1,...,n.$$

In the estimation of (39), the $\check{\lambda}_i$'s should be subjected to the same constraints as the those imposed on the λ_i's in the estimation of the system (22).[32] Concerning the δ_i's the non-negativity restrictions in (34) pose no problem; they can be implemented by means of the same method as the one employed to ensure positive values on the λ_i's and the $\check{\lambda}_i$'s. The inequality constraints (36) are more difficult to impose, however. The simplest way to proceed is probably to ignore them in a

---

[32] For clarity, it should be pointed out that "estimation of (22)" is equivalent to "estimation of (39) subject to the constraint that $\delta_i = 0$ for $i = 1,...,n$". We use the former expression for the obvious reasons that it is shorter and simpler.

first round estimation. Should a comparison with the estimates obtained in the estimation of (22) reveal that any of the δ_i's violate (36) for some observations on the input price vector and the time index, the lower bound for these parameters can be raised according to $\delta_i \geq \hat{\delta}_i + \kappa_i$ where $\hat{\delta}_i$ is the first round estimate and κ_i a suitably chosen positive number. It is not certain that the second-round estimates will satisfy (36) at all observations either but the inequalities can always be made to hold by repeating the procedure; the lower bounds of the δ_i's may be increased from zero to the highest value obtained for the partial derivative $\partial g(\Delta_d p,t)/\partial(\lambda_i p_i)$ in the estimation of the system (22).[33]

When both the systems (22) and (39) have been estimated a likelihood ratio test can be performed of the null hypothesis $H_o : \delta_i = 0,\ i = 1,...,n$. The test of H_o corresponds to a weak test of technical efficiency – the test is weak in the sense that the system under the null, i.e. (22), is consistent both with technical efficiency and radial technical inefficiency. Rejection of H_o implies, however, that the production process cannot be technically efficient.

At first, rejection of the hypothesis that all the δ_i's are equal to zero might seem as an implausible outcome. If the two decompositions of overall inefficiency, involving radial and non-radial technical inefficiency, respectively, are indeed equivalent, then why should the latter be preferred to the former ? However, this objection fails to recognize that the fact that there exists alternative decompositions of overall inefficiency which are *mathematically* equivalent does not imply that these alternatives are also statistically equivalent, in the sense of providing equally good fit to data. Rather, one would expect the more richly parameterized alternative to be preferred to the more parsimonious one. Hence, if the production process is technically inefficient, rejection of H_o should be more likely than acceptance.

The estimated versions of the systems (22) and (39), yield an estimate of *TE*

---

[33] These possible adjustments of the lower bounds only concern $\delta_1,...,\delta_{n-1}$. As remarked above, for $i = n$ the constraint (36) will be fulfilled automatically, on account of the restriction (35).

according to

$$(40) \qquad TE = \frac{\sum_{k=1}^{n} p_k \cdot \frac{\partial g(\Lambda_d p, t)}{\partial(\lambda_k p_k)}}{\sum_{k=1}^{n} p_k \cdot \left[\frac{\partial g(\check{\Lambda}_d p, t)}{\partial(\check{\lambda}_k p_k)} + \delta_k\right]},$$

cf. (29). Like the Farrell measure of the degree of allocative efficiency, (25), TE varies over time in response to changes in the relative input prices and the time index. Furthermore, (32) shows that ζ, the common degree of overutilization corresponding to neutral technical inefficiency, is given by

$$(41) \qquad \zeta = \zeta(\Lambda_d p, \check{\Lambda}_d p, t) = \frac{1}{TE} - 1 .$$

Thus, ζ is not a constant but a function, determined by the input prices and the time index. Given the estimate of ζ, the input utilization in the context of only allocative inefficiency, i.e. the $\tilde{x}_i$'s, can be computed by dividing the actually observed input usage x_i^a, $i = 1,...,n$, by $(1 + \zeta)$; cf. (31).[34] Finally, by inserting the so obtained estimates into (27) we obtain estimates of the cost-minimizing input demands, too. Hence, it is possible to compare the input usage actually observed with the cost-minimizing levels of utilization and to compute the minimum total costs, i.e. $C \equiv \Sigma_{k=1}^{n} p_k x_k$, in spite of the presumed lack of output measure.

---

[34] Comparison of (34) and (20) shows that, in principle, the $\tilde{x}_i$ can also be obtained as

$$\tilde{x}_i = x_i^a \cdot [\partial g(\check{\Lambda}_d p,t)/\partial(\check{\lambda}_i p_i) + \delta_i]^{-1} \cdot [\partial g(\Lambda_d p,t)/\partial(\lambda_i p_i)] , \quad i = 1,...,n.$$

However, due to random errors this procedure may violate the condition that $\tilde{x}_i/x_i^a$ be equal for all i, in which case it is inconsistent with the definition of neutral technical inefficiency. While this method and the one advocated in the text are equivalent in expectation, the latter method has the advantage of avoiding this potential inconsistency problem.

4.3. *Computation of price and substitution elasticities, and effects of technical change.* To enable comparisons between the price and substitution elasticities prevailing under cost-minimization, i.e. (6) and (7), and those corresponding to the input utilization actually observed, we show here how the latter elasticities should be computed. Likewise, we consider the effects of technical change corresponding to the x_i^a, $i = 1,...,n$.

We first derive the elasticities of substitution. These should be defined in terms of the cost function from which the x_i^a have been derived. As $x_i^a = \partial\check{C}/\partial(\check{\lambda}_i p_i)$, $i = 1,...,n$, this means that the cost function (37) should be used. In order to obtain a formula for the actual elasticities which is analogous to (6) we need the input cost shares which are minimum for this cost function. Denoting these by $\check{S}_i$ we obtain

$$\check{S}_i \equiv \frac{(\check{\lambda}_i p_i)\cdot\dfrac{\partial\check{C}}{\partial(\check{\lambda}_i p_i)}}{\check{C}} = \frac{(\check{\lambda}_i p_i)\cdot\left[\dfrac{\partial g(\check{A}_d p,t)}{\partial(\check{\lambda}_i p_i)} + \delta_i\right]}{g(\check{A}_d p,t) + \delta'\check{A}_d p}, \tag{42}$$

$i = 1,...,n$. To compute the numerator of $\check{S}_i$ we simply multiply the estimated numerator of S_i^a by $\check{\lambda}_i$; cf. (39). And, as usual, the denominator is equal to the sum of the numerators. In analogy with (6), the actual elasticities, which we denote by σ_{ij}^a, can be expressed in terms of the $\check{S}_i$'s, according to

$$\sigma_{ij}^a = \left[\check{S}_i\check{S}_j + (\check{\lambda}_j p_j)\cdot\frac{\partial\check{S}_i}{\partial(\check{\lambda}_j p_j)}\right]\cdot\frac{1}{\check{S}_i\check{S}_j}. \tag{43}$$

Because of the proportionality between the shadow prices and the input prices actually observed the price elasticities can be obtained as follows

$$\eta_{ij}^a \equiv \frac{\partial x_i^a}{\partial p_j}\frac{p_j}{x_i^a} = \frac{\partial x_i^a}{\partial(\check{\lambda}_j p_j)}\frac{(\check{\lambda}_j p_j)}{x_i^a} = \check{S}_j\sigma_{ij}^a, \tag{44}$$

where the first equality is due to the chain rule and the second equality follows by analogy with (7); this analogy is justified because the $(\check{\lambda}_j p_j)$'s are the prices for which the cost function (37) is defined.

Concerning technical change, its effects on input utilization are given by

$$(45) \qquad \tau^a_{x_i} \equiv \frac{\partial x^a_i(y,\check{A}_d p,t)}{\partial t}\,\frac{1}{x^a_i(y,\check{A}_d p,t)}$$

$$= \frac{\partial^2 g(\check{A}_d p,t)}{\partial(\check{\lambda}_i p_i)\,\partial t}\left[\frac{\partial g(\check{A}_d p,t)}{\partial(\check{\lambda}_i p_i)}\right]^{-1}, \qquad i = 1,\ldots,n.$$

Regarding the effects on total costs, two aspects are relevant. On the one hand, it is of interest to consider the influence that technical change has had on the total costs actually observed, i.e.

$$(46) \qquad \tau^a_C \equiv \frac{\partial C^a}{\partial t}\,\frac{1}{C^a} = \sum_{k=1}^{n} S^a_k\,\tau^a_{x_i}.$$

On the other hand, from the producer's point of view it is relevant to compute the effects of technical change on the total cost function (37), as (37) is the dual representation of the production technology (under the false perception that the input price vector is given by $\check{w} = \check{A}_d p$ rather than p). This means that in the aggregation of the $\tau^a_{x_i}$'s the $\check{S}_i$'s should be used as weights according to

$$(47) \qquad \check{\tau}_C \equiv \frac{\partial \check{C}}{\partial t}\,\frac{1}{\check{C}} = \sum_{k=1}^{n} \check{S}_k \tau^a_{x_i}.$$

Like τ_C [defined by (15)] $\check{\tau}_C$ has a dual interpretation: $-\check{\tau}_C$ is the rate of technical change in the production function to which (37) is dual. No dual

interpretation is possible with respect to τ_C^a, however, since C^a is not a regular cost function. Finally, by analogy with (12), the relative changes in S_i^a and $\check{S}_i$ brought about by technical change are equal to $\tau_{x_i}^a - \tau_C^a$ and $\tau_{x_i}^a - \check{\tau}_C$, respectively.

5. SUMMARY AND CONCLUSIONS

What can you learn about a production process for which no output measures are available ? This is the question we have tried to answer with the help of duality theory. Our results show that the possibilities to characterize production by means of input data only are indeed much greater than could be expected intuitively.

The fundamental property upon which we base our analysis, i.e. the fact that for a homothetic technology the input cost shares can be completely specified without any information about output, was implicit already in Shephard (1953). The importance of this result for applied production theory seems not to have been recognized, however, which is surprising considering the tremendous growth that has occured since then in the production of services, where the output measurement problems are especially severe. It is significant that Hulten's (1984) study of productivity change in the public sector, which is the only previous attempt to characterize a production process econometrically without explicit measures of the price or quantity of output, did not escape the output measurement problem by considering the input cost shares. Instead, Hulten chose to regard communities as generalized households, thereby making it possible to apply the analytical apparatus of the household production model to the production of public services.

In contrast to Hulten's framework, our method can be applied to any production activity. Moreover, our analysis goes beyond Hulten's in that it is not limited to the issue of estimating productivity growth. We show that given a homothetic

technology, knowledge of input prices and input cost shares makes it possible to estimate elasticities of substitution and factor demand, analyze productivity effects of technical change, and study (deviations from) efficiency in production. Although the homotheticity assumption might be too restrictive in the context of goods production we argue that is is more easily motivated in the production of services, primarily because of the limited possibilities to routinize services.

Concerning the relationship between technical change and productivity growth, we show that the relative effects of technical change on total costs always can be estimated but that these correspond to estimates of the dual rate of growth in total factor productivity (TFP) only if constant returns to scale are assumed, as in Hulten's study. If, instead, homogeneity of degree $r \neq 1$ is assumed the rate of growth in TFP can be estimated up to an initial condition or bench-mark value, while homotheticity allows only the sign of the TFP growth rates to be determined.

We also demonstrate how possible deviations from cost minimization can be taken into account parametrically. Here, we make use of the fact that for a large class of technologies overall inefficiency can be decomposed either into independent measures of *neutral* technical inefficiency and allocative inefficiency according to Farrell (1957), or into two interdependent measures of *non-neutral* technical inefficiency and allocative inefficiency.[35] Since the input cost shares are invariant with respect to neutral technical inefficiency, the Farrell decomposition results in a share system which can take allocative inefficiency explicitly into account but which only allows for (neutral) technical inefficiency implicitly, making it impossible to quantify the latter. The second decomposition, on the other hand, yields a system of cost shares by means of which overall inefficiency can be measured but which cannot separate clearly between technical and allocative inefficiency. We show, however, that by estimating *two* share systems, one for each decomposition, Farrell

[35] The measures of allocative inefficiency will, of course, differ between the two alternative decompositions.

measures of technical, allocative, and overall inefficiency can be obtained. Moreover, the increases in total costs brought about by the inefficiencies can be estimated as well as the cost-minimizing input demands, in spite of the presumed lack of output data.

REFERENCES

Allen, R.G.D., 1959, Mathematical analysis for economists (Macmillan, London).

Atkinson, S.E. and R. Halvorsen, 1980, A test of relative and absolute price efficiency in regulated utilities, Review of Economics and Statistics 62, 74–80.

Atkinson, S.E. and R. Halvorsen, 1984, Parametric efficiency tests, economies of scale and input demand in U.S. electric power generation, International Economic Review 25, 647-662.

Atkinson, S.E. and R. Halvorsen, 1986, The relative efficiency of public and private firms in a regulated environmement: The case of U.S. electric utilities, Journal of Public Economics 29, 281-294.

Barten A.P., 1969, Maximum likelihood estimation of a complete system of demand equations, European Economic Review 1, 7-73.

Berndt, E.R. and M.S. Khaled, 1979, Parametric productivity measurement and choice among flexible functional forms, Journal of Political Economy 87, 1220-1245.

Binswanger, H.P., 1974, The measurement of technical change biases with many factors of production, American Economic Review 64, 964-976.

Blackorby, C., C.A.K. Lovell, and M.C. Thursby, 1976, Extended Hicks neutral technical change, Economic Journal 86, 845-852.

Byrnes, P., S. Grosskopf, and K. Hayes, 1986, Efficiency and ownership: Further evidence, Review of Economics and Statistics 68, 337-341.

Christensen, L.R., D.W. Jorgenson, and L.J. Lau, 1973, Transcendental logarithmic production frontiers, Review of Economics and Statistics 55, 28-45.

Diewert, W.E., 1971, An application of the Shephard duality theorem: A generalized Leontief production function, Journal of Political Economy 79, 481–507.

Eakin, B.K. and T.J. Kniesner, 1988, Estimating a non-minimum cost function for hospitals, Southern Economic Journal 54, 583-597.

Färe, R., 1974, On linear expansion paths and homothetic production functions, in:

W. Eichhorn, R. Henn, O. Opitz, and R.W. Shephard, R.W. eds., Production theory, (Springer-Verlag, Berlin-Heidelberg, New York).

Färe, R., 1975, Efficiency and the production function, Zeitschrift für Nationalökonomie 35, 317-324.

Färe, R., S. Grosskopf, and J. Logan, 1985, The relative performance of publicly-owned and privately-owned electric utilities, Journal of Public Economics 26, 89-106.

Färe, R., S. Grosskopf, and C.A.K. Lovell, 1985, The Measurement of efficiency of production (Kluwer Nijhoff, Boston).

Färe, R. and C.A.K. Lovell, 1978, Measuring the technical efficiency of production, Journal of Economic Theory 19, 150-162.

Farrell, M.J., 1957, The measurement of productive efficiency, Journal of the Royal Statistical Society, Series A 120, 253-281.

Feigenbaum, S. and R. Teeples, 1983, Public versus private water delivery: A hedonic cost approach, Review of Economics and Statistics 65, 672-677.

Førsund, F.R., 1975, The homothetic production function, Swedish Journal of Economics 77, 234-244.

Fuss, M. and D. McFadden 1978, Flexibility versus efficiency in *ex ante* plant design, in M. Fuss and D. McFadden eds., Production economics: A dual approach to theory and applications, Vol. 1 (North-Holland, Amsterdam) 311-364.

Hicks, J.R., 1932, The theory of wages, (Macmillan, London).

Hulten, C.R., 1984, Productivity change in state and local governments, Review of Economics and Statistics 66, 256-266.

Jorgenson, D.J. and Z. Griliches, 1967, The explanation of productivity change, Review of Economic Studies 34, 249-282.

Kopp, R.J., 1981, Measuring the technical efficiency of production: A comment, Journal of Economic Theory 25, 450-452.

Kopp, R.J. and W.E. Diewert, 1982, The decomposition of frontier cost function deviations into measures of technical and allocative efficiency, Journal of Econometrics 19, 319-331.

Lau, L.J., 1978, Testing and imposing monotonicity, convexity, and quasi-convexity constraints, in M. Fuss and D. McFadden eds., Production economics: A dual approach to theory and applications, Vol. 1 (North-Holland, Amsterdam) 409-453.

Lau, L.J. and P.A. Yotopoulos, 1971, A test for relative efficiency and application to Indian agriculture, American Economic Review 61, 94-109.

Lovell, C.A.K. and R.C. Sickles, 1983, Testing efficiency hypotheses in joint

production: A parametric approach, Review of Economics and Statistics 65, 51-58.

Mellander, E. and L. Jansson, 1987, CONRAD – A maximum likelihood program for estimation of non-linear simultaneous equations models,IUI Research Report No. 30 (Almqvist & Wicksell International, Stockholm).

Nadiri, I. and M. Schankerman, 1980, The structure of production, technological change, and the rate of growth of total factor productivity in the U.S. Bell system, in: T.G. Cowing and R.E. Stevenson, eds., Productivity measurement in regulated industries (Academic Press, New York) 219-247.

Ohta, M., 1974, A note on the duality between production and cost functions: Rate of returns to scale and rate of technical progress, Economic Studies Quarterly 25, 63-65.

Parks, R.W., 1971, Price responsiveness of factor utilization in Swedish manufacturing, 1870-1950, Review of Economics and Statistics 53, 129-139.

Register, C.A. and E.R. Bruning, 1987, Profit incentives and technical efficiency in the production of hospital care, Southern Economic Journal 53, 899-914.

Rhoades, S.A. and A.P. White, 1984, Output in relation to labor input in the banking and savings and loan industries 1927-1979, Journal of Banking and Finance 8, 119-130.

Sato, R., 1980, The impact of technical change on the holotheticity of production functions, Review of Economic Studies 67, 767-776.

Schwab, R.M. and E.M. Zampelli, 1987, Disentangling the demand functions from the production function for local public services: The case of public safety, Journal of Public Economics 33, 245-260.

Shephard, R.W., 1953, Cost and Production Functions (Princeton University Press, Princeton, New Jersey).

Toda, Y., 1976, Estimation of a cost function when the cost is not minimum: The case of Soviet manufacturing industries, 1958-1971, Review of Economics and Statistics 58, 259-268.

Toda, Y., 1977, Substitutability and price distortion in the demand for factors of production: An empirical estimation, Applied Economics 9, 203-217.

Uzawa, H., 1962, Production functions with constant elasticities of substitution, Review of Economic Studies 29, 291-299.

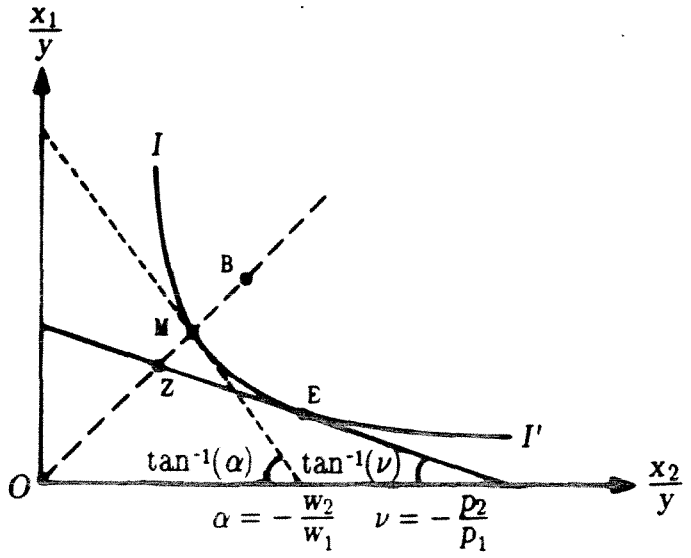

Fig. 1. Farrell measures of allocative, technical, and overall efficiency.

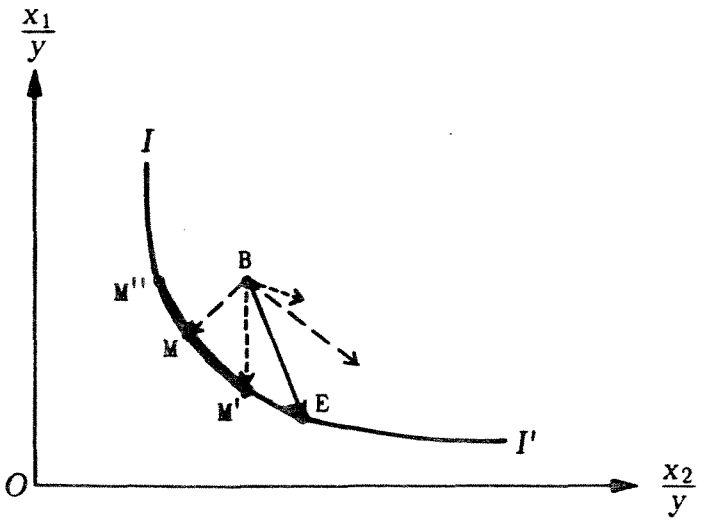

Fig. 2. Equivalent decompositions of overall efficiency.