Chapter 2

PRODUCTIVITY AND EFFICIENCY MEASUREMENT USING PARAMETRIC ECONOMETRIC METHODS

Subal C. Kumbhakar

Abstract

In this chapter we survey the econometric approaches to productivity and efficiency measurement. Both primal and dual approaches are considered. More specifically, we examine the production function and distance functions (multiple outputs), cost functions (with single and multiple outputs), standard and alternative profit functions (with single and multiple outputs). Possible extensions of the traditional productivity and efficiency measures in the light of non-traditional inputs are also discussed.

Keywords and Phrases: Partial and total factor productivity; technical change; returns to scale; technical efficiency; technical efficiency change; production function; distance function; cost function; profit function; alternative profit function; mixing models.

JEL Classification No.: C23, C33, D24, O30

“Productivity cannot be measured directly. Instead, it must be measured indirectly as a relationship between physical outputs and inputs that can be assembled.”

(John W. Kendrick 1984, p.9)

“To measure the productivity change of a firm or an industry, we first have to define what we mean by a productivity change.”

(W. Erwin Diewert 1992, p. 163)
1. INTRODUCTION\textsuperscript{1}

The most commonly used definition of productivity is the amount of physical output produced by one unit of a given factor of production at a stated period of time. Thus, productivity indicators ordinarily relate output to a single factor of production, creating measures such as labor productivity, capital productivity, etc. These are also defined as partial factor productivity. In contrast, multifactor (total factor) productivity measures output per unit of a set of combined factors of production (such as labor capital, land, etc.) and gives a single overall measure of productivity. These partial and multifactor productivity measures are based on quantities of inputs and outputs and are called primal measures of productivity. Instead of using input and output quantities, productivity can also be defined using cost, profit and price information. For example, average cost can be used to measure productivity. Note that these measures of productivity (average product and average cost) can be measured directly from the data.

Productivity (no matter how it is defined) is likely to change over time. In simple terms, productivity change/growth means more output is produced with a given level of inputs, which occurs when technology improves over time. Since inputs also change over time, productivity change is often defined as output growth net of input growth rate. Accordingly, productivity change can take place without technical change. In a single output single input case, productivity change is simply the rate of change in average product. We will give a formal definition of productivity change with single output and multiple inputs technology. This will be followed by multiple outputs multiple inputs technology. In addition to measuring productivity change, we also consider the sources of productivity growth.

Productivity and change in productivity can be measured using different techniques. Diewert (1992) showed that productivity change can be calculated using an index number approach (Fisher (1922) or Törnqvist (1936) productivity index. Both indices require quantity and price information, as well as assumptions concerning the structure of technology and the behavior of producers, but neither requires estimation of anything econometrically. Productivity change can also be calculated using the Divisia index, which is nonparametric. Finally, it can be estimated using econometric techniques.

A disadvantage of index number techniques and the Divisia index is that they do not provide sources of productivity growth, whereas nonparametric and econometric techniques do. Although nonparametric and econometric techniques are capable of measuring productivity change and its sources, only the econometric

\textsuperscript{1} Parts of this section are drawn from Chapter 8 of Kumbhakar and Lovell (2000).
The approach is capable of doing so in a stochastic environment. Here we show how to use econometric techniques to estimate the magnitude of productivity change, and then to decompose estimated productivity change into its various sources.

Productivity is affected in the presence of inefficiency. This is likely to affect productivity growth as well, unless inefficiency is time-invariant. Traditional econometric models of productivity change ignored the contribution of efficiency change to productivity change. Productivity change was decomposed into technical change and scale economies. However, if inefficiency exists, then efficiency change provides an independent contribution to productivity change. If efficiency change is omitted from the analysis, its omission leads to an erroneous allocation of productivity change to its sources. Accordingly, it is desirable to incorporate the possibility of efficiency into econometric models of productivity and productivity change.

To get a flavor of the issues involved, we begin with a quantity-based (primal) approach for the estimation and decomposition of productivity change. In doing so we allow for the possibility that production plans can be technically inefficient. The general structure of the primal approach is illustrated in Figure 2.1, in which a single input is used to produce a single output. Assume that for a producer at time $t$ the production plan is given by the input-output combination $(x^t, y^t)$ and the production frontier (the maximum possible output function given input quantities) is $f(x, t)$. Productivity for this input-output combination is defined by the ratio of output to input, viz., $y^t / x^t$, which can be easily measured from input and output data. Note that $y^t < f(x^t, t)$, which means that the production plan is technically inefficient. We define technical efficiency as $y^t / f(x^t, t)$, which is at most unity. Since the production frontier, $f(x, t)$, is not known,
the measurement of technical efficiency requires estimation of the frontier. Furthermore, productivity is reduced in the presence of technical inefficiency, i.e., \( \frac{y^t}{x^t} < f(x^t, t)/x^t \). Alternatively, the more efficient a firm is, the higher is its productivity, *ceteris paribus*. If we assume that the producer expands the production plan from \((x^t, y^t)\) to \((x^{t+1}, y^{t+1})\), and technical progress has occurred between periods \(t\) and \(t+1\), this would imply that \( f(x, t + 1) > f(x, t) \). Ignoring noise, it is clear that production is technically inefficient in both periods, since \( y^t < f(x^t, t) \) and \( y^{t+1} < f(x^{t+1}, t+1) \), and that technical efficiency has improved from period \(t\) to period \(t+1\), since \([y^t / f(x^t, t)] < [y^{t+1} / f(x^{t+1}, t+1)]\). It is also clear that productivity growth has occurred, since \((y^{t+1} / x^{t+1}) > (y^t / x^t)\). The estimated rate of productivity growth can then be decomposed into returns to scale, technical change and change in technical efficiency.

It is clear from the above discussion that in a single-output single-input case, productivity at a point in time is measured by \( \frac{y^t}{x^t} \), which is nothing more than the average product of \(x\) at time \(t\). Productivity change (treating time as a continuous variable) is measured by the rate of change in the average product of \(x\), i.e., \( \frac{\partial \ln \left( \frac{y^t}{x^t} \right)}{\partial t} = \dot{y} - \dot{x} \), where the dot over a variable indicates its rate of change. Thus both productivity and productivity change can be measured from observed data without estimating anything. If inefficiency is present, then the productivity of an input is lower than its maximum possible value. The degree to which productivity is lowered depends on the degree of inefficiency, which has to be estimated. Similarly, the effect of inefficiency on productivity change will depend on the temporal behavior of inefficiency.

If there are multiple inputs, one can obtain the partial factor productivity of each input \(j\), defined as \( \frac{y}{x_j} \), and the partial factor productivity change defined as \( \frac{\partial \ln \left( \frac{y}{x_j} \right)}{\partial t} = \dot{y} - \dot{x}_j \). Observing that both the partial factor productivity and its change (with or without technical inefficiency) depend on the usage of other factors, which are likely to differ among inputs, a single measure of productivity (i.e., total factor productivity, TFP) and productivity change is needed. In practice, labor productivity is often used as the measure of productivity, although its level depends on the amount of other factors being employed, such as capital, energy, and materials used, in the production process.

The objective of this paper is to survey some issues related to productivity measurement and the decomposition of TFP change in the context of a panel data framework. The focus is on the parametric econometric models based on production, distance, cost, and profit functions. TFP change is decomposed into technical change, scale economies, and technical efficiency components. Several alternative strategies in modeling technical change and production technology (in both primal and dual contexts) are considered.

We begin by using a production frontier approach to obtain estimates of productivity change, and to decompose estimated productivity change into a
technical change, a returns to scale, and a component associated with change in technical efficiency. We also consider the factor augmenting approach to technical change, which allows the researcher to decompose technical change into contributions due to each individual input. To allow for the possibility of multiple technologies and to measure the technology gap across regions, countries, etc., we discuss mixing (latent class) models and estimation of a best practice (meta) frontier. Furthermore, we use the input and output distance functions as an alternative to the production function approach. We repeat this investigation using a dual cost function approach, which has the advantage of accommodating multiple outputs. With multiple outputs, an additional component of productivity growth, viz., markup in output prices, is added. We also use a profit function approach with both single and multiple outputs. The profit function approach is extended to accommodate markups in output prices. Finally, we consider the alternative profit function in which output prices (instead of output quantities) are treated as endogenous.

2. THE PRODUCTION FUNCTION APPROACH

2.1. Time Trend Representation of Technical Change

We summarize the technology of the firm in time period $t$ by its period $t$ production function $f_t(\cdot)$. Assuming that there is one output, the production function can be represented as $y_{it} = f_t(x_{1it}, \ldots, x_{Jit})$, where $y_{it}$ is the output of the $i^{th}$ firm ($i = 1, \ldots, N$) in period $t$ ($t = 1, \ldots, T$), $f_t(\cdot)$ is the production technology, and $x$ is a vector of $J$ inputs. In order to estimate the parameters of such a production function econometrically, it is necessary to relate the production function in period $t$ to the corresponding production function for other periods. A common approach is to assume that the production function is atemporal, meaning that its form and the parameters do not depend on the time index $t$. Exploiting this fact and accommodating technical inefficiency, the production function can be represented by

$$y_{it} = f(x_{it}, t)\exp(-u_{it}),$$

where $u_{it} \geq 0$ is output-oriented technical inefficiency. Technical inefficiency, $u_{it}$, measures the proportion by which actual output ($y_{it}$) falls short of maximum possible output $f(x_{it}, t)$. Technical efficiency is then defined by $y_{it}/f(x_{it}, t) = \exp(-u_{it}) \leq 1$. The time trend variable $t$ in (1) represents technical change (a shift in the production function over time, ceteris paribus).
When input quantities change, productivity change is measured by what is popularly known as TFP change (or the Divisia index of productivity change, denoted by $\dot{TFP}$) and is defined as

$$\dot{TFP} = \dot{y} - \dot{x} \equiv \dot{y} - \sum_j S_j \dot{x}_j,$$  \hfill (2)\end{equation}

where $S_j = \frac{w_j x_j}{C^a}$ and $C^a = \sum_j w_j x_j$, $w_j$ being the price of input $x_j$. Here the rate of change in the composite input is defined as the weighted average of the rate of change on individual inputs. By using this definition, the TFP index can be constructed from

$$\dot{TFP}_t = \frac{\dot{TFP}_{t-1}(1 + \dot{TFP}_t)}{t}, \text{ for example, } TFP_1 \text{ for the base year } (t = 1) \text{ is } 100, \text{ i.e., } TFP_1 = 100.$$

To decompose $\dot{TFP}$ we differentiate (1) totally and use the definition of TFP change in (2) to obtain

$$\dot{TFP} = TC - \frac{\partial u}{\partial t} + \sum_j \left( \frac{f_j x_j}{f} - S_j \right) \dot{x}_j$$

$$\hfill (3)$$

where $TC = \frac{\partial \ln(f(x,t))}{\partial t}$, $TEC = \frac{\partial u}{\partial t}$ and $RTS = \sum_j \frac{\partial \ln f_x}{\partial \ln x_j} = \sum_j \frac{\partial \ln f_x}{\partial \ln x_j}$. $\dot{x}_j$ is the measure of returns to scale. Finally, $\lambda_j = \{f_j x_j / \sum_k f_k x_k = 1\}$ is the marginal product of input $x_j$. The relationship in (3) decomposes TFP change into scale ($\text{(RTS - 1)} \sum_j \lambda_j \dot{x}_j$), technical change ($\frac{\partial \ln(f_x,t)}{\partial t}$), technical efficiency change ($\frac{\partial u}{\partial t}$), and price effect ($\sum_j \{\lambda_j - S_j\} \dot{x}_j$) components. This last component captures either deviations of input prices from the value of their marginal products, i.e., $w_j \neq pf_j$, or the departure of the marginal rate of technical substitution from the ratio of input prices ($f_j/f_k \neq w_j/w_k$). Thus, the last component can be dropped from the analysis if one assumes that firms are allocatively efficient.\footnote{TFP growth is computed using the Divisia index. Therefore, the sum of the TFP growth components obtained from the parametric model (for example, from (3)) has to be equal to the TFP growth obtained using the Divisia index. But in practice, a wide gap between the two measures is often observed, especially when TFP growth (which is the sum of the TFP growth components) is estimated using a parametric model. This problem can be avoided by using the TFP growth equation as an additional equation in the system. See Kumbhakar and Vivas (2004a) for details on this issue in the context of a dual cost function estimation.}

If technical inefficiency is time-invariant (i.e., $-\frac{\partial u}{\partial t} = 0$), the decomposition in (3) shows that TEC does not affect TFP change. Under the assumption of

\footnote{Subscripts $i$ and $t$ are omitted to avoid notational clutter.}
constant returns to scale, the TFP change formula in (3) is identical to the one derived in Nishimizu and Page (1982), viz.:

$$\dot{T\text{FP}} = TC - \frac{\partial u}{\partial t} + \sum_j \left[ e_j - S_j^e \right] \dot{x}_j.$$  \hfill (4)

The TFP growth formula with input-oriented technical inefficiency (for which the production function is written as $$y = f(x \exp(-\tau), t), \ \tau \geq 0$$ being input-oriented technical inefficiency) is presented in Appendix A. Technical inefficiency can also be non-neutral. The TFP growth formulae for non-neutral (input- and output-oriented) technical inefficiency are presented in Appendices B and C.

2.2. The Factor-Augmenting Model of Technical Change

Since technical progress is defined in terms of shifts in the production function or the isoquants, it means that either a given level of output can be produced with fewer inputs, or more output can be produced using the same amount of inputs. This, in turn, implies that technical progress increases the productivity of at least one input. Does technical progress increase productivity of all or some of the inputs? Is it possible that the productivity of some inputs increases while those of other inputs remain constant or decline? What is the contribution of a particular input to overall technical change? The standard time trend model (discussed in Section 2.1) gives an estimate of the overall effect of technical change on output but it cannot answer the above questions. Answers to these questions can be obtained by specifying technical change in factor augmenting (FA) form (Beckmann and Sato (1969), Sato and Beckmann (1969), Kumbhakar (2002, 2003)), viz.:

$$y = f(Ax) \exp(-u) = f(A_1(t)x_1, \ldots, A_J(t)x_J) \exp(-u)$$

$$\equiv f(\tilde{x}_1, \ldots, \tilde{x}_J) \exp(-u) = f(\tilde{x}) \exp(-u),$$  \hfill (5)

where $$\tilde{x}_j = A_j(t)x_j$$ is the $$j$$th variable input measured in efficiency units, and $$f(\cdot)$$ is the production technology. $$A_j(t) > 0$$ is the efficiency factor associated with input $$j \ (j = 1, \ldots, J).$$ It can also be viewed as an input-specific productivity/efficiency index. If $$A_j(t)$$ increases over time, then the productivity of input $$j$$ rises. Thus, $$A_j(t) - A_j(t-1)$$ measures the productivity change of input $$j$$ from period $$(t-1)$$ to period $$t$$. Consequently, productivity growth in input

---

4 Although we are making $$A(\cdot)$$ a function of time, in principle, it can depend on other exogenous variables.
j (i.e., \( A_j(t) - A_j(t-1) > 0 \)) implies an increase in partial factor productivity \( (y/x_j) \) but not vice-versa.

Using the same definition of technical change as before, \( TC \) in the FA model can be expressed as

\[
TC_p = \sum_j \frac{\partial \ln f(\tilde{x})}{\partial \ln \tilde{x}_j} \cdot \frac{\partial \ln \tilde{x}_j}{\partial t} = \sum_j \frac{\partial \ln f(\tilde{x})}{\partial \ln \tilde{x}_j} \cdot \dot{A}_j \equiv \sum_j TC_p^j. \tag{6}
\]

\( TC_p^j \) represents the contribution of the \( j \)th input to the aggregate (overall) primal technical change \( TC_p \). It is clear from (6) that \( TC_p^j \) depends on the rate of change of input productivity \( \dot{A}_j \) and \( \frac{\partial \ln f(\tilde{x})}{\partial \ln \tilde{x}_j} \), which under competitive market conditions, is the cost share of input \( j \) in total revenue.

The essential difference between the specifications of the production technology in (1) and (5) is that technical progress in (1) shifts the production function over time, \textit{ceteris paribus}, whereas in (5) it enhances input quantities in efficiency units \( (\tilde{x}) \), thereby causing a movement along the production function \( y = f(\tilde{x}) \). Such a movement may be caused by factors other than time.

The input-specific productivity indices \( (A_j) \) can be functions of time as well as variable and quasi-fixed inputs.\(^5\) Thus, depending on the specification of \( A_j \), technical change can be purely exogenous (functions of only time), purely endogenous (functions of choice/decision variables), or a mixture of the two. However, in the context of estimating a single equation production function, endogeneticity of inputs is typically not taken into account. In such a case the distinction between endogenous and exogenous technical change is not clear.

TFP growth in this setup is

\[
\dot{TFP} = (RTS - 1) \sum_j \lambda_j \tilde{x}_j + TC_p + \sum_j (A_j - S^o_j) \tilde{x}_j - \frac{\partial u}{\partial t}, \tag{7}
\]

To examine these components in detail, we assume a translog functional form to represent the underlying production technology, viz.:

\[
\ln y = \alpha_0 + \sum_j \alpha_j \ln \tilde{x}_j + \frac{1}{2} \sum_j \sum_k \alpha_{jk} \ln \tilde{x}_j \ln \tilde{x}_k - u, \tag{8}
\]

where \( \tilde{x}_j = A_j(t) x_j \). It is necessary to specify \( A_j(t) \) in order to estimate the above model. To illustrate this, we consider the following two forms for \( A_j(t) \). First, we specify \( A_j(t) \) as a function of time with the following parameterization

\[
\ln A_j(t) = a_j t + b_j t^2 \tag{9}
\]

\(^5\) Variables such as R&D expenditure that is considered a choice variable in the dynamic model also affects technical change. Technical change caused by choice variables is often labeled endogenous technical change.
where \( a_j \) and \( b_j \) are parameters which are to be estimated along with the parameters of the production function. In the second specification, the \( A_j \)'s are functions of time as well as other \( x \) variables, viz.:

\[
\ln A_j = t \left( a_j + \sum_k b_{jk} \ln x_k \right) \tag{10}
\]

where \( a_j \) and \( b_{jk} \) are parameters to be estimated.

From the above specifications one can easily test whether the rate of change in efficiency factors are constant or not, by restricting \( b_j = 0 \) \( \forall j \) in (9) and \( b_{jk} = 0 \) in (10). Similarly, the hypothesis of no change in productivity of inputs can be tested from the restrictions \( a_j = b_j = 0 \) \( \forall j \) in (9) and \( a_j = b_{jk} = 0 \) \( \forall j,k \) in (10). Specification (10) assumes that productivity of an input at a point in time depends not only on time but also quantities of inputs being used at that point of time. That is, the productivity gain associated with an input is not purely exogenous. For example, productivity of labor in an environment with computers (higher level of capital) might be higher than a similar person working in an environment with fewer computers. Thus, labor productivity might depend on the stock of capital, and efficiency of capital might depend on the quantity of labor. Using this specification we can test whether one input affects the productivity of another input (i.e., whether some inputs are complementary to others from an efficiency point of view).

The measure of overall technical change \( (TC_p) \) using the above production function is

\[
TC_p = \frac{d \ln y}{dt} = \sum_j R_j \dot{A}_j \equiv \sum_j TC^j_p, \tag{11}
\]

where

\[
R_j = \frac{\partial \ln y}{\partial \ln \tilde{x}_j} = \alpha_j + \sum_k \alpha_{jk} \ln \tilde{x}_j, \tag{12}
\]

\[
TC^j_p = R_j \dot{A}_j(t). \tag{13}
\]

2.3. Latent Class Model

So far it has been assumed that there is a unique technology employed by every firm. However, firms in particular industries may use different technologies. In such a case, estimating a single frontier function encompassing every sample observation may not be appropriate in the sense that the estimated technology is
not likely to represent the “true” technology. That is, the estimate of the underlying technology may be biased. Furthermore, if the unobserved technological differences are not taken into account during estimation, the effects of these omitted unobserved technological differences might be inappropriately labeled as inefficiency.

To reduce the likelihood of such misspecification, the sample observations are often classified into certain categories using exogenous sample separation information, and a separate technology is estimated for each group. For example, Mester (1993) and Grifell and Lovell (1997) grouped banks into private and savings banks. Kolari and Zardkoohi (1995) estimated separate costs functions for banks grouped in terms of their output mix. Mester (1997) grouped sample banks in terms of their location. In the above studies, estimation of the technology using a sample of firms is carried out in two stages. First, the sample observations are classified into several groups. This classification is based on either some \textit{a priori} sample separation information (e.g., ownership of firms (private, public and foreign), location of firms, etc.) or applying cluster analysis to variables such as output and input ratios. In the second stage, separate analyses are carried out for each class/sub-sample.

To exploit the information contained in the data more efficiently, a latent class model approach (hereafter LCM) is often used, in which technological heterogeneity can easily be incorporated by estimating a mixture of production functions.\footnote{Finite mixture/LCMs are widely used in marketing, biology, medicine, sociology, psychology, and many other disciplines. For applications in social sciences, see, Hagenaars, J.A. and McCutcheon (2002). Statistical aspects of the mixing models are discussed in detail in McLachlan and Peel (2000).} In the standard finite mixture model, the proportion of firms (observations) in a group is assumed to be fixed (a parameter to be estimated), see, for example, Beard and Gropper (1991) and Caudill (2003). However, the probability of a firm using a particular technology can be explained by some covariates and may vary over time, and these technologies along with the (prior) probability of using them (that might vary across firms) can be estimated simultaneously. Once the technologies are estimated, one can define the best practice technology (metafrontier) by taking the outer envelope of the individual technologies (see Battese \textit{et al.} (2004)). The advantage of defining the best practice technology is that it can be compared to the individual technologies to measure technology gaps among countries, industries, firms, etc.

Assume that there are $j$ technologies that can be represented by the density functions $f_j(v_j), j = 1, \ldots, J$. Each producer in the sample uses one of these technologies (that is each firm at a point in time belongs to a particular class). The analyst does not know who is using what technology. Sometimes the analyst
follows some *ad hoc* criteria such as the size of firm, region in which the firm is located, etc., to group the sample firms. In the LCM one assumes that each firm has a probability of belonging to any group. Reintroducing the firm subscript $i$, these probabilities ($\pi_{ij} \geq 0$, $j = 1, \ldots, J \forall i$ and $\Sigma_j \pi_{ij} = 1 \forall i$) can be either constants or functions of covariates. Thus, in a LCM the technology for the $j^{th}$ class is specified as

\[
\ln y_i = \ln f(x_i, t)|_j + v_i|_j
\]

where $j = 1, \ldots, J$ stand for class. For each class, the stochastic nature of the frontier is modeled by adding a two-sided random error term $v_i|_j$. The noise term for class $j$ is assumed to follow a normal distribution with mean zero and constant variance ($\sigma^2_{vj}$). Thus, the conditional likelihood function for firm $i$ given that it belongs to class $j$ is

\[
f(i|j) = \phi \left[ \frac{v_i|_j}{\sigma_{vj}} \right]
\]

where $\phi(\cdot)$ is the pdf of a standard normal variable. Consequently, the unconditional likelihood function for firm $i$ is

\[
l(i) = \sum_j \pi_i \phi \left[ \frac{v_i|_j}{\sigma_{vj}} \right]
\]

where $\pi_{ij}$ is the prior probability assigned by the analyst on firm $i$ to be using the technology of type $j$. Sometimes firm-specific covariates are used to explain

![Fig. 2.2. The Metafrontier](image-url)
the prior probabilities. To impose the constraint that these probabilities are non-
negative and sum to unity, the multinomial form is often used. That is

$$\pi_{ij} = \frac{\exp(z'_j \eta_j)}{\sum_j \exp(z'_j \eta_j)} \cdot \eta_j = 0$$  \hspace{1cm} (17)

where $z_j$ are covariates explaining the prior probabilities and $\eta_j$ are the associated parameters.

The log likelihood function is then

$$\log L = \sum_{i=1}^{n} \log l(i)$$  \hspace{1cm} (18)

which is maximized with respect to $\beta_j$ (which is the vector of parameters associated with the class $j$ technology), $\eta_j$ and $\sigma_{ij}$ ($j = 1, \ldots, J$). Maximization of this likelihood function can be done with conventional gradient methods. The main problem is the possibility of multiple optima. Thus, one should use different starting values to make sure that the global maximum is achieved. The EM algorithm is a particularly useful device in this setting.\footnote{See Dempster, Laird and Rubin (1977) and McLachlan and Peel (2000).}

Furthermore, the EM algorithm is simple and intuitive (see Caudill (2003) for details). The above approach can easily be extended to incorporate technological inefficiency (Caudill (2003), Orea and Kumbhakar (2004), Kumbhakar and Tsionas (2003)).

Given that there are $J$ groups (industries, regions, countries, etc.), the LCM estimates $J$ different technologies.\footnote{In estimating a latent class model one has to address the problem of determining the number of groups/classes. The AIC and BIC (Schwartz’s criterion) are the most widely used criterion in standard latent class models to determine the appropriate number of classes. Both statistics measure the model’s goodness of fit but penalize by the complexity (number of parameters) of the model. Hence, they can be used to compare models with different numbers of classes. The best model is the one with lowest AIC or highest BIC.}

The highest posterior probability can be used to assign a group for each producer, i.e., firm $i$ is classified into group $k$ ($= 1, \ldots, J$) if $P(k|i) = \max_j P(j|i)$. The technology of class $k$ is then used as the reference technology to estimate technical efficiency of firm $i$. The outer envelope of these group technologies (frontier
for each group) is used to define the metafrontier (best practice technology), 
\( f(x) = \max_j \{ f(x) | j \} \). Finally, the group technologies can be compared with 
the best practice technology to obtain measures of technology gap, defined as 
the ratio of \( f(x)_j \) to \( f(x) \).

### 2.4. Multiple Outputs

One can express the multiple output production technology either in input or output augmenting form, viz., (i) \( F(y, \tilde{x}) = a \) and (ii) \( F(\tilde{y}, x) = a \), where \( \tilde{y} = D(t)y \) is a vector of \( M \) outputs and \( D(t) \) is the corresponding vector of efficiency factors while \( a \) is a constant. These models can be further extended to accommodate technical inefficiency by attaching the one-sided inefficiency term to either inputs or outputs. One way to interpret the formulation of technical change in (i) is that technical progress shifts the isoquants inward to produce a given level of output (thereby meaning that less inputs are needed) because input quantities in efficiency units are higher (\( A_j(t) > 0 \)). In formulation (ii) technical progress shifts the production possibility frontier (PPF) outward, thereby meaning that more outputs are produced (which implies that \( D_m < 0 \), given the input quantities.

If the specification in (ii) is chosen, then the production function can be written as

\[
y_1 = f(D_2 y_2, \ldots, D_M y_M, x_1, \ldots, x_J) \exp(-u) \tag{19}
\]

where output 1 plays an asymmetric role. This is a major disadvantage because estimation of the production technology is not invariant to the choice of output 1. Furthermore, there is the endogeneticity problem. If output 1 is endogenous, why not output 2, etc? Because of these problems, multiple output production functions are not estimated econometrically. These functions will be considered again in the dual set-up where behavioral assumptions (cost minimization and profit maximization behaviors) are introduced explicitly into the model.

A convenient way of modeling multiple outputs in a primal framework is to use distance functions. In this framework the homogeneity property (degree one) is used to solve the asymmetry problem that was encountered in the multiple output production functions. Both input and output distance functions are specified as \( F(x, y) = a \). The homogeneity property separates one from the other. The input (output) distance function is homogeneous of degree one in inputs (outputs). We now discuss the distance function approach.
3. DISTANCE FUNCTIONS

When many inputs are used to produce many outputs, Shephard’s (1953, 1970) distance functions provide a functional characterization of the structure of the production technology. Input distance functions characterize input sets, and output distance functions characterize output sets. Not only do distance functions characterize the structure of production technology, but also they are intimately related to the measurement of technical efficiency.

An input distance function is a function \( D_I(y, x, \lambda) = \max \{ \lambda : x/\lambda \in L(y) \} \), where \( L(y) \) describes the sets of input vectors that are feasible for each output vector. It adopts an input-saving approach to the measurement of the distance from a producer to the boundary of production possibilities. It gives the maximum amount by which an input vector can be radially contracted while still being able to produce the same output vector. In Figure 2.3, the scalar input \( x \) is feasible for output \( y \), but \( y \) can be produced with smaller input \((x/\lambda^*)\), and so \( D_I(y, x, \lambda) = \lambda^* > 1 \).

An output distance function is a function \( D_O(x, y, \mu) = \min \{ \mu : y/\mu \in P(x) \} \), where \( P(x) \) describes the sets of output vectors that are feasible for each input vector \( x \). It takes an output-augmenting approach to the measurement of the distance from an observed input bundle to the boundary of production possibilities. It gives the minimum amount by which an output vector can be deflated and still remain producible for the same vector of inputs. In Figure 2.4 scalar output \( y \) can be produced with input \( x \), but so can larger output \((y/\mu^*)\), and so \( D_O(y, x, \mu) = \mu^* < 1 \).

Fig. 2.3. An Input Distance Function (\( N = 2 \))
Fare and Primont (1995) discussed properties of both output and input distance functions. They show how to derive returns to scale (RTS) using these distance functions. These results are used next in the derivation of the TFP growth formulae.

### 3.1. Output Distance Function

In the presence of technical inefficiency, the output distance function is written as, $D_O(x, y, t) \leq 1 \Rightarrow D_O(x, y, t) \exp(u) = 1$, where $u \geq 0$ is output-oriented technical inefficiency. Differentiating totally, we get

$$
\sum_m \frac{\partial \ln D_O(x, y, t)}{\partial \ln y_m} \dot{y}_m + \sum_j \frac{\partial \ln D_O(x, y, t)}{\partial \ln x_j} \dot{x}_j + \frac{\partial \ln D_O(x, y, t)}{\partial t} + \frac{\partial u}{\partial t} = 0
$$

$$
\Rightarrow \sum_m R_m \dot{y}_m - \sum_j S_j \dot{x}_j (RTS) + \partial \ln D_O/\partial t + \partial u/\partial t = 0
$$

$$
\Rightarrow \hat{T}\hat{FP} = (RTS - 1) \sum_j S_j \dot{x}_j - \partial \ln D_O/\partial t - \partial u/\partial t
$$

(20)

where

$$
\hat{T}\hat{FP} = \sum_m R_m \dot{y}_m - \sum_j S_j \dot{x}_j, \quad RTS = - \sum_{j=1}^J \frac{\partial \ln D_O(x, y, t)}{\partial \ln x_j},
$$

$$
S_j = \frac{w_j x_j}{\sum_j w_j x_j} = \frac{\partial \ln D_O/\partial \ln x_j}{\sum_j \partial \ln D_O/\partial \ln x_j}, \quad R_m = \frac{p_m y_m}{\sum_m p_m y_m} = \frac{\partial \ln D_O/\partial \ln y_m}{\sum_m \partial \ln D_O/\partial \ln y_m} = \frac{\partial \ln D_O}{\partial \ln y_m}
$$
since $\sum_m \frac{\partial \ln D_O}{\partial \ln y_m} = 1$ (which follows from the homogeneity property of $D_O$). To estimate the above model, we use the homogeneity property and write the output distance functions as $D_O(x, y, t) = D_O(x, \tilde{y}, t)$ where $\tilde{y} = (y_2/y_1, \ldots, y_J/y_1)$. Finally, we rewrite it as $\ln D_O(x, y, t) - \ln y_1 = \ln D_O(x, \tilde{y}, t) \Rightarrow -\ln y_1 = \ln D_O(x, \tilde{y}, t) + u$ where $\ln D_O = -u \leq 0$. After this, a parametric function is assumed for $\ln D_O(x, \tilde{y}, t)$ and a stochastic noise term is added prior to estimation. The standard frontier technique of maximum likelihood method can then be used to estimate the output distance function.

### 3.2. Input Distance Function

In the presence of technical inefficiency the input distance function is written as, $D_I(x, y, t) \geq 1 \Rightarrow D_O(x, y, t) \exp(-\tau) = 1$, where $\tau \geq 0$ is input-oriented technical inefficiency. Total differentiation of the input distance function yields

$$\sum_m \frac{\partial \ln D_I(x, y, t)}{\partial \ln y_m} \dot{y}_m + \sum_j \frac{\partial \ln D_I(x, y, t)}{\partial \ln x_j} \dot{x}_j + \frac{\partial \ln D_I(x, y, t)}{\partial t} - \frac{\partial \tau}{\partial t} = 0$$

$$\Rightarrow (-\text{RTS}^{-1}) \sum_m R_m \dot{y}_m + \sum_j S_j \dot{x}_j + \partial \ln D_I/\partial t - \partial \tau/\partial t = 0$$

$$\Rightarrow TFP = (1 - \text{RTS}^{-1}) \sum_m R_m \dot{y}_m + \partial \ln D_I/\partial t - \partial \tau/\partial t$$

(21)

where

$$TFP = \sum_m R_m \dot{y}_m - \sum_j S_j \dot{x}_j, \text{ RTS}^{-1} = -\sum_{m=1}^{M} \frac{\partial \ln D_I(x, y, t)}{\partial \ln y_m}$$

$$R_m = \frac{p_m y_m}{\sum_m p_m y_m} = \frac{\partial \ln D_I/\partial \ln y_m}{\sum_m \partial \ln D_I/\partial \ln y_m}, \quad S_j = \frac{w_j x_j}{\sum_j w_j x_j} = \frac{\partial \ln D_I/\partial \ln x_j}{\partial \ln x_j}$$

since $\sum_j \partial \ln D_I/\partial \ln x_j = 1$ (which follows from the homogeneity property of $D_I$). Using the same property, we can express $D_I(x, y, t)$ as $D_I(x, y, t)/x_1 = D_I(\tilde{x}, y, t)$ where $\tilde{x} = (x_2/x_1, \ldots, x_J/x_1)$. Thus, $\ln D_I - \ln x_1 = \ln D_I(\tilde{x}, y, t)$ which in turn implies that $-\ln x_1 = \ln D_I(\tilde{x}, y, t) - \tau$ where $\ln D_I = \tau \geq 0$. Therefore, the second component in (21) (i.e., $\partial \ln D_I(x, y, t)/\partial \tau$) encapsulates technical change, where $\partial \ln D_I(x, y, t)/\partial \tau = \partial \ln D_I(\tilde{x}, y, t)/\partial \tau \geq 0$ ($\leq 0$) implies technical progress (regress). The last component of (21) (i.e., $-\partial \tau/\partial t$) represents the effects of technical efficiency change. Once $\tau$ is estimated from the

---

9 See Brummer et al. (2002) for details on the decomposition.
model $-\ln x_1 = \ln D_j(\bar{x}, y, t) - \tau$, after a parametric function is assumed for $\ln D_j(\bar{x}, y, t)$ and a stochastic noise term is added, $\partial \tau / \partial t$ can be easily estimated. Thus, by estimating an input distance function, all three components of TFP growth can be obtained. For example, if a translog form is assumed for $\ln D_j(\bar{x}, y, t)$, viz.:

$$
-\ln x_1 = \alpha_0 + \sum_{j=2}^{J} \alpha_j \ln(x_j / x_i) + \sum_{m=1}^{M} \beta_m \ln y_m + \frac{1}{2} \sum_{j=2}^{J} \sum_{h=1}^{j-1} \alpha_{jh} \ln(x_j / x_i) \ln(x_h / x_i) \\
+ \frac{1}{2} \sum_{m=1}^{M} \sum_{l=1}^{M} \beta_{ml} \ln y_m \ln y_l + \sum_{j=1}^{J} \sum_{k=1}^{J} \gamma_{jk} \ln(x_j / x_i) \ln y_k \\
+ \psi_0 t + \frac{1}{2} \psi_0 t^2 + \sum_{j=2}^{J} \xi_{jt} \ln(x_j / x_i) + \sum_{m=1}^{M} \xi_{mt} \ln y_m + v - \tau
$$

(22)

Standard frontier techniques can be used to estimate the above model from which $RTS, R_m, S_j$ and $\partial \tau / \partial t$ can be computed. These are then used to obtain the components of TFP growth in (21).10

4. THE COST FUNCTION APPROACH11

4.1. Single Output

While modeling inefficiency in a cost model, we can argue that inefficiency increases cost and therefore the cost function can be written as $C_a = C^0 \times CE$ where $C_a$ is the observed cost, $C^0$ is the minimum (frontier) cost, and $1/CE \leq 1$ is cost efficiency. While this argument is true for input-oriented technical inefficiency, it is worthwhile to derive the above result from the firm’s optimization problem, which is

$$
\min \ w'x \text{ subject to } y = f(x e^{-\tau}, t)
$$

where $\tau \geq 0$ is the input-oriented technical inefficiency. Assuming that firms are allocatively efficient, $f_j / f_i = w_j / w_i, \ j = 2, \ldots, J$, where $f_j$ is the marginal product (MP) of input $j$. The solution of the above problem gives inefficiency

10 See Karagiannis et al. (2004) for details.
11 See Kumbhakar (2000) for details on this model with allocative inefficiency.
adjusted input demand functions, $x_j(t)e^{-\tau}$ as a function of $(w, y, t)$. These input
demand functions can be used to define actual (observed) cost, $C^a$, which is

$$C^a = \sum_j w_jx_j \Rightarrow C^a e^{-\tau} = \sum_j w_jx_j e^{-\tau} = C^0(x, y, t)$$ (23)

where $C^0$ (which is the minimum cost function without technical inefficiency).

The above cost function can be written as

$$\ln C^a = \ln C^0(w, y, t) + \tau,$$ (24)

where, $\tau \geq 0$ can be interpreted as the percentage increase in cost due to technical
inefficiency.\(^{12}\)

Following Denny and Waverman (1981), it can be shown that the TFP growth
formula in a cost minimizing set up is

$$T\dot{FP} \equiv \dot{y} - \sum_j S^w_j \dot{x}_j = \dot{y} \left(1 - \frac{\partial \ln C}{\partial \ln y}\right) - \dot{C}_t = \frac{\partial \tau}{\partial t}$$

$$= (1 - 1/RTS)\dot{y} - \dot{C}_t - \frac{\partial \tau}{\partial t},$$ (25)

which decomposes TFP growth into scale, $TC$, and $TEC$ components.

The cost function, which is dual to the factor augmenting production function,
augmented to accommodate input-oriented technical inefficiency ($\tau$) can be
written as

$$C^a = C(\bar{w}, y) \exp(\tau),$$ (26)

where $\bar{w}_j = B_j(t)w_j$. If $A_j$ depends only on time, then $B_j(t) = 1/A_j(t) \forall j$.

Thus, an increase in efficiency of an input is equivalent to a decrease in its
effective price ($\bar{w}$). The overall technical change in a cost model is expressed
as $TC_c = -\frac{\partial \ln C}{\partial \ln y} = -\sum_j \frac{\partial \ln C_j}{\partial \ln w_j} \dot{B}_j(t)$, which can be expressed as $TC_c = \Sigma_j TC^j_c$,
where $TC^j_c = -\frac{\partial \ln C_j}{\partial \ln w_j} \dot{B}_j(t)$. Thus, $TC_c$ is a weighted average of input productivity
change ($A_j(t) = -\dot{B}_j(t)$), where the weights are the cost shares ($S^w_j$).

Thus, the TFP growth formula can be written as

$$T\dot{FP} \equiv \dot{y} - \sum_j S^w_j \dot{x}_j = \dot{y} \left(1 - \frac{\partial \ln C}{\partial \ln y}\right) + TC_c - \frac{\partial \tau}{\partial t}$$

$$= (1 - 1/RTS)\dot{y} + TC_c - \frac{\partial \tau}{\partial t}$$ (27)

\(^{12}\) See Kumbhakar (1996) and Kumbhakar and Wang (2005) for the derivation of the cost function
with output-oriented technical inefficiency.
The TFP growth formulae with output-oriented technical (along with and without allocativity) inefficiency are given in Appendix D. In Appendix E, we derive the TFP growth formulae with non-neutral output-oriented technical inefficiency.

4.2. Multiple Output Cost Function

If there are multiple outputs, TFP growth is defined as

$$\dot{T\bar{F}P} = \sum_m R_m \dot{y}_m - \sum_j S_j \dot{x}_j,$$

where $R_m = p_m y_m / \sum_l p_l y_l$ with $p_m$ representing the price of output $y_m (m = 1, \ldots, M)$. It can be shown (Denny et al. (1981)) that the components of TFP are

$$\dot{T\bar{F}P} = TC + (1 - RTS^{-1}) \dot{Y}_C + (\dot{Y}_P - \dot{Y}_C) - \frac{\partial \tau}{\partial t},$$

where $\dot{Y}_C = \sum_l \left( \frac{\varepsilon_{y_l}}{\sum_l \varepsilon_{y_l}} \right) \dot{y}_l + \sum_l p_l y_l$, and $\varepsilon_{y_l} = \partial \ln C / \partial \ln y_l$. TC and RTS in (28) can be obtained from a parametric cost function ($C = C(w, y, t)$) when RTS is defined as $RTS^{-1} = \sum_l \partial \ln C / \partial \ln y_l$. Thus, the first component of TFP growth is TC and the second component is the Scale component (related to RTS), which is zero if RTS is unity. The third component is non-zero if output markets are non-competitive. That is, if output prices depart from their respective marginal costs then $\dot{Y}_P \neq \dot{Y}_C$. Finally, the last component is due to technical efficiency change ($TEC = -\frac{\partial \tau}{\partial t}$). It is positive (negative) if efficiency improves (deteriorates), i.e., $\tau$ declines (increases) over time.

4.3. The Factor-Augmenting Model of Technical Change

Here we consider the multiple output production technology $F(y, \bar{x}) = a$ with input-oriented technical inefficiency, for which the cost function is $C^a = C(w, y, \exp(\tau))$. Following the procedure for the time trend model, the TFP growth formula for the FA model can be written as

$$\dot{T\bar{F}P} = TC_c + (1 - RTS^{-1}) \dot{Y}_C + (\dot{Y}_P - \dot{Y}_C) - \frac{\partial \tau}{\partial t},$$

where as before $TC_c = -\frac{\partial \ln C}{\partial t} = - \sum_j \frac{\partial \ln C(J)}{\partial \ln w_j} \dot{B}_j(t)$. 

To pursue this decomposition in detail, we assume a translog form for \( C(\tilde{w}, y) \) and write it as

\[
\ln C_a = \alpha_0 + \sum_j \alpha_j \ln \tilde{w}_j + \sum_m \beta_m \ln y_m + \sum_{j m} \alpha_{jm} \ln \tilde{w}_j \ln y_m
\]

\[
+ \frac{1}{2} \left\{ \sum_{j k} \alpha_{jk} \ln \tilde{w}_j \ln \tilde{w}_k + \sum_{l m} \beta_{lm} \ln y_l \ln y_m \right\} + \tau, \quad (30)
\]

where \( \tilde{w}_j = B_j(t)w \). The cost function in (30) is assumed to satisfy symmetry and linear homogeneity (in \( \tilde{w} \)) restrictions. We specify the \( B_j \) as quadratic functions of time, i.e.:

\[
\ln B_j = t(a_j + b_j t)
\]

We then use formula \( TC_c = -\sum_j \frac{\partial \ln C}{\partial \ln \tilde{w}_j} B_j(t) \) to estimate technical change where

\[
\frac{\partial \ln C}{\partial \ln \tilde{w}_j} = \alpha_j + \sum_k \alpha_{jk} \ln \tilde{w}_k + \sum_m \beta_{jm} \ln y_m
\]

with the appropriate forms for \( \dot{B}_j \) derived from \( \dot{\tilde{w}}_j \). Furthermore, technical change as defined above can be decomposed into pure, scale and input price components.

In banking applications, multiple outputs are almost always used. One common problem with multiple outputs is the presence of zero values for some outputs (meaning that some banks do not produce all the outputs). Zero outputs create problems for the Cobb Douglas and translog cost functions, since the log of zero is not defined. To avoid this problem, researchers often replace the zero values with a small positive number. Some replace them by unity so that log values are zero, while others add the small positive number to all observations. For example, if output \( m \) has zero values for some banks, \( \ln y_m \) is redefined as \( \ln y_m = \ln(y_m + c) \), where \( c \) is a positive number supplied by the user. This procedure, although widely used in practice, has two problems. First, zero values for output(s) for a bank might be due to the fact that the bank specializes in a few outputs, and adding small numbers for outputs that are never produced puts the specialized banks in the same group as others. In fact, this procedure does not recognize the fact that some banks can be specialized in certain outputs. Thus, no matter how small \( c \) is, this procedure is not innocuous. Second, the output values are changed (either from zero to a positive constant \( c \) or from \( y_m \) to \( (y_m + c) \)) without changing the cost. That is,
the cost of the extra output \( c \) is zero. Looking at it from a purely mathematical point of view, either \( C = C(w, y_1, \ldots, y_m, t) \) or \( C = C(w, y_1, \ldots, y_m, \sum_{m=1}^{M} (y_m + c), t) \) is true, but not both at the same time.

Replacing the zero values by unity makes sense if the technology of the banks that are specialized in fewer outputs is the same as those that produce all the outputs. For example, if the specialized banks produce \( Q \) outputs while the other banks produce \( M (M > Q) \) outputs and the production technology for both types of banks is Cobb-Douglas, i.e.:

\[
\ln C_i = \alpha_0 + \sum_{j=1}^{J}\alpha_j \ln w_{ji} + \sum_{m=1}^{Q}\beta_m \ln y_{mi} + v_i, \quad i = 1, \ldots, N_i
\]

\[
\ln C_i = \alpha_0 + \sum_{j=1}^{J}\alpha_j \ln w_{ji} + \sum_{m=1}^{M}\beta_m \ln y_{mi} + v_i, \quad i = N_i + 1, \ldots, N
\]

then we can simply combine them and write the technology as

\[
\ln C_i = \alpha_0 + \sum_{j=1}^{J}\alpha_j \ln w_{ji} + \sum_{m=1}^{M}\beta_m \ln y_{mi} + v_i, \quad i = 1, \ldots, N
\]

where \( y_{mi} = 1 \) for \( m = Q + 1, \ldots, M \) and \( i = 1, \ldots, N_i \). Since the hypothesis, that the technology for the specialized and non-specialized banks is the same, is testable, there is no point in making the assumption \textit{a priori}. Thus the problem is not as innocuous as it seems.

This zero value problem is not endemic to the multi-output cost functions. It applies equally to output and input distance functions, as well as the latent class models discussed before.

### 4.4. Latent Class Models

The existence of multiple technologies can easily be accommodated in each of the above cases. Orea and Kumbhakar (2004) used a multiple output cost function approach and found four separate technologies used by the Spanish banks. Their approach can easily be extended to construct the metacost function, which is the inner envelope of the group cost frontiers. Such a metacost function can then be used to measure technology gaps among banks using different technologies. The Orea-Kumbhakar approach also accommodates technical inefficiency that varies across banks and differs among technologies. Furthermore, these inefficiencies can vary over time in a flexible manner. Because of the similarity of the approach with the production function approach discussed earlier, no further detail is given.
5. THE PROFIT FUNCTION APPROACH

5.1. Single Output

When modeling inefficiency by using a profit model, we start with the argument that inefficiency increases cost and therefore the profit function can be written as \( \pi^u = \pi^0 \times PE \), where \( \pi^u \) is the observed profit, \( \pi^0 \) is the maximum (frontier) profit, and \( PE \leq 1 \) is profit efficiency, which is modeled as a one-sided error term. Furthermore, it is assumed to be independent of the regressors (input and output prices and time). Here it is shown that this assumption is generally untrue.

The profit function, in the presence of technical inefficiency corresponding to the production function in (1), can be written as

\[
\pi(w, pe^{-u}, t) = \max_{ye^u, x} \left\{ py - w'x \right\}
\]

which also equals the actual profit \( \pi^u = py - w'x \). It means that profit, when price is \( p \) and output equals \( y \), is the same as profit when output equals \( f(x, t) \) but price equals \( pe^{-u} \). That is, a 10% reduction in output given inputs has the same effect on profit as a 10% reduction in output price holding output constant.

Using \( p' \equiv pe^{-u} \), the profit function can be implicitly written as \( \pi(w, p', t) \). Then the Hotelling’s lemma, \( \partial \pi / \partial p' = ye^u \) and \( \partial \pi / \partial w_j = -x_j \) can be applied to obtain the input demand and technical inefficiency adjusted output supply functions. Note that the argument of the profit function is \( p' \), which in turn implies that technical inefficiency may not appear additively in the log profit function. For example, in the case of the translog profit function

\[
\ln \pi = \alpha_0 + \sum_j \alpha_j \ln w_j + \sum_j \beta_{pj} \ln p_j + \alpha_j t
\]

\[
+ \frac{1}{2} \left\{ \beta_{ppj} \ln p_j \ln p_j + \sum_j \sum_k \alpha_{jk} \ln w_j \ln w_k + \alpha_{ttj} t^2 \right\}
\]

\[
+ \sum_j \gamma_{pj} \ln w_j \ln p_j + \beta_{pj} \ln p_j t + \sum_j \delta_{pj} \ln w_j t,
\]

(31)
which can be rewritten as

\[
\ln \pi = \alpha_0 + \sum_j \alpha_j \ln w_j + \beta_p \ln p + \alpha_t t + \frac{1}{2} \left( \beta_{pp} \ln p \ln p + \sum_j \sum_k \alpha_{jk} \ln w_j \ln w_k + \alpha_{tt} t^2 \right) + \sum_j \gamma_{jm} \ln w_j \ln p + \beta_p p t + \sum_j \delta_j \ln w_j t - u \left( \beta_p + \sum_j \gamma_{jm} \ln w_j + \beta_{pt} t \right) + \frac{1}{2} \beta_{pp} u^2 \quad (32)
\]

\[
\equiv \ln \pi^0 - g(u, \ln p, \ln w, t)
\]

where \( \ln \pi^0 \) is the translog profit frontier, and \( g(u, \ln p, \ln w, t) = u(\beta_p + \sum_j \gamma_{jm} \ln w_j + \beta_{pt} t) - \frac{1}{2} \beta_{pp} u^2 \geq 0 \) is profit inefficiency (profit loss due to technical inefficiency). It is clear from the profit function above that \( g(u, \ln p, \ln w, t) \) is a non-linear function of \( u \) and it cannot assumed that \( g(u, \ln p, \ln w, t) \) is an independently and identically distributed random variable. As a result, the standard tools used to estimate the production frontier cannot be used here.\(^{13}\) A simple sign change is not enough to model technical inefficiency in the profit function, unless the underlying production function is homogeneous.\(^{14}\)

This will also affect the TFP growth formulation. Differentiating the profit function \( \pi = \pi(w, p', t) \) totally, we get

\[
\frac{d \ln \pi}{dt} = \frac{\partial \ln \pi}{\partial \ln p'} \dot{p'} + \sum_j \frac{\partial \ln \pi}{\partial \ln w_j} \dot{w}_j + \frac{\partial \ln \pi}{\partial t} = \frac{1}{\pi} \left( py \dot{p'} - C^a \sum_j S_j \dot{w}_j \right) + \frac{\partial \ln \pi}{\partial t},
\]

using \( \partial \pi / \partial p = y \) and \( \partial \pi / \partial w_j = -x_j \) (Hotelling’s lemma).

From the definition of profit \( \pi = py - \sum_j w_j x_j \), we get

\[
\frac{d \ln \pi}{dt} = \frac{py}{\pi} \left( \dot{y} + \dot{p} \right) - \frac{C^a}{\pi} \sum_j \left( \dot{S}_j \dot{x}_j + S_j \dot{w}_j \right)
\]

Equating the above two equations gives (after some algebraic manipulations),

\[
\pi \left( \frac{\partial \ln \pi}{\partial t} \right) = py \left( \dot{y} - \sum_j S_j \dot{x}_j \right) + py \sum_j S_j^a \dot{x}_j - C^a \sum_j S_j^a \dot{x}_j + py \frac{\partial \pi}{\partial t}
\]

\[
= py(TFP) - py(RTS') - \sum_j S_j^a \dot{x}_j + py \frac{\partial \pi}{\partial t}
\]

\(^{13}\) See Kumbhakar (2001) and Kumbhakar and Tsionas (2005) for details on estimation issues.

\(^{14}\) There are numerous banking papers that use translog profit functions as \( \ln \pi = \ln \pi^0 - u \) (for example, see Berger and Mester (2003) and the references cited in there), which is clearly inappropriate.
The above equation can be rearranged to give

\[
\dot{T}FP = \frac{\pi}{p_y} \left\{ \frac{\partial \ln \pi}{\partial t} \right\} + (RTS^i - 1) \sum_j S^j x_j - \{\partial u/\partial t\}. \tag{33}
\]

where \( RTS^i = 1 - \left\{\frac{\partial \ln \pi/\partial \ln p^i}{\partial \ln p^i}\right\}^{-1} \). Furthermore, it follows from the envelope theorem that \( \frac{\delta \pi}{\delta w} = p \frac{\partial \pi}{\partial \ln p} \), which in turn implies that \( \frac{\partial \pi}{\partial \ln p} = \frac{\partial \ln f}{\partial \ln w} = TC \), which is used in (3). There is one important difference between (3) and (33). In deriving (3) we did not rely on any behavioral assumptions explicitly, whereas in (33) we used the profit maximizing conditions to determine the allocation of inputs and the production of output. Consequently, input quantities in (33) (and therefore \( RTS \)) are affected by the presence of technical inefficiency.\(^\text{15}\)

Using the following results, the components of TFP growth can be expressed in terms of the profit function, viz.:

\[
\frac{p_y}{\pi^i} = \frac{\partial \ln \pi}{\partial \ln p^i},
\]

\[
\{RTS^i - 1\} = -\left\{\frac{\partial \ln \pi/\partial \ln p^i}\right\}^{-1}, \tag{34}
\]

\[
S^j = \left\{ \frac{\partial \ln \pi^j/\partial \ln w_j}{\partial \ln w_j} \right\} \left\{\frac{\partial \ln \pi}{\partial \ln w_j}\right\}^{-1} \frac{\partial \ln \pi}{\partial t} - \frac{\partial \ln w_j}{\partial t}.
\]

The problem with the above decomposition is that each component depends on technical inefficiency. This can be avoided by using the relation \( \ln \pi = \ln \pi^0 - g(u, \ln p, \ln w, t) \) in the above formula. In doing so, we can capture both direct and indirect effects (via input and output prices as well as time) of technical inefficiency on TFP growth. However, if we specify (erroneously) the translog profit function as \( \ln \pi = \ln \pi^0 - u \), then (i) the estimated parameters of the profit function are likely to be biased, and (ii) the indirect effect (via input and output prices and time) of technical inefficiency TFP growth would not be captured.

### 5.2. Multiple Output

As before, we define \( \dot{T}FP \) as \( \dot{T}FP = \Sigma_m R_m \dot{y}_m = \Sigma_j S^j \dot{x}_j \) when \( R_m = p_m y_m/R \) and \( R = \Sigma_m p_m y_m \). In the standard case, where firms are assumed to be both

\(^{15}\) See Kumbhakar (2000) for the TFP growth formula that takes into account both technical and allocative inefficiency. The decomposition without any inefficiency is given in Kumbhakar (2000b).
technically and allocatively efficient (so that the multiple output profit function is \( \pi = \pi(w, p, t) \)), the TFP change formula can be expressed as\(^\text{16}\)

\[
T\dot{FP} = \frac{\pi}{R} \frac{\partial \ln \pi}{\partial t} + (RTS - 1) \sum_j S_j \dot{x}_j,
\]

(35)

where \( RTS - 1 = -\left\{ \sum_m \frac{\partial \ln \pi}{\partial \ln p_m} \right\}^{-1} \).

To derive the TFP change in the present case, we denote \( p_s = e^{-u} \frac{1}{\prod_m p_m} \) and start from the profit function \( \pi = \pi(w, p_s, t) \), which is defined as

\[
\pi(w, p_s, t) = \sum_m p_m y_m - \sum_j w_j x_j = \sum_m p_m y_m e^u - \sum_j w_j x_j = \pi(w, p_s, t).
\]

Thus, the TFP change formula is

\[
T\dot{FP} = \frac{\pi_s}{R} \frac{\partial \ln \pi_s}{\partial t} - \frac{\partial u}{\partial t} + (RTS_s - 1) \sum_j S_j \dot{x}_j,
\]

(36)

in which we used the results \( \frac{\partial \pi}{\partial p_m} = y_m \) (Hotelling’s lemma) and \( \frac{\partial \pi}{\partial w_j} = -x_j \). All the components in (36) can be expressed in terms of profit. For example, \( R = \pi' \left[ \sum_m \frac{\partial \ln \pi'}{\partial \ln p_m} \right], \)

\[
RTS'_s - 1 = -\left\{ \sum_m \frac{\partial \ln \pi}{\partial \ln p_m} \right\}^{-1},
\]

which gives \( (RTS'_s - 1) = -\frac{\pi}{R} \). Using these results, coupled with those in (34) for \( S'_j \), everything in (36) can be expressed in terms of \( \pi \) and rates of change in input and output quantities.

The first term on the right-hand side of (36) can be expressed as

\[
\frac{\pi_s}{R} \frac{\partial \ln \pi_s}{\partial t} = \sum_m R_m \frac{\partial \ln y_m}{\partial t},
\]

(37)

where \( y_m(.) \) is the supply function of output \( y_m \). Since the expression in (37) is a weighted average (weights being the revenue share of each output) of rates of change of individual outputs (\( \frac{\partial \ln y_m}{\partial t} \)), holding everything else constant, it can be viewed as a measure of output technical change. Note that this measure of output technical change is different from profit technical change, defined as \( \frac{\partial \ln \pi'}{\partial t} \).

It can be seen from (36) that contribution of technical change and returns to scale depends on both technical inefficiency via \( p_s \). Since technical change, returns to scale, etc., are usually defined in terms of the profit frontier, it is possible to rewrite (36) in terms of the profit frontier so that the effects of technical change and returns to scale are directly observable.

\^\text{16} See Karagiannis (2000) for TFP growth decomposition with multiple outputs in a profit function framework with quasi-fixed inputs.
inefficiency on TFP change are separated. For example, in the translog case we can express actual profit as 
\[\ln \pi = \ln \pi^0 - g(u, \ln p, \ln w, t),\]
where \(\ln p\) and \(\ln w\) are vectors of output and input prices. This result can be used in the TFP growth formula to separate technical inefficiency effects in the TFP growth components.

5.3. Alternative Profit Functions

Use of the profit function approach is based on the assumption that prices are exogenous, and that producers seek to maximize profit (or variable profit) by selecting outputs and inputs under their control. One justification for exogeneticity of prices is that producers operate in competitive markets. If producers have some degree of monopoly power in their product markets, then demand would be exploited to determine output prices and quantities jointly, and only input prices would be exogenous.

Recently Humphrey and Pulley (1997), among others, have introduced the notion of an “alternative” profit frontier to bridge the gap between a cost frontier and a profit frontier. For example, Berger and Mester (1997) suggest that the alternative profit approach may be helpful when (i) there are substantial unmeasured differences in the quality of banking services; (ii) outputs are not completely variable; (iii) output markets are not perfectly competitive; and (iv) output prices are not accurately measured.

An alternative profit frontier is defined as
\[
\pi^A = \max_{p,x} \{p'y - w'x | F(x, y, t) = 0, g(p, y, x, t) = 0\}
\]

\[\Rightarrow \pi^A = \pi(y, w, t),\]

where the endogenous variables are \((p, x)\) and the exogenous variables are \((y, w, t)\). \(F(x, y, t) = 0\) is the production function (it can also be specified by an output distance function), and \(g(p, y, w, t) = 0\) represents what Humphrey and Pulley refer to as the producer’s “pricing opportunity set,” which captures the producer’s ability to transform exogenous \((y, w, t)\) into endogenous product prices \(p\). However, \(\pi^A(.)\) is not dual to the production function because it incorporates both the structure of production technology and the structure of the pricing opportunity set. Moreover, without specifying the properties satisfied by the function \(g(p, y, w, t) = 0\), it is not possible to specify the properties satisfied by \(\pi^A(.)\).

The main problem with the alternative profit function is that it has no theoretical foundation such as with the standard profit and/or cost functions. Consequently, the approach has no use other than measuring efficiency, which is also
problematic, as will be shown later. Since the whole idea is based on the pricing opportunity function \( g(p, y, w, t) = 0 \), it is worth exploring the issue further. First, the single output case is considered. The profit maximization problem is partitioned into two steps. In step 1 the cost is minimized, given the production function, to obtain the cost function, \( C^a = w'x = C(w, y, t) \). In step 2 the profit is maximized, viz., \( \max_p \{ py - C(w, y, t) \} [g(p, y, w, t) = 0] \). Note that \( g(p, y, w, t) = 0 \) can be expressed implicitly, as \( p = p(w, y, t) \), which is nothing but the inverse demand function. Thus, there is no need to solve the optimization problem in step 2 for \( p \). Consequently, there is no meaning to the alternative profit function because producers do not maximize profit to obtain \( p \). The so-called optimal price \( p \) is not related to the production technology (either the production or the cost function).

A similar result is obtained when multiple outputs are considered. The cost function from step 1 can still be written as \( C^a = w'x = C(w, y, t) \). The first-order conditions from step 2 are: \( p_m = \mu \frac{\partial g(\cdot)}{\partial p_m} \quad m = 1, \ldots, M \) where \( \mu \) is the Lagrange multiplier associated with the pricing opportunity function. Thus, the output prices can be solved from the above first-order conditions and the pricing opportunity function \( g(p, y, w, t) = 0 \), without any reference to the technology (production or cost functions). These solutions are, in implicit form, \( p_m = p_m(w, y, t) \). Thus, cost of production (namely, marginal cost) does not play any role in determining optimal output prices.

Now we examine the case with (input-oriented) technical inefficiency and write the single output production function as \( y = f(x \exp(-\tau), t) \), for which the cost function is \( C^a = w'x = C(w, y, t) \exp(\tau) \). The first-order conditions from the second step are exactly the same as before. Thus, the solution of \( p \) would not be affected by the presence of technical inefficiency. Consequently, revenue will not be unaffected by the presence of inefficiency. Improvement in technical inefficiency would not be transmitted to revenue through output prices. Using this optimal price \( p = p(w, y, t) \) in the definition of profit we get

\[
p^a = p(w, y, t)y - C(w, y, t) \exp(\tau) \neq p^A(w, y, t) \exp(-\tau) \quad (39)
\]

Thus, the above relationship cannot be expressed as \( \ln p^a = \ln p^0 - \tau \), as is common in the literature (e.g., Berger and Mester (2003), DeYoung and Hasan (1998), among others). That is, the observed profit is not necessarily reduced by the same proportion by which cost is increased. For example, if profit without inefficiency is \( \pi_0^a = p(w, y, t)y - C(w, y, t) = 100 - 80 = 20 \) and the profit with technical inefficiency is \( \pi_0^a = p(w, y, t)y - C(w, y, t) \exp(\tau) = 100 - 80 \cdot (1.10) = 12 \), profit is reduced by 40% although cost has increased by only 10%.

In the presence of multiple outputs, the cost function is \( C^a = w'x = C(w, y, t) \exp(\tau) \). Maximization of profit subject to \( g(p, y, w, t) = 0 \) gives the
solutions of output prices that can be implicitly expressed as $p_m = p_m(w, y, t)$.
Note that these prices are not affected by the presence of technical inefficiency.
Thus, given output quantities, revenue is not affected by the presence of inefficiency. Consequently, the argument that alternative profit function takes into account the effect of inefficiency on both cost and revenue does not hold. Furthermore, actual profit is

$$\pi^a = \sum_m p_m(w, y, t) y_m - C(w, y, t) \exp(\tau) \neq \pi^d(w, y, t) \exp(-\tau).$$  \hspace{1cm} (40)$$

That is, actual profit is not reduced by the same percent by which cost is increased. Therefore, the efficiency estimates based on the alternative profit function cannot be related to technical inefficiency based on the production technology. Since we cannot draw a meaningful interpretation of the inefficiency term in the model $\ln \pi^a = \ln \pi^d(w, y, t) - \tau$, and the whole idea of estimating the alternative profit function is to estimate profit inefficiency, the alternative profit function approach seems vacuous. However, this is not the case with the standard profit function, although care has to be taken in modeling profit inefficiency (as was shown in the previous section).

This brings us back to the standard profit function that can be extended to accommodate non-competitive behavior in the output markets. This issue is explored next (see Kumbhakar and Lozano-Vivas (2004) for details).

5.4. Modeling Markups in Variable Profit Functions

Assuming that the objective of the banks is to maximize profit, the bank’s optimization problem can be formulated with the following two steps. In step 1, a bank solves the following problem, given the output vector $y$, to determine the least cost (variable) input quantities, i.e.:

$$\min_w w'x \text{ subject to } F(y, x, z, t) = 0$$

where $z$ is the vector of $Q$ quasi-fixed inputs and output attributes. The solution to the above problem gives the conditional input demand functions $x_j = x_j(w, y, z, t)$ that are then substituted into the objective function to obtain the minimum cost function $C(w, y, z, t)$. In step 2, the bank’s problem is to maximize profit, i.e.:

$$\max_y \pi = p'y - C(w, y, z, t)$$
Productivity and Efficiency Measurement Using Parametric Econometric Methods

in which the choice variables are outputs. If the output markets are competitive, the first-order conditions (FOC) for profit maximization can be expressed as

$$ p_m = \frac{\partial C}{\partial y_m} \equiv MC_m, \ m = 1, \ldots, M, $$

where \( MC_m \) is the marginal cost associated with output \( y_m \). However, if the output markets are not competitive and the banks possess monopoly power, the FOC of the above problem become

$$ p_m^* = MC_m $$

where \( p_m^* = p_m \theta_m \), with \( \theta_m \geq 1 \) representing the markup factor. Thus, the presence of markups can be tested by restricting \( \theta_m = 1 \forall \ m \). In the presence of markups, the relevant output prices are \( p_m^* = p_m \theta_m \) and the output supply functions are

$$ y_m = y_m(w, p^*, z, t). $$

Consequently the variable profit function can be expressed as \( \pi^* = \pi^*(w, p^*, z, t) \) from which the input demand and output supply functions can be obtained by using Hotelling’s lemma, viz.:}

$$ \frac{\partial \pi^*}{\partial p^*} = y_m \text{ and } \frac{\partial \pi^*}{\partial w_j} = -x_j. $$

The profit function \( \pi^*(.) \) is often labeled as the shadow variable profit function. However, it should be noted that \( \pi^* \) is not observed and \( \pi^* \neq \pi^a \) where \( \pi^a \) is the actual profit, defined as \( \pi^a = \sum_m p_m y_m - \sum_j w_j x_j \). Invoking Hotelling’s lemma it can be shown that

$$ \pi^a = \pi^* \left\{ \sum_m SR_m^*/\theta_m + \sum_j SC_j^* \right\} \equiv \pi^a H, $$

where \( H = \left\{ \sum_m SR_m^*/\theta_m + \sum_j SC_j^* \right\} \), \( \frac{\partial \ln \pi^a}{\partial SR_m} = SR_m^* \), and \( \frac{\partial \ln \pi^a}{\partial SC}_j = -SC_j^* \). The shadow shares \( SC_m^* \) and \( SC_j^* \) are not observed. These shadow shares are related to the actual (observed) shares as follows:

$$ SR_m^a = \frac{p_m y_m}{\pi^a} = \frac{SR_m^*}{H} \left\{ \frac{1}{\theta_m} \right\} $$

$$ SC_j^a = \frac{w_j x_j}{\pi^a} = -\frac{SC_j^*}{H}. $$

The above model can be estimated using either the variable profit function in (42) along with the associated share equations in (43), or using the shadow share equations in (44).
equations strictly by themselves. Before proceeding we need to specify the behavior of the markup factors ($\theta_m$) over time. First, $\theta_m$ may be specified as

$$\theta_m = \exp(b_m + c_m t), \quad m = 1, \ldots, M$$

where $b_m$ and $c_m$ are parameters to be estimated. The exponential function is often chosen to guarantee $\theta_m \geq 0$. Depending on the signs of $b_m$ and $c_m$, $\theta_m$ can decrease or increase over time. We are primarily interested in testing whether the $\theta$'s approach unity over time.

The system of equations in (42) and (43) can be operational only when a parametric form of the shadow variable profit function $\ln \pi^*$ is assumed. To minimize a priori restrictions on the underlying production technology, we use a translog form of the shadow variable profit function, $\ln \pi^*$, and write it as

$$\ln \pi^* = \alpha_o + \sum_j \alpha_j \ln w_j + \sum_q \beta_q \ln z_q + \alpha_t t + \frac{1}{2} \left\{ \sum_{jk} \alpha_{jk} \ln w_j \ln w_k + \sum_{ql} \beta_{ql} \ln z_q \ln z_l + \sum_{mn} a_{mn} \ln p_m^* \ln p_n^* + \alpha_{tt} t^2 \right\} + \sum_j b_{jq} \ln w_j \ln z_q + \sum_m c_{jm} \ln w_j \ln z_m^* + \sum_q d_{mq} \ln p_m^* \ln z_q + \sum_j \alpha_j \ln w_j t + \sum_m a_{mt} \ln p_m^* t + \sum_q b_{qt} \ln z_q t$$

The above shadow profit function is assumed to satisfy the symmetry and convexity conditions. Furthermore, it is homogeneous of degree one in input and shadow output prices. The symmetry and homogeneity restrictions are imposed in estimating the parameters of the model. The symmetry restrictions on (45) are

$$\alpha_{jk} = \alpha_{kj}, \quad \beta_{ql} = \beta_{lq}, \quad a_{mn} = a_{nm},$$

and the homogeneity restrictions are

$$\sum_j \alpha_j + \sum_m a_m = 1, \quad \sum_k \alpha_{jk} + \sum_m c_{jm} = 0 \quad \forall \ j, \quad \sum_n a_{mn} + \sum_j c_{jm} = 0 \quad \forall \ m$$

$$\sum_j b_{jq} + \sum_m d_{mq} = 0 \quad \forall \ q, \quad \sum_m a_{mt} + \sum_j \alpha_j = 0 \quad \forall \ t.$$
Normalizing the shadow profit function in (45) will impose the homogeneity restrictions with respect to one price. Since the objective is to estimate markups for both outputs, we use the $J$th input price to normalize the other prices. Then we calculate $SR^*_m$ and for $m = 1, 2, \ldots, M$, and $SC^*_j$ for $j = 1, 2, \ldots, J - 1$.

From the shadow profit function in (45), $SR^*_m$ and $SC^*_j$ can be derived as

$$SR^*_m = \frac{\partial \ln \pi^*}{\partial \ln p^*_m} = a_m + \sum_n a_{mn} \ln p^*_n + \sum_j c_{jm} \ln w_j + \sum_q d_{mq} \ln z_q + a_{mt}, \quad (46)$$

and

$$-SC^*_j = \frac{\partial \ln \pi^*}{\partial \ln w_j} = \alpha_j + \sum_k \alpha_{jk} \ln w_k + \sum_q b_{jq} \ln z_q + \sum_m c_{jm} \ln p^*_m + \alpha_j t. \quad (47)$$

Finally, using (46) and (47) above, $H = \{ \Sigma_m SR^*_m / \theta_m + \Sigma_j SC^*_j \}$ can be expressed in terms of the unknown parameters and observed data. Consequently, the profit system (after adding classical error terms to each equation) becomes

$$\ln \pi^\alpha = \ln \pi^* + \ln H + v$$

$$SR^\alpha_m = \frac{P_m Y_m}{\pi^\alpha} = \frac{SR^*_m}{H} \left\{ \frac{1}{\theta_m} \right\} + v_m \quad (48)$$

$$SC^\alpha_j = \frac{W_j X_j}{\pi^\alpha} = -\frac{SC^*_j}{H} + e_j$$

which can be estimated using the iterative non-linear seemingly unrelated regression technique. One of the share equations has to be dropped because the shares sum to unity.17

It should be noted here that the above approach does not recognize endogeneity of output prices. If one assumes that producers have monopoly power in the output market, then we have to add the demand (or inverse demand) functions to take care of the endogeneity of output prices. Adding $M$ such inverse demand functions (one for each output price) will make the above system consistent in the sense that the number of equations equals the number of endogenous variables ($x, y, p$).

One endemic problem in estimating profit functions (standard or alternative), using a translog or Cobb-Douglas functional form, is that profit has to be positive. This is, however, not the case in reality. In many studies, especially in

---

17 Productivity decomposition formula for this problem is left to the readers.
banking, negative profits force researchers to use other functional forms that allow negative profits. A majority of the banking studies, however, avoid this problem by adding a positive number large enough to make profit for every bank positive (see, for example, Berger and Mester (1997, 2003), Berger and DeYoung (2001), and many others). This is clearly inadmissible, both economically and econometrically. Profit used in the profit function should be defined as revenue minus cost (which may be different from accounting profit). So any change in profit has to be reflected in the profit function. That is, if we assume no inefficiency and start from the profit function \( \pi^a = \pi(p, w, t) \) then \( \pi^a + c = \pi(p, w, t) + c \) and, therefore, \( \ln(\pi^a + c) = \ln(\pi(p, w, t) + c) \). However, what is used in practice is \( \ln(\pi^a + c) = \ln\pi(p, w, t) \), i.e., no adjustment is made on the right-hand side of the equation. This ad hoc procedure cannot be economically justified. Adding something to the dependent variable and then taking the log of it changes the intercept as well as the coefficients of all the right-hand side variables (regressors) in the profit function. Thus, the estimates of technological parameters can easily be manipulated (and therefore, estimates of returns to scale, input substitutability, technical change, etc.) simply by changing the positive constant that is added to profit. Since the alternative profit function approach uses the same procedure to avoid taking log of negative profits, it is subject to the same criticism.

### 6. SOME NEW ISSUES AND CHALLENGES

#### 6.1. Competitiveness, Deregulation and Efficiency

Economists since Adam Smith have argued in favor of the virtues of competitive markets. The competitive framework is used as a benchmark because it leads to socially efficient outcomes. Any departures from competitive input and/or output markets lead to the inefficient allocation of resources and production of outputs, resulting in a deadweight loss to society. There might be many reasons why firms in some markets may not be competitive. For example, firms might possess some degree of market power in selling their products and/or buying their inputs. Consequently, the main driving force behind deregulatory effort is to increase competition in the hope of reducing the deadweight loss to the society. Here we examine the competitive and efficiency issues in the context of banking.

Like many other countries, the banking industry in Europe has faced important changes. These changes were specifically designed to liberalize the provision of services, bringing increased competition to the banking industry. Additionally, the establishment of the economic and monetary union (EMU), along
with developments in information technology and removal of entry barriers are other important changes faced by the European banking markets. Many of these changes have important implications for the competitive structure of the banking and financial sectors. Some of the techniques proposed in this survey can be used to examine whether output markets are competitive or not, as well as the contribution of non-competitive output markets on TFP growth. We can also estimate the degree and the temporal behavior of non-competitiveness from the estimates of markup factors. In a cross-country study the information on markups may be useful to the policymakers (at the EU level), in the sense that they can examine whether banks from the new EU member (or to-be EU member) countries can survive if such markups are eliminated due to competition. More work is needed in this area.

Given that the banking system plays a unique role in an economy, the efficiency of that system can hardly be ignored. Studying the efficiency of the individual banks, as well as the banking system, can provide feedback on ways to improve policy-making decisions. Models of efficiency studies discussed in this survey can answer many questions, some of which are: Does regulation increase efficiency of banks? Are de novo banks more efficient than existing banks? Can banks from the countries joining the EU survive in competition with the foreign banks, if left without subsidy or other assistance? Can efficiency and/or scale economies explain bank mergers? These questions can be addressed using micro data from the EU countries. One can also use the aggregate industry level data from the EU member countries to examine differences in efficiency, productivity, technology, and competitiveness of the banking industry. If the industry is not competitive and one erroneously makes the assumption that it is competitive, efficiency and productivity measures are likely to be biased (especially when an econometric technique is used).

If one uses value added per worker as the measure of productivity and the sole indicator of performance, the measure cannot separate the contribution of competitiveness to productivity improvement. That is, if one compares value added per worker from two different industries, the difference would not tell us anything about the contribution of competitiveness, even if we know that one industry is more competitive than the other. This is because difference in value added per worker captures the effect of many other things such as difference in technology, size and scale differences, price differences, etc. Contribution of these factors to the overall productivity growth (whether using a partial or total factor productivity index) cannot be separated without estimating the technology directly or indirectly. For example, if cost-minimizing behavior is used, the contribution of competitiveness on TFP growth can be computed once the cost function is estimated. The benchmark is the competitive output markets. The
The main advantage of the cost function approach is that we do not have to assume a particular form of market structure to estimate this effect. It can also handle multiple outputs, and can disentangle the effect of non-competitive behavior in each output market. This is especially useful if the industries under investigation produce more than one output and the degree of competition varies across outputs.

The cost function approach cannot test any hypotheses about competitive output markets. For this, a profit function approach has to be used that can allow distortions in output prices due to regulation, non-competitive behavior, etc. Thus, if we want to model price distortion (and decomposes its effect on productivity growth), it is necessary to use a profit function approach.

Another challenge is to incorporate quality differences in outputs that can be confounded with non-competitive behavior, because quality difference is reflected in the prices. If output qualities differ across firms and these are not taken into account, deviations in the \( p = MC \) rule might be thought of as non-competitive behavior. Similar to quality differences, there might be heterogeneity in products, behaviors, etc., among firms/industries operating in different countries. Such differences are to be taken into account. Furthermore, cross-country data may not be compared directly, unless the outputs, inputs, behavior, etc., are homogeneous. Another complication is that efficiency/performance of firms is likely to change over time and a model should be used that allows for his possibility.

Multi-dimensional indices of productivity and efficiency measures, such as labor productivity, employment growth, competitiveness, globalization, foreign investment, etc., are often constructed to examine performance. Some firms may be champions based on, for example, one or some of these criteria. This does not, however, mean that the same firms will be champions when the other criteria are used. Thus, an overall productivity index has to be constructed from these individual indices to make a meaningful comparison. That is, to get a macro outlook, we have to dig down into the microanalysis.

In all the approaches discussed so far, the effect of competition or lack of it is examined from the producers’ point of view. However, it is well-known that non-competitive markets contribute to deadweight loss to the society meaning that the society would be better off under competitive markets. By focusing only on the production side, we fail to see the compete picture and miss the deadweight loss arising out of, for example, monopolistic markets. The deadweight loss can be measured from the markups in output prices and difference in the outputs produced with and without competitive market conditions. Both the output differential and markups can be obtained from the estimates of shadow profit function. No one needs to look into the deadweight loss component in TFP growth decomposition.
6.2. Data Requirement for Improving the Efficiency and Productivity Measures

To address the issues mentioned above, for example in banking, extensive data is required, some of which are not reported in the balance sheets and income statements. For a meaningful efficiency analysis, for example, data is needed on different types of loans (consumer loans, business loans, real estate loans, etc.), different types of deposits, assets, capital, foreign exchange reserves, labor, wages and salaries, interest paid on different types of deposits, and rates charged on different types of loans. Quality measure of loans might be important for disentangling efficiency and productivity measures from output quality. Variables on quality of inputs, R&D expenditure, expenditure on information technology, etc., are important and useful in disentangling productivity differential. Depth of the study will depend on the type and extent of data made available to do the analysis.

In this survey we discussed several methods, some of which can be used to crosscheck results from competing models. Data requirements of these models also vary, and so does behavioral assumptions. For example, the distance functions require data on input and output quantities, while estimation of profit functions require data on input and output prices as well as profit. While discussing alternative profit functions in Section 5.3, we mentioned that profitability might be affected through revenue and cost. For example, loan quality and non-interest net income affect revenue, and cost inefficiency increases cost. Thus, excluding quality of loans from the analysis is likely to mis-measure revenue, which is likely to bias the results of the profit function approaches. If information on the quality of output variables is obtained and these quality variables are used in the cost function, estimates of the efficiency results would be much more reliable. If the quality variables are not included (due to non-availability of data), estimates of efficiency scores are likely to be contaminated. High efficiency score might be due to bad output quality and vice versa.

To put some of these in a broader perspective, consider some of the simpler non-econometric performance measures that are widely used in practice. For example, partial factor productivity, such as value added per worker, is often used as an indicator of performance. Another indicator is employment growth, although comparison between growth rates of value added and employment across industries, sectors, regions, etc., might not give similar performance measures. For example, value added per worker can increase without increasing employment growth if workers become more productive by using more capital (substituting labor with capital) and/or better technology. In this respect, growth in value added per worker is a better measure than growth in employment. However, even with this simple measure, we would like to decompose growth in
value added per worker into returns to scale and technical change components. To do so, we have to estimate the underlying technology using a primal or dual approach. In estimating the technology and using it to measure growth rates of value added per worker, we can control for employment growth (indirectly in the regression equation), instead of using it as a separate measure of productivity. Again, we can use the metafrontier approach to explain differences in productivity growth across regions, industries, and over time. The advantage of the metafrontier approach, as mentioned before, is that we can examine the role of differences in technology (labeled as technology gap) and efficiency change in the overall productivity change. This will be helpful in examining whether the top performing firms (champions) are using the best practice technology (metafrontier) and are also technically efficient. For some industries/regions, the champion firms might be technically efficient and may also be using the best practice technology. But this may not be true for every sector/region. The typical growth accounting procedure (although useful and simple) cannot capture the sources of productivity growth differentials across regions, industries, size of firms in an industry, etc. Therefore, the simple value added approach cannot separate the effects of technical efficiency from the technology.

6.3. Use of Aggregated and Micro Data

Since the number of EU member countries is relatively small, we have to pool time series data on the member countries to estimate efficiency, productivity, etc., especially if we want to use the aggregated macro data. This puts a limit on allowing technological heterogeneity among EU member countries, especially in econometric models. If we assume a single technology for all member countries, the estimated inefficiency will be relative to the single frontier defined for all member countries. Thus, we cannot address the question of technology gap. That is, technology gap and inefficiency will be lumped together, yielding higher estimates of inefficiency. Furthermore, the estimated technology might be biased, if more than one technology is used in practice. This problem can be avoided if we use either a mixing model approach or somehow group the countries based on some a priori information, and estimate the technology for each group separately.

Another potential problem in using cross-country data is consistency of the data. That is, all the input and output variables, prices, and other control variables (if any) need to be defined in the same way. Sometimes the variables are indeed defined in the same way but certain components might be missing. This is hardly the case in reality. For example, data on loans might have ten classifications in
one country but for another country there might be five categories. Thus, the analyst has to know whether similar procedures are used in constructing the variables that are used to estimate the technology. This problem is avoided if we use bank level data and estimate a separate frontier function for each country.

The data definition/construction problem mentioned above is absent when we use cross-sectional (panel) data on banks from each country because the same procedure is used to construct the variables for each bank. Thus, we can estimate the frontier function for each country and measure efficiency and productivity relative to the estimated frontier. However, sometimes we want to compare bank efficiency across countries. Since each country has a frontier of its own, for a valid comparison of efficiency we have to take into account the fact that the technologies are different. Spanish banks might be efficient relative to their own frontier, but they might not be using the best practice technology. Thus, there might be a technology gap. To estimate the technology gap, we have to construct a metafrontier using the country-specific frontiers. The deviation of country frontiers from the metafrontier is then viewed as a technology gap. Thus, for example, Swiss banks might be technically efficient (relative to their own frontier) but when the technology gap is taken into account, they may not be as so efficient. That is, if a technology gap exits, efficiency of Swiss banks relative to the metafrontier (global frontier) would not be as efficient. In other words, to get a full picture we have to take into account both the country-specific frontiers and the metafrontier. This would be a better indicator of bank performance than that which is currently used in the banking literature.

While working with aggregate country level data, we often look at things beyond the traditional production/cost/profit function approach. For example, a new literature on productivity measurement (productivity and the new economy (Nordhous, 2001)) has recently emerged. Following this approach, variables can be included, such as technological capacity, human capital, financial capacity, enterprise and innovation, openness, adaptability, etc. to enrich the productivity measures. These variables (for which indices are to be constructed) can affect productivity either indirectly by enhancing efficiency of traditional inputs such as capital and labor, or directly through improving efficiency, i.e., increasing outputs holding inputs constant.

7. CONCLUSIONS

In this survey we focused on the parametric models to estimate and decompose TFP growth into scale, technical change, and technical efficiency change components. For modeling inefficiency we concentrated on output-oriented technical
inefficiency. The TFP growth formulas for input-oriented technical inefficiency (for some selected models) are given in the appendix. Throughout the analysis, we assumed that firms are allocatively efficient to make things simple. For modeling technical change we used both time trend and factor-augmenting approaches. In the primal (quantity based) models we discussed the single output production function approach, as well as input and output distance functions (that accommodates multiple outputs). To allow for the possibility that sample firms might use more than one technology and the analysts might not know who is using which technology, we considered the latent class modeling approach.

In addition to estimating technical inefficiency for each firm, this approach is capable of measuring technology gap (distance between the individual/group frontiers from the metafrontier).

In the dual cost function approach, our TFP decomposition analysis was done for single and multiple output cost functions. We briefly discussed the modeling issues for the latent class models. Finally, we considered the profit function approach in both single and multiple output frameworks. We also examined the alternative profit function and discuss problems associated with this approach. Finally, we discussed extensions of the standard profit function to accommodate markups in the output markets.

Estimation issues are not addressed in this survey. Since one standard technique would not fit all the models, and some of the techniques (especially when the cost and profit function models that require use of a system approach) are quite involved, we decided not to discuss estimation techniques in this survey. However, econometric techniques are available to estimate every model discussed in this survey.

BIBLIOGRAPHY


Productivity and Efficiency Measurement Using Parametric Econometric Methods


Appendix A

TFP growth formula with input-oriented technical inefficiency

Write the production function as

\[ y = f(\hat{x}, t) = f(x \exp(-\eta), t) \]

\[ \Rightarrow \dot{y} = \frac{1}{f} \sum_j f_j \hat{x}_j \left( \dot{x}_j - \frac{d\eta}{dt} \right) + \frac{\partial \ln f}{\partial t} \quad (A.1) \]

\[ \Rightarrow TFP = (RTS - 1) \sum_j \lambda_j \dot{x}_j + \sum_j (\lambda_j - s^\eta) \dot{x}_j - RTS \sum_j \lambda_j \frac{d\eta}{dt} + \dot{f}, \]

where \( \hat{x}_j = x_j \exp(-\eta) \). \( RTS = \sum_j \frac{\partial \ln f}{\partial \ln x_j} = \frac{1}{f} \sum_j f_j x_j e^{-\eta} \), \( \lambda_j = f_j x_j / \sum_k f_k x_k \), and the dot over a variable represents its rate of change.

Since \( RTS \) depends on \( \eta \) (so is \( \lambda_j \)), the above formula can be rewritten as

\[ TFP = (RTS^0 - 1) \sum_j \lambda_j^0 \dot{x}_j + \dot{j}^0 - \frac{d\eta}{dt} + \text{misc} \quad (A.2) \]

where \( RTS^0 (=RTS|\eta = 0) \), \( TC^0 (=TC|\eta = 0) \) and \( \dot{j}_i^0 (=\dot{j}_i|\eta = 0) \). Finally, the miscellaneous component can be further decomposed into deviations of \( RTS \) from \( RTS^0 \), \( TC \) from \( TC^0 \), etc.

Note: The OO and IO measures are identical if the production function is homogeneous. In other words, one is a constant multiple of the other and there is no difference in estimating these models.
Appendix B

TFP growth formula with output-oriented non-neutral technical inefficiency

The production function is

\[ y = f(x, t) \exp(-u(z, t)) \text{ where } z \text{ does not include } x. \]

\[ \Rightarrow \dot{y} = \frac{1}{f} \sum_j f_j x_j \dot{x}_j + \sum_k \frac{\partial \ln e^{-u}}{\partial \ln z_k} \dot{z}_k + \frac{\partial u}{\partial t} + \dot{f}_t \]

\[ = \text{RTS} \sum_j \lambda_j \dot{x}_j + \sum_k I_k \dot{z}_k - \frac{\partial u}{\partial t} + \dot{f}_t \]

where \( \frac{\partial \ln e^{-u}}{\partial \ln z_k} = I_k \) is efficiency change induced by \( z_k \)

Thus, TFP = \((\text{RTS} - 1) \sum_j \lambda_j \dot{x}_j + \dot{f}_t - \frac{\partial u}{\partial t} + \sum_k I_k \dot{z}_k + \sum_j (\lambda_j - s^p_j) \dot{x}_j \) (B.1)

= Scale + TC + TEC + induced by the \( z \) variables + price/allocative
Appendix C

TFP growth formula with input-oriented non-neutral technical inefficiency

The production function is:

\[ y = f(\hat{x}, t) = f(x \exp(-\eta(z, t), t) \]

\[ \Rightarrow \dot{y} = \frac{RTS}{\sum_k f_k \hat{x}_k} \left[ \sum_j f_j \hat{x}_j \left\{ \dot{x}_j + \sum_k I_k \dot{z}_k - \frac{d\eta}{dt} \right\} \right] + \dot{f}_t \]

where \( RTS = \sum_j \frac{\partial \ln y}{\partial \ln x_j} = \frac{1}{2} \sum f_j \hat{x}_j \) (depends on \( \eta \))

Thus, \( \dot{TFP} = (RTS - 1) \sum_j \lambda_j \dot{x}_j + \dot{f}_t + RTS \sum_k I_k \dot{z}_k \sum_j \lambda_j \]

\[ -RTS \sum_j \lambda_j \frac{\partial \eta}{\partial t} + \sum_j (\lambda_j - \dot{x}_j^\circ) \dot{x}_j \]

\[ = \text{Scale} + TC + \text{Induced by } z + TEC + \text{Price/Allocative} \]

Note: Each component of TFP growth depends on \( \eta \). The formula can be rewritten so that the above components, except a residual component, are free from \( \eta \). The residual term will contain \( \eta \), which can be further decomposed into scale, TC, etc.
Appendix D

**TFP growth formula with output-oriented technical inefficiency**

**Without allocative inefficiency**

\[
\begin{align*}
\text{Min}_x \ w'x \ s.t. \ y &= f(x, t)e^{-u} \\
\Rightarrow C^u &= C(w, ye^u) \\
\Rightarrow TFP &= \ddot{Y}(1 - E_{cy}) - \dot{C}_t - E_{cy} \frac{\partial u}{\partial t} \\
\end{align*}
\]

(D.1)

where \( E_{cy} = \partial \ln C / \partial \ln y \).

**With allocative inefficiency**

Production function: \( ye^u = f(x, t) \)

The first-order conditions are \( \frac{\partial f_j}{\partial x_1} = \frac{w_j e^{\xi_j}}{w_1} \equiv \frac{w_j}{w_1} (\xi_1 = 0) \)

\[
\Rightarrow x_j = x_j(w^e, ye^u) \\
\Rightarrow C^v = \sum(w_j x_j) = C^v(w^e, ye^u, t)
\]

Shephard’s Lemma: \( \frac{\partial \ln C^v}{\partial w_j} = S_j = \frac{w_j x_j}{C^v} \)

Thus, \( C^u = \sum w_j x_j = C^v \sum_j w_j S_j / w_j' \equiv C^v \cdot G(\cdot) \)

\[
\Rightarrow \ln C^u = \ln C^v + \ln G(\cdot).
\]

\[
\Rightarrow TFP = \ddot{Y}(1 - E_{cy}) + \sum_j (S_j^u - S_j^v) \dot{w}_j - \dot{C}_t - E_{cy} \frac{\partial u}{\partial t} - \frac{\partial \ln G(\cdot)}{\partial t} \\
\]

(D.2)

where \( E_{cy} = \partial \ln C^v / \partial \ln y. \)
Appendix E

TFP growth formula with non-neutral output-oriented technical inefficiency

With Allocative Inefficiency

Production function: \( y = f(x, t) \exp(-u(z, t)) \)

The first-order conditions: \( f_j/f_1 = w_j, e^{\xi_j}/w_1 = w_j^*/w_1 \)
\( \Rightarrow \ln C^* = \ln C^\alpha (w^*, y \exp(u, t), t) + \ln G(\cdot) \)
\[
\dot{TFP} = \dot{y}(1 - E^{\dot{y}}_c) + \sum (S^a_j - S^c_j) \dot{w}_j - \dot{C}^c_j - E^{\dot{y}}_c \frac{\partial u}{\partial t} + \sum I_k \dot{z}_k E^{\dot{y}}_c - \frac{\partial \ln G}{\partial t} \quad (E.1)
\]

where \( I_k = \left\{ \frac{\partial \ln e^{u(t)}}{\partial \ln \gamma_k} \right\} \)

Note: (i) As before, the above formula can be expressed in terms of scale, TC, TEC, etc., defined at the frontier \( u = 0 \). (ii) Cost function with non-neutral IO technical inefficiency will be similar to the neutral case except for one extra term involving \( (I_k \dot{z}_k) \). (iii) Every component depends on \( u \).

Rewrite \( \dot{TFP} \) as
\[
\dot{TFP} = \dot{y}(1 - E^{\dot{y}}_c^0) - \dot{C}^{\dot{y}}_c - E^{\dot{y}}_c \frac{\partial u}{\partial t} + \rho(\xi, w, u, y) \quad (E.2)
\]

where the miscellaneous component, \( \rho(\xi, w, u, y) \), can be further decomposed into deviations of technical change, RTS, TEC, etc., from their respective values at the frontier.
Please provide citation for fig. 2.2.