CONTRIBUTI DI RICERCA CRENOS

MEASURING PRODUCTIVITY

Massimo Del Gatto
Adriana Di Liberto
Carmelo Petraglia

WORKING PAPERS

2008/18

CENTRO RICERCHE ECONOMICHE NORD SUD
(CRENOS)
UNIVERSITÀ DI CAGLIARI
UNIVERSITÀ DI SASSARI

Il CRENoS è un centro di ricerca istituito nel 1993 che fa capo alle Università di Cagliari e Sassari ed è attualmente diretto da Raffaele Paci. Il CRENoS si propone di contribuire a migliorare le conoscenze sul divario economico tra aree integrate e di fornire utili indicazioni di intervento. Particolare attenzione è dedicata al ruolo svolto dalle istituzioni, dal progresso tecnologico e dalla diffusione dell'innovazione nel processo di convergenza o divergenza tra aree economiche. Il CRENoS si propone inoltre di studiare la compatibilità fra tali processi e la salvaguardia delle risorse ambientali, sia globali sia locali.
Per svolgere la sua attività di ricerca, il CRENoS collabora con centri di ricerca e università nazionali ed internazionali; è attivo nell'organizzare conferenze ad alto contenuto scientifico, seminari e altre attività di natura formativa; tiene aggiornate una serie di banche dati e ha una sua collana di pubblicazioni.

www.crenos.it
info@crenos.it

CRENOS – CAGLIARI
VIA SAN GIORGIO 12, I-09100 CAGLIARI, ITALIA
TEL. +39-070-6756406; FAX +39-070- 6756402

CRENOS - SASSARI
VIA TORRE TONDA 34, I-07100 SASSARI, ITALIA
TEL. +39-079-2017301; FAX +39-079-2017312

Titolo: MEASURING PRODUCTIVITY

ISBN: 978-88-8467-494-4

Prima Edizione: Dicembre 2008

Via Is Mirrionis, 1
09123 Cagliari
Tel./Fax 070291201
www.cuec.it

Measuring productivity*

Massimo Del Gatto
G.d'Annunzio University
and
CRENoS

Adriana Di Liberto
University of Cagliari
and
CRENoS

Carmelo Petraglia
University of Napoli
Federico II

December 2008

Abstract

Quantifying productivity is a *conditio sine qua non* for empirical analysis in a number of research fields. The identification of the measure that best fits with the specific goals of the analysis, as well as being data-driven, is currently complicated by the fact that an array of methodologies is available. This paper provides economic researchers with an up-to-date overview of issues and relevant solutions associated with this choice. Methods of productivity measurement are surveyed and classified according to three main criteria: i) macro/micro; ii) frontier/non-frontier; iii) deterministic/econometric.

Keywords: productivity measurement, TFP, Solow residual, endogeneity, simultaneity, selection bias, Stochastic Frontier Analysis, DEA, Growth accounting,, GMM, Olley-Pakes, firm heterogeneity, price dispersion.
J.E.L. Classification: O40, O33, O47, C14, C33, C43.

1 Introduction

Quantifying productivity is a *conditio sine qua non* for empirical analysis in a number of research fields. What is usually needed is a measure of output differences which is not explained by different input choices and occurs, instead, through marginal product increases. This "quantity", usually referred to as Total Factor Productivity (henceforth TFP), is the essence of the economic notion of productivity. To put it formally, what the economists have in mind when they talk about productivity is a production function of the type

$$Y_{it} = A_{it}\ F(\mathbf{X}_{it}) \tag{1.1}$$

relating the output (Y) of a generic unit (firm/industry/country) i at time t to a (1xL) vector of inputs $\mathbf{X}$ and the term A saying how much output a given unit is able to produce from a certain amount of inputs, given the technological level. The state of technology, embodied by the function $F(\cdot)$, is given and common to all i's. Hence, the TFP index at time t is the ratio of produced output and total inputs employed:

$$TFP_{it} \equiv A_{it} = \frac{Y_{it}}{F(\mathbf{X}_{it})} \tag{1.2}$$

*Contacts: m.delgatto@unich.it, adriana.diliberto@gmail.com, petragli@unina.it.
The first two authors acknowledge financial support from the European Community under the FP7 SSH Project "Intangible Assets and Regional Economic Growth"grant n. 216813.

The idea is quite simple, but giving it an operational content is not an easy task. An array of methodologies is available, and researchers have to make a choice that, even when the estimation is only propaedeutical to the main analysis, it is likely to represent most of the story of an article.

This paper aims at contributing to this choice by reviewing most of the available methodologies for productivity estimation, which we classify according to different criteria. In Figure 1, we distinguish between *deterministic* methodologies, whose output is a "calculated"measure of TFP, and *econometric*, providing us with "estimated"productivity levels and/or growth rates. Within these, we discriminate between *Frontier* and *Non-Frontier* Approach.

However, the first distinction one should keep in mind when approaching this field is between methodologies used in macro studies, that is methods concerned with *aggregate* (countries/regions/industry) productivity and methodologies used in micro studies, aimed at measuring *individual* (firm/plant) productivity. Thus, in Figure 1 we also indicate if a specific technique has been applied to macro or micro data sets, or to both. Although in principle one would expect some form of aggregation of the latter to exactly reproduce the former, as we will discuss in section 2, this is not the case. Thus, it is not surprising that the two strands of literature develop along different lanes and are rather difficult to compare.

	Deterministic Methodologies	**Econometric Methodologies**	
		Parametric	**Semi- Parametric**
Frontier	DEA (Data Envelopment Analysis) (micro-macro) FDH (Free Disposal Hull) (micro-macro)	Stochastic Frontier Analysis (micro-macro)	
Non-Frontier	Growth Accounting (macro) Index Numbers (micro-macro)	Growth Regressions (macro)	Proxy-variables (micro)

Figure 1: Surveyed methodologies

Aggregate studies are mainly concerned with identifying the role of TFP on growth dynamics, the main goal consisting of explaining the still wide differences in economic performance across countries. This literature started with the Solow growth theory, in which the pattern of productivity growth essentially mirrors that of the so called technological progress (i.e. Solow residual). This approach goes under the name of *Growth accounting* and it has been the first deterministic methodology proposed to estimate TFP and has been used to estimate TFP at both aggregate and sectoral levels. The first evidence dates back to the fifties (Abramovitz, 1956; Solow, 1957) and, despite its age, still represents one of the most popular ways to estimate TFP. New methodologies have been suggested to improve traditional Solow residual estimates. In particular, a recent extension of the growth accounting methodology is the level of development accounting decomposition (Klenow and Rodriguez Clare, 1997; Hall and Jones, 1999; Caselli, 2005). This methodology has the advantage to produce estimates of TFP levels instead of their growth rates as in traditional growth accounting. The focus on TFP levels instead of rates of change is particularly important in growth models where technology transfers represent the main engine for growth and convergence (Parente and Prescott, 1994; Benhabib and Spiegel, 1994 and 2005).

Among macro studies, we also describe alternative parametric methods to estimate TFP: the so called

growth regressions. Like growth accounting, these are extensions of the standard Solow growth model. However, unlike the former, they use a model-based approach, as its main contribution is to identify a structural equation to estimate TFP levels from aggregate data (Islam, 1995; Caselli et al. 1996). Therefore, with respect to the Growth Accounting approach, the advantage of growth regressions is that TFP is not estimated as a residual from a calibration exercise and we do not need to trust specific functional form assumptions.

Frontier models are applied to both aggregate and individual data. They differ from the *Non-Frontier* models for the assumption that observed production units do not fully utilize their existing technology. In the presence of inefficiency, productivity measurement is affected and so it will be productivity change, unless inefficiency does not vary over time. Since in many contexts it is relevant to provide evidence on the contribution of efficiency change to productivity change, a main advantage of these models is that they allow for the presence of time varying technical inefficiency in production. The main reason leading to the adoption of frontier models is then their capability to disentangle two main sources of productivity growth: technological change and technical efficiency change. Technological progress is assumed to push the frontier of potential production upward, while efficiency change reflects the capability of productive units to improve production with a set of given inputs and available technology. An advantage of frontier models is that they can provide useful information to the policy maker for the design of productivity-enhancing policies. For instance, if the main source of a productivity slowdown is detected to be technological regress, this would suggest orienting policies towards measures that induce technological innovation. We will focus on the key features of both deterministic and econometric branches of this field of research based on *Data Envelopment Analysis* (DEA) and *Stochastic Frontier Analysis* (SFA) respectively. DEA and SFA will be reviewed within other deterministic and econometric methodologies respectively.

DEA (Farrell, 1957; Charnes et al., 1978) can be seen as an attempt to overcome some of the specific weaknesses of the growth accounting approach: a particular functional form for technology, particular assumptions on market structure, the hypothesis that markets are perfect. The basic idea of this approach consists of enveloping the data (the observed input-output combinations) in order to obtain an approximation of the production frontier (or "best-practice" frontier) and using this to identify the contribution of technological change, technological catch-up, and inputs accumulation to productivity growth. SFA also starts by assuming that firms can not produce using the most efficient possible way but, differently from DEA, accommodate for shortfall from potential output due to random shocks beyond the control of producers. In order to describe the main departures of SFA from other econometric approaches to the estimation of TFP growth, our survey will first focus on the estimation of technical inefficiency in a cross-sectional framework and then on the SFA approach to the decomposition of TFP in a panel context proposed by Kumbhakar (2000).

As for micro level studies, the interest in estimating firm-level productivity received early this century an extra-kick from two simultaneous and related circumstances: the development of a theoretical literature in which firms are assumed to be heterogeneous (and in which heterogeneity is thought of in terms of productivity), and the increasing availability of micro-level data. Recent developments in growth theory focused on more sophisticated mechanisms describing the channels through which firms competition and selection affect innovation incentives (Aghion et al., 1999) or where the organisation of firms and production should be different in industries that are closer to the world technology frontier (Acemoglu et al., 2006). However, the main focus of this strand of literature is on the relationship between the productivity distribution of firms and the integration process (Melitz, 2003; Bernard et al., 2003; Melitz and Ottaviano, 2008; Bernard et al. 2007; Chaney, 2008). Related to this, the empirical literature is of course interested in understanding firm-level differences in performance, as well as in studying the determinants of these differences (see e.g. Clerides et al., 1998; Bernard and Jensen, 1999; Aw et al., 2003; Pavcnik, 2002; Bernard et al., 2006; Roberts and Tybout, 1997; Syverson, 2004; Del Gatto et al., 2008). To deal with these questions, methods providing a measure of productivity in "levels" are needed. Studies in this field usually rely on semi-parametric methods, based on proxy variables. These methods are conceived for keeping into account the main problems associated with estimating productivity at the firm level, namely: simultaneity, selectivity, and price-dispersion. Apart from the selection bias, these problems are not specific to the micro context, but, in such context, several methods have been developed to cope with them. The key points of these (semi-parametric) methods are i) the identification of a proxy variable, which is function of the observed (by the firm) TFP, and ii) the definition of the conditions under which this function can be inverted in order to express TFP as a function

of the proxy variable itself. For example, Levinsohn and Petrin (2003) suggest using intermediate goods as a function of TFP and capital. This function is invertible provided that, with given capital, the utilization of intermediate goods increases with the growth in TFP. Olley and Pakes (1996) suggest using investment instead. Although this idea of recovering TFP by the traces it leaves in the (observed) behavior of the firm is the main novelty of this approach, the implied "invertibility conditions" can represent a weakness, as it has to hold for all firms, regardless of their size and market position.

The described methodologies can be regarded as the main blocks of methods for productivity estimation. However, they do not exhaust the array of available techniques.[1] In particular, it is worth noting that most methods[2] surveyed in this paper require data on inputs and output. This not only causes the problems discussed in section 5.2, when the reference market structure is imperfect competition, but it is troublesome also in a perfectly competitive framework and at the aggregate level. Expressing the amount of inputs in physical terms is in fact not straightforward. Opportune deflators are always needed, and evaluating the stock of capital used in production becomes very data demanding if one opts for the perpetual inventory method. Although this might not be decisive if the focus is on a single country, when one aims at comparing productivity across sectors and/or countries, data on inputs and output must be comparable, and this is usually not the case. It is well known that cross-country heterogeneities in the quality of data are quite large, and this is particularly true for data on capital. An alternative approach has been developed in order to overcome these shortcomings. The idea is that trade flows, via comparative advantages, embody cross-country differences in sectoral productivity. To the extent in which this information can be drawn out, productivity levels, this time comparable across countries and sectors, can be easily obtained. Finicelli et al. (2008) apply this reasoning to the probabilistic ricardian framework of the Eaton and Kortum (2002) model, while Fadinger and Fleiss (2008) move in a monopolistic competition framework. For its nature of model-based analysis, a detailed description of this line of research is very space-demanding. Since these promising techniques are still under the evaluation of the scientific community, we will leave them out of the analysis.

The following exposition is articulated as follows. Sections 2 and 3 open with a general discussion of macro versus micro and frontier versus non frontier issues. Section 4 describes deterministic methodologies to estimate TFP, namely Growth Accounting (section 4.1), Index numbers (section 4.2) and DEA (section 4.3). All these methodologies are applied in both macro and micro contexts. Section 5 is devoted to the econometric estimation strategies. Section 5.1 describes the techniques with macro datasets. Section 5.2 describes the estimation strategies for micro studies: Proxy-variables methods (5.2.1) and methods correcting for price dispersion under imperfect competition (section 5.2.2). Section 5.3 describes SFA. Section 6 concludes.

2 Macro vs micro TFP measures

Attention is increasingly growing away from the study of TFP at the aggregate and industry level of detail, towards the firm/plant level. This recent shift in focus may be explained by different factors. First of all, data availability and computing power have improved. Furthermore, from the theoretical point of view there has been a shift from competitive to non-competitive models of analysis. In growth theory, models of endogenous growth focus on increasing returns to scale, non-competitive markets, externalities, creative destruction processes together with the idea that innovation (and, thus, productivity) is not "manna from heaven" but it is best seen as an endogenous part of the economic development. In particular, early shumpeterian growth models (Aghion and Howitt, 1992) argue that monopoly rent induces firms to innovate and thus positively affects productivity while further developments focused on more sophisticated mechanism describing the channels through which firms competition and selection affect innovation incentives (Aghion *et al.*, 1999) or where the organisation of firms and production should be different in industries that are closer to the world technology frontier (Acemoglu et al., 2006). On the other hand, new trade models in which firms are modeled as heterogenous in terms productivity (Melitz, 2003; Bernard et al., 2003; Melitz and Ottaviano, 2008; Bernard, Redding and Schott, 2007; Chaney, 2008) focus on the relationship between

[1] Van Biesebroeck (2007) provides a description of several methods dealt with in this survey, focusing on their robustness to: i) measurement error in inputs; ii) mis-specifications in the deterministic portion of the production technology; iii) erroneous assumptions on the evolution of unobserved productivity. Concerning *Frontier* models in particular, a recent and up-to-date review is provided by Fried et al. (2008).

[2] The only exception is represented by the growth regression approach where data inputs are not required.

the TFP distribution of firms and the integration process. Many of these hypothesis and mechanisms can be only investigated at the micro level.

Despite their growing importance, micro studies results may be sometimes difficult to generalise. Because of data availability, these studies usually investigate productivity patterns for a single economy or across groups of developed economies only, while they do not investigate TFP dynamics at global level. Therefore, one should be very cautious in extending the results obtained from a sample of one or few industrialized economies to, for example, less developed economies. Even for developed countries, datasets are often not comprehensive and lack information for important sectors of the economy as most studies focus on comparisons of the production functions of manufacturing industries. This may be misleading in cross-section comparisons since, as stressed by Caselli (2005), even data on the agricultural sector show large within sector TFP differences across countries and this helps to explain the observed large GDP differences in these economies. Macro and micro approaches can thus be considered as complementary as they ask different questions and produce different pieces of information.

In principle, the relationship between aggregate, industry and firm/plant level estimated productivity should include a mutually consistent measure of productivity at each level of analysis (see Hulten, 2001). In practice, things are not as obvious as it may seem. As discussed above, provided that one is able to estimate the productivity level of all the firms in a given industry, one would expect industry TFP and, then, aggregate TFP measures to result from some form of aggregation of each level of the hierarchy. And *vice versa*. In a top down analysis, we should be able to decompose aggregate TFP at industrial and then firm level. However, how to aggregate (bottom up) or, conversely, decompose TFP depends on many factors. For example, a well known problem when one needs to aggregate TFP from sectors to the whole economy is that aggregate studies assume that GDP is produced by a single sector and this implies that the role of inter-industry flows, that is, of intermediate goods, cancels out.[3]

But similar problems can be found at other level of the hierarchy. How to aggregate TFP from firm/plant to industry analysis depends, for instance, on the purposes of the analysis. In general, let us start with the following Cobb-Douglas specification of (1.1):

$$Y_{it} = A_{it} \prod_{n=1}^{N} \left(X_{n,it}\right)^{\beta_n} \tag{2.1}$$

where $X_{n,it}$ is the amount of input n (with $n = 1, .., N$) used, and β_n is the relevant production coefficient. Equation (2.1), which assumes implicitly Hicks neutral technical change, expresses firm i's output at time t as a function of a bundle of N inputs times the TFP component A_{it}. At the level of the single firm, A_{it} encapsulates (Griliches and Mairesse, 1995) unmeasured components such as R&D stocks and other intangibles, technology levels and marginal efficiency, input quality and effort.

In order to go from the firm to the industry, the simplest form of aggregation one can conceive is the sum. In this case, aggregate TFP turns out to be

$$A_t^Z = \sum_i w_{it} A_{it} = \frac{\sum_i Y_{it}}{\sum_i \left[\prod_{n=1}^{N} \left(X_{n,it}\right)^{\beta_n}\right]} \tag{2.2}$$

where the weights w_{it} represent firm's input share with respect to industry's total inputs ($\frac{\prod_{n=1}^{N}(X_{n,it})^{\beta_n}}{\sum_i\left[\prod_{n=1}^{N}(X_{n,it})^{\beta_n}\right]}$).

However, this simple weighting scheme does not reproduce aggregate/industry estimates, as the latter are commonly obtained as

$$A_t = \frac{\sum_i Y_{it}}{\prod_{n=1}^{N} \left(\sum_i X_{n,it}\right)^{\beta_n}}. \tag{2.3}$$

The two measures coincide only if the aggregation is made by using as weights $w_{it} = \prod_n \left(\frac{X_{n,it}}{\sum_i X_{n,it}}\right)^{\beta_n}$, whose sum, however, is not equal to one. Van Biesebroeck (2008) shows that a number of advantages is associated with using this weighting scheme, but, as far as we know, there are no other applications of such weighting scheme.

[3] A methodology to aggregate data from sectoral to country level TFP has been proposed by Domar (1961) and Hulten (1978).

This simple exercise shows how the choice to start from either macro or micro data on input and output leads to different notions of productivity. It is in fact evident that the economic meaning of A_t^Z and A_t is intrinsically different, the latter being the output per unit of input of the economy thought of as a single "big firm".

Thus, the choice about how to aggregate cannot be made irrespective of the objective of the analysis. A decomposition of the aggregate productivity is particularly important with respect to disentangling the contribution to the latter of firms of different size, market share, productivity, etc. In this respect, Olley and Pakes (1996) suggest a decomposition of weighted aggregate productivity (wa_t) into two parts: the unweighed aggregate productivity measure ($\bar{a}_t$) and the total covariance between a firm/plant's share of the industry output and its productivity:

$$wa_t = \sum_i w_{it} a_{it} = \bar{a}_t + \sum_i (w_{it} - \bar{w}_t)(a_{it} - \bar{a}_t) \tag{2.4}$$

where the bar over a variable denotes the mean over all plants in a given year. The covariance term is interesting because it represents the contribution to wa_t resulting from market shares and resource reshuffling from less productive to more productive firms. Of course the choice of the weights is not indifferent. Olley and Pakes (1996), and more recent studies such as Pavcnick (2002), use output shares, while Bartelsman and Dhrymes (1998) use input shares. Van Biesebroeck (2008) shows that using output shares amplifies the relative importance of the correlation term.

More complex specifications have been used to decompose productivity growth (see Foster et al. (2001) for a review). In particular, Baily et al. (1996) suggest the following decomposition:

$$\dot{a}_t = \sum_{i \in S} w_{it-1} \dot{a}_{it} + \sum_{i \in S} \dot{w}_{it} (a_{it} - A_{t-1}) + \sum_{i \in S} \dot{w}_{it} \dot{a}_{it} \tag{2.5}$$

$$+ \sum_{i \in N} w_{it} (\dot{a}_{it} - A_{t-1}) + \sum_{i \in X} w_{it-1} (\dot{a}_{it} - A_{t-1}) \tag{2.6}$$

where: w_{it} is the industry (output or input) share of firm i at time t; S, N and X are the sets of surviving, entering and exiting firms respectively; productivity is expressed in logs. According to this formulation, the productivity change between two periods results from five components: i) within-firm growth, weighted by initial output shares; ii) changing output shares weighted by the deviation of final firm productivity and initial aggregate productivity; iii) firm TFP growth times plant share change; iv) weighted sum of the difference between final TFP of entering firms and initial industry TFP; v) weighted sum of the difference between initial TFP of exiting firms and initial industry TFP. Note that, apart from the first term, which is the productivity change that would be caught by comparing two different estimations in levels, the other four terms account for complications arising only in a dynamic context. Following Bartelsman and Doms (2000), note how the contribution to aggregate productivity growth of continuing firms with an increasing share is positive only if they start with a productivity level higher than the industry average. On the other hand, entering (exiting) firms contribute only if they have lower (higher) productivity than the initial industry average. As noted by Bartelsman and Doms (2000), this treatment of births and deaths ensures that the contribution to the aggregate does not arise because the entering plants are larger than exiting plants, but because of productivity differences.

In general, the links between the micro and macro levels of TFP analysis need to be further developed. While this is considered (Hulten, 2001) "... one of the greatest challenges facing productivity analysis today", the empirical literature on productivity estimate is still evolving rapidly in both directions.

3 Frontier vs non-frontier TFP measures

Traditional *Non-Frontier* methodologies shared the common assumption and interpretation that production is always fully efficient: the observed output — either produced by firms/plants or by regions/countries — equates the potential level of production at each moment in time. The formulation originally introduced by Solow (1957), who provided the original analysis of the growth accounting approach (to be discussed in section 4.1), starts from eq. (1.2) (index i dropped) and assumes that TFP growth between time t and $t+1$ is evaluated using the following expression:

$$\frac{TFP_{t+1}}{TFP_t} = \frac{A_{t+1}}{A_t} \tag{3.1}$$

In such derivation observed output is assumed to be equal to the frontier output and the implied measure of TFP growth solely captures shifts in A, i.e. technological change (Grosskopf, 1993). However, such an estimate will be biased in the presence of inefficiency. The recent aggregate *Non-Frontier* literature on developing accounting and growth regressions focuses on explaining the wide differences in economic performance across countries and produces estimates of TFP levels instead of growth rates. This framework focuses on catching-up mechanisms where regions/countries' output is not assumed to be equal to the frontier output and thus, differently from the traditional approach, TFP estimates are interpreted as broad measures of the efficiency with which regions/nations transform their factors of production into output (more on this in sections 4.1 and 5.1), that is, they do not identify TFP with technology. Nevertheless, these studies do not estimate separately the contribution of different sources of TFP change.

An alternative approach has been introduced by scholars within the *Frontier* approach to the measurement of TFP: observed output and potential output might differ due to the presence of technical inefficiency in productive processes of observed units. This implies the adoption of a new perspective with respect to *Non-Frontier* methodologies, since estimated TFP will now explicitly result from a decomposition of productivity growth in technological change and efficiency change. Technological progress is assumed to push the frontier of potential production upward, while efficiency change will reflect the capability of productive units to improve production with a set of given inputs and available technology. Assuming the presence of technical inefficiency in productive processes leads to a discrepancy between observed output and maximum feasible output:

$$Y_t < A_t \; F(\mathbf{X}_t) \tag{3.2}$$

$$Y_{t+1} < A_{t+1} \; F(\mathbf{X}_{t+1}) \tag{3.3}$$

The concept of distance function (Malmquist, 1953; Shephard, 1970) is introduced into the analysis in order to bring observed output up to its efficient level. The output distance function D_t^0 is given by:

$$D_t^0\left(\mathbf{X}_t, \mathbf{Y}_t\right) = \inf\left\{\theta : \left(\mathbf{X}_t, \frac{\mathbf{Y}_t}{\theta}\right) \in \mathbf{S}_t\right\} = \left(\sup\left\{\theta : \left(\mathbf{X}_t, \theta\mathbf{Y}_t\right) \in \mathbf{S}_t\right\}\right)^{-1} \tag{3.4}$$

where $\mathbf{S}_t$ models the transformation of $\mathbf{X}_t \in \mathbb{R}_+^N$ inputs in $\mathbf{Y}_t \in \mathbb{R}_+^M$. The output distance function is hence defined as the reciprocal of the maximum expansion in output vector — given available inputs — such that production is still feasible, i.e., $\left(\mathbf{X}_t, \theta\mathbf{Y}_t\right) \in \mathbf{S}_t$.

The definition of the distance function in (3.4) completely characterizes the technology. Indeed, the following is true:

$$D_t^0\left(\mathbf{X}_t, \mathbf{Y}_t\right) \leq 1 \quad \text{if and only if} \quad \left(\mathbf{X}_t, \theta\mathbf{Y}_t\right) \in \mathbf{S}_t \tag{3.5}$$

and the value taken by the distance function will be 1 if and only if production is technically efficient.

From the concept of distance function and (3.2)-(3.3), it follows that:

$$D_t^0\left(\mathbf{X}_t, Y_t\right) = \frac{Y_t}{A_t F\left(\mathbf{X}_t\right)} \tag{3.6}$$

$$D_{t+1}^0\left(\mathbf{X}_{t+1}, Y_{t+1}\right) = \frac{Y_{t+1}}{A_{t+1} F\left(\mathbf{X}_{t+1}\right)} \tag{3.7}$$

where, at each moment in time, in the presence of technical inefficiency, maximum potential output $A_t F\left(\mathbf{X}_t\right)$ will be equal to the observed output Y_t corrected for the output distance function $D_t^0\left(\mathbf{X}_t, Y_t\right)$.

The TFP indexes at time t and $t+1$ will be given respectively by:

$$TFP_t = \frac{Y_t}{F\left(\mathbf{X}_t\right)} = A_t D_t^0\left(\mathbf{X}_t, Y_t\right) \tag{3.8}$$

and

$$TFP_{t+1} = \frac{Y_{t+1}}{F\left(\mathbf{X}_{t+1}\right)} = A_{t+1} D_{t+1}^0\left(\mathbf{X}_{t+1}, Y_{t+1}\right) \tag{3.9}$$

which yields the following expression for the TFP growth index between the two periods:

$$\frac{TFP_{t+1}}{TFP_t} = \frac{A_{t+1}}{A_t} \frac{D^0_{t+1}\left(\mathbf{X}_{t+1}, Y_{t+1}\right)}{D^0_t\left(\mathbf{X}_t, Y_t\right)} \tag{3.10}$$

that is, in the presence of technical inefficiency, TFP growth is due to technological progress (the first ratio on the right hand side of (3.10)) and change in technical efficiency (the distance functions ratio). Hence, the measure of TFP growth obtained in (3.10) will be equivalent to the one obtained following the growth accounting approach in (3.1) only in the absence of inefficiency, i.e. only if TFP change can be explained solely in terms of technological change. On the other hand, in the presence of inefficiency, measurements of TFP growth based on *Non-Frontier* methods will lead to biased results.

As for the employed estimation techniques, applied works within the frontier approach — dealing with both micro and macro settings — have adopted either deterministic linear programming techniques (DEA) or the stochastic frontier approach (SFA). A less popular deterministic method is the Free Disposal Hull (FDH) model proposed by Deprins, Simar and Tulkens (1984). FDH is a more flexible model with respect to DEA as it only relies on the free disposability assumption of the production set, while DEA also assumes convexity. However, although providing efficiency estimates within a more general framework, FDH has not gained as much success as DEA in applied works.[4]

In coherence with the adopted deterministic/econometric classification criterion, we will focus on the key features of the deterministic frontier approach to the measurement of efficiency and productivity in section 4.3, while we will provide a brief review of the SFA in section 5.3. The implementation of both techniques suffers from limitations but has advantages. The main weakness of the first class of techniques is due to the fact that they are solely based on input and output data and to their deterministic nature, which implies that any discrepancy between actual and potential output is attributed to inefficiency. Any other feasible sources of technical inefficiency, i.e., omitted variables, unobserved measurement errors and stochastic noise are neglected, resulting in a possible upward bias of inefficiency scores.[5] Furthermore, large datasets are required, since the "best practice"frontier obtained with small samples may be too rough an approximation of the real production frontier.[6] On the other hand, DEA does not require the imposition of any functional form for the technology set and allows technical change to vary across decision-making units. SFA is able to distinguish between inefficiency and other possible causes of the discrepancy between observed and maximum potential output. This is made possible by separating two components of the error term in the stochastic production function and the distributional assumptions may significantly affect the results. Moreover, the need to specify a functional form for the production frontier together with the assumption of a common technical change across production units represent two important limitations.

A comprehensive comparison of advantages and limitations of DEA and SFA goes beyond the scope of our survey, which is to underlie their main departures from *Non-Frontier* models. To the best of our knowledge, two recently published volumes will provide the most recent and up-to-date reading to researchers interested in frontier models, that is, Fried *et al.* (2008) and Daraio and Simar (2007). In particular, Greene (2008) reviews the econometric approach to efficiency analysis. Daraio and Simar (2007, pp. 25-42) propose a general taxonomy of frontier models (according to three criteria: specification of the functional form for the production function, the presence of noise in the sample data and the type of data analyzed) and provide an instructive picture of the latest methodological developments within non-parametric frontier approach.

4 Calculating TFP (deterministic methodologies)

4.1 Calculating TFP: growth accounting

Given its popularity and long-term existence, the growth accounting literature shows a plethora of extensions and different results. This approach has been used to estimate TFP at both aggregate and sectoral level

[4]DEA and FDH models are described in detail and compared against each other by Daraio and Simar (2007).

[5]For most recent methodological advances in the field of statistical inference within non-parametric frontier models the reader is referred to Daraio and Simar (2007).

[6]In particular, since the latter is likely to be above the former, this methodology may "read"as "technology regress"something that is, in fact, an efficiency decline. Kumar and Russell (2002) found this result mainly for countries with low capital-labour ratios.

and applied to both within and across countries analysis. The more traditional growth accounting approach focuses more on within rather than across country analysis and decomposes output growth into components using time series data. The first evidence dates back to the fifties with early studies usually finding that a very large fraction of output growth was due to TFP growth. In particular, using US data, Abramovitz (1956) and Solow (1957) found that almost 90% of output growth was associated with TFP growth. Later studies[7] have modified the basic framework and usually find a smaller role of TFP to GDP growth. In this section we provide a brief description of the standard methodology and its more recent developments.

Growth accounting measures TFP indirectly, as the residual component of GDP growth that cannot be explained by the growth of the assumed inputs of production. Let us start the analysis from the standard Hicks neutral aggregate production function described by (1.1). Taking logs (lowercase letters) and derivatives with respect to time (and dropping time dependence) eq. (1.1) becomes:

$$\frac{\dot{y}}{y} = \frac{\dot{a}}{a} + \sum_{1}^{N} \beta_n \frac{\dot{x}_n}{x_n} \tag{4.1}$$

where $(\dot{a}/a)$ is the productivity, or TFP, growth rate and the β_ns are input social marginal products $(F_X X/Y)$. Thus, if we can compute the factor's growth rates and their social marginal products, the TFP growth rate would be easily calculated as a residual, or Solow residual (SR henceforth), from:

$$SR = \frac{\dot{a}}{a} = \frac{\dot{y}}{y} - \sum_{1}^{N} \beta_n \frac{\dot{x}_n}{x_n}. \tag{4.2}$$

The rates of change of TFP represent the change in national income that is not explained by changes in the level of inputs used. Assuming perfect competition and constant return to scale, eq. (4.2) becomes:

$$SR = \frac{\dot{a}}{a} = \frac{\dot{y}}{y} - \sum_{1}^{N} s_n \frac{\dot{x}_n}{x_n}. \tag{4.3}$$

where $s_n = (wX/Y)$ is the fraction of Y used to pay input n. Given the assumptions that $\sum_n \beta_n = 1$, in the Cobb-Douglas case these input shares are constant over time and correspond to the exponents in the production function.

Overall, these assumptions imply that social marginal products can be measured by (observable) factor prices and that to compute SR we only need to calculate the growth rates of output, inputs and, with only L and K (labor and capital respectively) as inputs, the value of the share of physical capital. Estimation is commonly carried out in the two inputs, and the estimate of s_K is usually assumed equal to approximately 1/3, a value based on studies that directly calculate the share of physical capital in aggregate output from national account for developed countries[8] data by computing the remuneration of capital as a share of GDP.

Alternatively, the Solow residual can be measured from growth rates of factor prices, the residual of the dual cost function, rather than from growth rates of factor quantities as in (4.3). This dual approach to growth accounting has been firstly developed by Jorgenson and Griliches (1967). They show that using the equality between output and factor incomes $Y = rK + wL$, the dual or price approach[9] implies:

$$SR = s_K \left(\frac{\dot{r}}{r}\right) + s_L \left(\frac{\dot{w}}{w}\right) \tag{4.4}$$

A second contribution of Jorgenson and Griliches (1967) to growth accounting was to identify the importance of possible errors of aggregation. In general, in this framework, stocks of physical capital are usually generated using the perpetual inventory method and this procedure has been accused to mismeasure actual stocks of physical capital.[10] Nevertheless, even assuming that the perpetual inventory method is appropriate,

[7] See Denison (1985), Maddison (1995), Klenow and Rodriguez-Clare (1997), Hall and Jones, 1999; Aiyar and Feyrer, 2002) among the many others.

[8] For example, 1/3 is the value obtained with US time series data on the capital-share. See Caselli (2005).

[9] For the dual approach see Hsieh (1999).

[10] Pritchett (2000) stresses as measures of K-stocks are sensitive to different assumptions on the K depreciation rate. Further, the assumptions made to calculate the initial capital stock K_0 may be too restrictive, as they usually use the steady state condition of the Solow growth model. While rich countries may approximately satisfy this condition, for poorer countries this assumption is less plausible and this may thus cause their estimate on K_0 to be too high.

errors of aggregation arise when we fail to control for the different quality of factors in the data. In this case, using (4.3) or (4.4) to calculate the Solow residual, unmeasured inputs quality improvements will end up in SR,[11] causing TFP estimates to be upward biased. Conversely, taking into account for differences in the quality of inputs reduces the estimated contribution of TFP growth to output growth.

This problem is also strictly related to the error that arises in aggregating investment goods of different vintages by simply adding together quantities of investment goods of each vintage. There is a long-running debate on how to deal with embodied technological change when we calculate productivity measures that has recently intensified due to the rapid development and impact of new technologies on capital and labour markets.[12] If the quality of investment goods, as measured by the marginal productivity of capital, is not constant over all vintages, this procedure results in aggregation errors. The vintage capital bias in growth accounting may be critical since technological progress tends to be embodied in new forms of capital and different types of capital equipment have different R&D contents.[13] As shown by Caselli and Wilson (2004) this problem is particularly relevant when we compare SR estimates from different countries since there are significant differences in terms of what kinds of capital equipment they use. An appropriate index of capital services may be constructed by treating each vintage of investment goods as a separate commodity. One obvious extension to the standard growth accounting framework is thus to include new and improved measures of factors of productions since (1.1) may be easily modified to include differences in inputs quality. As long as we have measures of each factor's price (and thus each type of factor is weighted by its specific income share) an extended version of (4.3) may correctly measures the TFP growth rate.

Another and more recent extension of the growth accounting methodology is the level of development accounting decomposition.[14] In particular, like growth accounting, development accounting tries to quantify a decomposition of output into inputs contribution and productivity, where the latter is calculated as a residual. This methodology has the advantage of producing estimates of TFP levels instead of estimating their growth rates. Hall and Jones (1999) stress that the focus on TFP levels instead of growth rates has important implications as, theoretically, many growth models imply that differences in levels are the interesting differences to explain and, empirically, cross-country differences in growth rates have often been estimated as mostly transitory. We develop further this issue below in section 5.1. Moreover, levels analysis should capture the differences in long-run economic performances that are relevant to welfare as measured by the consumption of goods and services.

This framework introduces a production function augmented by human capital with Harrod neutral technology.[15] For each country/region i we may write:

$$Y_i = K_i^{\alpha} \left(A_i H_i\right)^{1-\alpha} \tag{4.5}$$

where H is the stock of human capital-augmented labour. The latter may be calculated by $H_i = e^{\phi(E_i)} L_i$ where L is row labour and $\phi(E)$ represents the efficiency of a unit of labour with E years of schooling attendance relative to one with no schooling. Further, $\phi(0) = 0$ and $\phi^{'}(E)$ corresponds to the return to education estimated in a standard individual Mincerian wage equation.[16] A benefit of this framework is that there exists a vast empirical literature offering countries estimates of returns to schooling that can be used

[11] More precisely, "Quality change in this sense occurs whenever the rates of growth of quantities within each separate group are not identical. For example, if high quality items grow faster than items of low quality, the rate of growth of the group is biased downward relative to an index treating high and low quality items as separate commodities." (Jorgenson and Griliches, 1967, p. 259).

[12] On this see Hercowitz (1998), Jorgenson (2005), Greenwood and Krusell (2007) and Oulton (2007).

[13] See Jorgenson (2005) on US and G7 countries and Jorgenson et al. (2007). These studies focus on the role of information technology as the possible driving force behind the acceleration of productivity growth that began in the 1990s.

[14] It has been proposed by Hall and Jones (1999) and Klenow-Rodriguez Clare (2001) among others. Recently Aiyar and Dalgaard (2005) have discussed on how to extend the dual approach to this level accounting methodology.

[15] In this case the production function is augmented by human capital. Mankiw et al. (1992) have been the first to use the less restrictive case where $Y = K^{\alpha} H^{\beta} \left(AL\right)^{1-\alpha-\beta}$. In terms of equation (1.1), the Harrod-neutral technology implies $Y = F\left[A, K, L\right] = \tilde{F}\left(K, AL\right)$ and $g = \left(\frac{F_L L}{Y}\right)\left(\frac{\dot{a}}{a}\right)$.

[16] In its standard form the Mincerian wage equation is defined by $\ln w_i = X_i^{'}\gamma + \phi E_i$, where X is a set of demographic controls. For more on this see Mincer (1974).

for the calibration exercise.[17] Equation (4.5) can then be rearranged as:

$$\frac{Y_i}{L_i} = A_i \left(\frac{K_i}{Y_i}\right)^{\frac{\alpha}{(1-\alpha)}} \left(\frac{H_i}{Y_i}\right) \tag{4.6}$$

In order to estimate A_i, that is, TFP's levels, from eq. (4.6) we thus need data on output, labour, educational attainments, physical capital, capital shares and returns to human capital.

Overall, the recent empirical literature on growth or development accounting estimates a smaller role of efficiency to output growth with respect to early studies. Still, the contribution of efficiency is usually estimated as highly significant and the consensus view in aggregate cross country data is that TFP is at least as important as factors of productions to explain differences in economic performance across countries.[18] Nevertheless, there are exceptions to this point of view. In particular, Young's studies on East Asian countries[19] have been influential in claiming that GDP growth in is mostly explained by inputs accumulation rather than TFP growth. More recently, using US and G7 countries data Jorgenson (2005) found the contribution of inputs exceeds that of TFP, and similar results may be also found in Baier, Dwyer and Tamura (2006).

As well as large consensus,[20] growth accounting techniques have also received some criticism. The first concerns the imposition of too many assumptions such as constant returns to scale and perfect competition.[21] As shown by Hall (1988), relaxing the perfect competition assumption produces a difference between the estimated Solow residual from (4.3), $\widetilde{SR}$, and true productivity growth, SR. Defining $y = \frac{Y}{K}$ and $l = \frac{L}{K}$, the Solow residual from eq. (4.3) may be rewritten as:

$$\Delta y = \mu\theta\Delta l + SR \tag{4.7}$$

where θ is the labour factor share $\mu = \frac{P}{MC}$ is the markup ratio, with MC the marginal cost, and P the price level. Under perfect competition, $P = MC$ and the Solow residual correspond to the true growth rate of productivity, that is, may be correctly estimated from the usual formula $SR = \Delta y - \theta\Delta l$, where $\Delta y = (\Delta \log y)$, and $\Delta l = (\Delta \log l)$. Conversely, if firms have market power,[22] the Solow residual estimated by $\widetilde{SR} = \Delta y - \theta\Delta l$ will be different from the true growth rate of productivity.

Similar problems may be identified when we relax the assumption of constant returns to scale in the production function. Increasing returns and spillovers may be represented by:

$$Y_i = AK_i^{\alpha} K^{\beta} L_i^{1-\alpha} \tag{4.8}$$

Equation (4.8) represents firm i production function that depends not only on private inputs, K_i and L_i, but also on aggregate capital stock K. If $0 < \alpha < 1$ and $\beta > 0$ this represents a production function with CRS in the private inputs and positive spillovers. Assuming that in equilibrium each firm adopts the same capital-labour ratio, aggregating across firms it can be shown that the economy-wide production function can be written as:

$$Y = AK^{\alpha+\beta} L^{1-\alpha} \tag{4.9}$$

Therefore, even in this case the estimated standard Solow residual $\widetilde{SR}$ would be biased upwards, as in the previous case:

$$\widetilde{SR} = SR + \beta\frac{\dot{K}}{K} \tag{4.10}$$

[17] For example, using survey evidence, Hall and Jones (1999) calculate H assuming that is piecewise linear with slope 0.13 for$E \leq 4$, 010 for $4 < E \leq 0.07$ for $E > 8$.

[18] "A sentence commonly used to summarize the existing literature sounds something like differences in efficiency account for at least 50% of differences in per capita income"(Caselli, 2005, p. 2).

[19] See, for example, Young (1995).

[20] "...we take a standard neoclassical approach....This is a natural benchmark. It ignores externalities from physical and human capital. We believe there is little compelling evidence of such externalities..."(Hall and Jones, 1999, p. 89).

[21] Another disputed assumption of the growth accounting approach that we do not discuss here is the absence of factor hoarding. For more on this, see Roeger (1995).

[22] And under constant returns to scale.

where SR is the true growth rate of TFP. Standard growth accounting techniques have also been criticized since they are not informative about casual relationships that connect the different inputs to growth. For this reason, many regard them as an "ad hoc measure with little economic content"(Greenwood and Krusell, 2007, p. 1301). In particular, accounting decompositions may easily attribute to capital accumulation something that should be attributed to technological progress and vice-versa. This is certainly true if capital is endogenous and responds to technological progress or if improvements in educational attainment have indirect effects on output through changes in labour force participation or R&D and, thus, on TFP growth.[23]An alternative approach may be found in the "quantitative theory". This calibration approach defines the impulses affecting the economy but has the drawback of being significantly more burdensome than growth accounting as requires a fully specified general equilibrium model (Greenwood and Krusell, 2007).

Finally, another drawback of traditional aggregate studies is that the role of the sectorial composition of output is ruled out by assumption since, as said above, they assume GDP to be produced by a single sector. Thus, it is usually difficult to disentangle how much of TFP differences across countries are due to sectorial specialization rather than to other factors that make some countries less efficient than others. More precisely, even assuming that within-sector productivity are identical, aggregate differences in TFP across countries may be explained by cross-countries differences in the sectorial composition of the economy. Note that these two hypotheses have different policy implications (Caselli, 2005). In fact, if TFP differences across countries are due to differences in sectorial composition we should focus mainly on barriers to the mobility of factors across sectors. Conversely, if sectorial composition does not contribute to explain TFP differences we should focus, more broadly, on barriers to technology (or work practices) adoption across countries. The aggregate approach usually focuses only on the latter.

In sum, even if nowadays there is a growing preference for econometric modelling of the factors causing productivity change (Hulten, 2001), it is probably fair to say that TFP estimated as the Solow residual by this aggregative deterministic approach"... should be understood as a diagnostic tool, just as medical tests can tell one whether or not he is suffering from a certain ailment, but cannot reveal the causes of it. This does not make the test any the less useful"(Caselli, 2005).

4.2 Calculating TFP: Index Numbers issues

In this section we briefly describe the basic idea of these index numbers issues, borrowing from Van Biesebroeck (2007) and Hulten (2001). The underlying concept is the same illustrated in section 4.1 for the SR. As seen above, the Solow residual (SR) is in fact a measure of efficiency that uses a deterministic index number approach with TFP computed directly from prices and quantities. The main indexes used to measure productivity are the Laspeyr's, the Paasche, the Fisher and the T*ö*rnqvist. We only briefly describe the latter, leaving a full description of these indexes to other surveys.[24]

In continuous time equation (4.3) is exact and this may represent a problem since data to estimate SR are in discrete time. In order to avoid possible biases, it is thus necessary to find alternative discrete-time approximations to (4.3). Consider equation (4.3) and replace s_n with the average of current and lagged factor shares ($\frac{s_{n,t}+s_{n,t-1}}{2}$). The T*ö*rnqvist Index Numbers in this equation has been shown (Diewert, 1976) to give an exact expression for the second term in (4.3), under the condition that the production function is translog.

Index numbers are also used in the frontier approach described below. According to Caves et al. (1982b), the Malmquist productivity index (4.17) (described in the following section) exactly equals the difference between a T*ö*rnqvist output index and the corresponding input index with a scale factor to account for non-constant returns to scale:

$$\frac{\dot{a}}{a} = \frac{\dot{y}}{y} - \sum_{1}^{N} \left(\frac{s_{n,t} + s_{n,t-1}}{2} \right) \frac{\dot{x}_n}{x_n} - \sum_{1}^{N} \left(\frac{s_{n,t}(1-\beta_{n,t}) + s_{n,t-1}\beta_{n,t-1}}{2} \right) \frac{\dot{x}_n}{x_n}. \tag{4.11}$$

[23]On this, see in particular Barro and Sala-i-Martin (2004) and Temple (2001).

[24]"The Laspeyr's index is the value of period 1 output measured using period 0 prices divided by the value of period 0 output measured using period 0 prices. The Paasche index measures the value of output in the two periods using period 1 prices. The Fisher index is the average of the Laspeyre's and Paasche indexes."(Carlaw and Lipsey, 2003).

That is, Caves et al. (1982b) show that the Törnqvist index has a more general validity and allows for technical change that is not Hicks-neutral and variable returns to scale in production.

With micro data, the same is true for multilateral productivity comparisons (i.e. subscripts t and $t+1$ can be replaced by i and j, denoting firms). In this case, as Törnqvist indices are not transitive, each firm is compared with the average firm (the firm with average output and input shares):

$$a_{it} - \bar{a}_t = (y - \bar{y}_t) - \sum_1^N \left(\frac{s_{it} + \bar{s}_t}{2} \right) (x_{n,it} - \bar{x}_{n,t}) \,. \tag{4.12}$$

Evidently, Index Numbers are straightforward to compute. The disadvantages associated with them are those illustrated in section 4.1 for Growth Accounting. It is worth noting that, since the aggregation is exact only if the production function is translog, the procedure cannot be regarded as "fully" non-parametric. However, under different assumptions about the production function, the Törnqvist index above can still be thought of as a "second order"approximation.

4.3 Calculating (and decomposing) TFP: DEA and the Malmquist index

Figure 2 provides the intuition for the use of distance functions in measuring efficiency and productivity, as described in section 3.

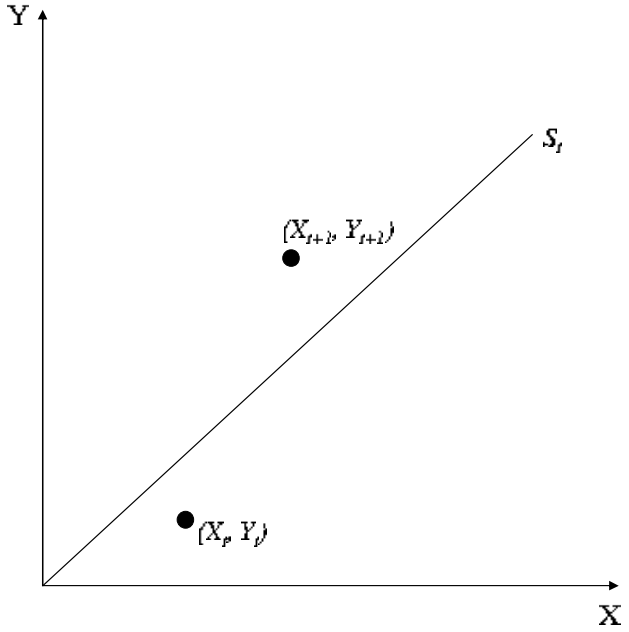

Figure 2: The Malmquist Index

Since $(X_{t+1}, Y_{t+1}) > (X_t, Y_t)$, productivity has increased over time. A Malmquist productivity index quantifies productivity growth by taking technology at time t — S_t — as a benchmark and by comparing the distances of (X_{t+1}, Y_{t+1}) and (X_t, Y_t) to S_t. Such distances can be measured either vertically or horizontally. Indeed, as pointed out by Caves et al. (1982a), productivity differences over time may be interpreted in two ways: as changes in maximum output conditional on a given level of inputs (output-oriented productivity indexes) or as changes in minimum input requirements, conditional on a given level of output (input oriented productivity indexes). Given the assumption one has made on the producer orientation, the ratio of these two distances will provide the measure of productivity change. In the example of Figure 2 such a ratio is greater than 1 both in the input and in the output oriented cases.

The Malmquist productivity index introduced by Caves et al. (1982a) uses the two distance functions defined above in (3.6) and (3.7) and the two following mixed period distance functions:

$$D_t^0\left(\mathbf{X}_{t+1}, Y_{t+1}\right) = \frac{Y_{t+1}}{A_t F\left(\mathbf{X}_{t+1}\right)} \tag{4.13}$$

$$D_{t+1}^0\left(\mathbf{X}_t, Y_t\right) = \frac{Y_t}{A_t F\left(\mathbf{X}_t\right)} \tag{4.14}$$

On the basis of the above output distance functions, Caves et al. (1982a) define their output oriented Malmquist productivity indexes for period t and $t+1$ as

$$M_t^0\left(\mathbf{X}_t, Y_t, \mathbf{X}_{t+1}, Y_{t+1}\right) = \frac{D_t^0\left(\mathbf{X}_{t+1}, Y_{t+1}\right)}{D_t^0\left(\mathbf{X}_t, Y_t\right)} \tag{4.15}$$

evaluated with respect to technology at time t; and

$$M_t^0\left(\mathbf{X}_t, Y_t, \mathbf{X}_{t+1}, Y_{t+1}\right) = \frac{D_{t+1}^0\left(\mathbf{X}_{t+1}, Y_{t+1}\right)}{D_{t+1}^0\left(\mathbf{X}_t, Y_t\right)} \tag{4.16}$$

evaluated with respect to technology at time $t+1$.

In order to avoid the subjective choice of the reference technology, an additional productivity index was defined as the geometric mean of (4.15) and (4.16):[25]

$$M_t^0\left(\mathbf{X}_t, Y_t, \mathbf{X}_{t+1}, Y_{t+1}\right) = \left[\frac{D_t^0\left(\mathbf{X}_{t+1}, Y_{t+1}\right)}{D_t^0\left(\mathbf{X}_t, Y_t\right)} \frac{D_{t+1}^0\left(\mathbf{X}_{t+1}, Y_{t+1}\right)}{D_{t+1}^0\left(\mathbf{X}_t, Y_t\right)}\right]^{\frac{1}{2}} \tag{4.17}$$

Starting from their seminal contribution, the literature following Caves et al. (1982a) has been dealing with two main issues. From the theoretical point of view, research has been devoted to the definition of possible decompositions of the Malmquist index. The aim is to measure the contribution of different sources of productivity change, that is, technological change, efficiency change, scale economies and changes occurred to the environment faced by producers. Secondly, the Malmquist index is a theoretical index based on the definition of distance functions which in turn are defined on unknown technologies. Hence, scholars have been dealing with empirical implementations aiming at approximating the Malmquist index (and its components, as delivered in theoretical studies). We focus on the decomposition of the Malmquist index which allows for decomposing productivity gains in technological progress and efficiency change and on its empirical estimation based on DEA.

Färe et al. (1994a) provide the following decomposition of (4.17) under the assumption of CTRS:

$$M_t^{0c}\left(X_t, Y_t, X_{t+1}, Y_{t+1}\right) = \frac{D_{t+1}^{0c}\left(X_{t+1}, Y_{t+1}\right)}{D_t^{0c}\left(X_t, Y_t\right)} \left[\frac{D_t^{0c}\left(X_{t+1}, Y_{t+1}\right)}{D_{t+1}^{0c}\left(X_{t+1}, Y_{t+1}\right)} \frac{D_t^{0c}\left(X_t, Y_t\right)}{D_{t+1}^{0c}\left(X_t, Y_t\right)}\right]^{\frac{1}{2}} \tag{4.18}$$

The first ratio on the right hand side of (4.18) represents the change in technical efficiency between period t and period $t+1$, while the term in brackets measures the shift in technology between the two periods. M^{0c} greater than 1 indicates that productivity has risen between period t and $t+1$ and this can be explained in terms of technical efficiency improvement and/or technological progress. A value of the index smaller than 1, will indicate a TFP slowdown between the two periods. It is important to notice that the two components may move in opposite directions. For instance, if neither input ($X_t = X_{t+1}$) nor output ($Y_t = Y_{t+1}$) change between the two periods, M^{0c} will be equal to 1 and technical change and efficiency change will be reciprocal but not necessarily both equal to 1.

The graphical example provided by Färe et al. (1994b) gives the key intuition of the decomposition of the Malmquist index in (4.18).

In Figure 3, S_t and S_{t+1} represent the production frontiers at time t and $t+1$ respectively. In this simple graphical example, the level of production observed at time t is not efficient. Indeed, observed output $Y = a$

[25] Notice that the Malmquist index does not require price information and assumptions on the structure of the technology and the behavior of producers as the Fisher and Törnqvist indexes do.

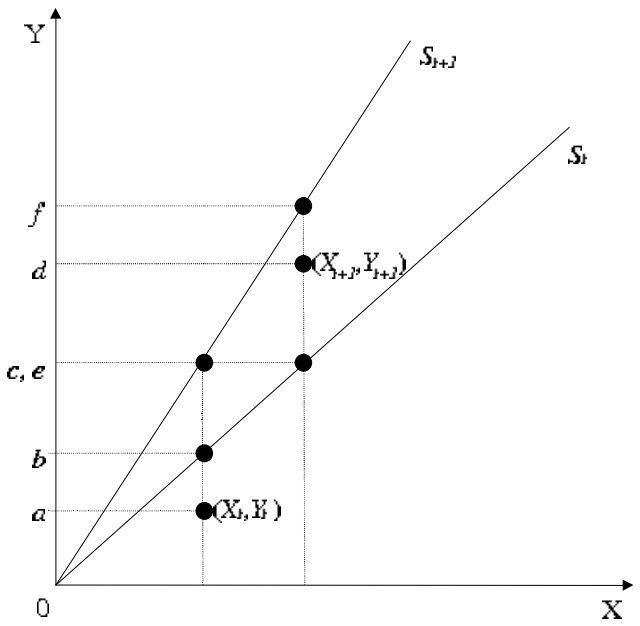

Figure 3: The Decomposition of the Malmquist Index

is below the frontier and maximum potential output is given by $Y = b$. According to the definition in (3.6), the output distance function is defined by the ratio $0a/0b < 1$. Since $S_t \subset S_{t+1}$, a technical advance has taken place between time t and $t+1$. However, observed production at time $t+1$ ($Y = d$) is still technically inefficient and the output distance function in the new period is equal to $0d/0f < 1$.

Given the two distance functions $0a/0b$ and $0d/0f$ and recalling the decomposition of the Malmquist productivity index in (4.18), we can define the ratio:

$$\frac{0d/0f}{0a/0b} = \text{efficiency change} \tag{4.19}$$

which for values greater than one will indicate that production is closer to its efficient level in period $t+1$ than in period t, i.e. an efficiency improvement occurred between the two periods. From Figure 3, we can also derive the graphical counterparts of the two mixed output distance functions in (4.13) and (4.14), which are necessary to obtain the decomposed Malmquist index of our example:

$$D_t^0\left(X_{t+1}, Y_{t+1}\right) = \frac{Y_{t+1}}{A_t F\left(X_{t+1}\right)} = \frac{0d}{0e} \tag{4.20}$$

$$D_{t+1}^0\left(X_t, Y_t\right) = \frac{Y_t}{A_{t+1} F\left(X_t\right)} = \frac{0a}{0c} \tag{4.21}$$

where the ratio in (4.20) represents the highest proportional change in output requirements to make (X_{t+1}, Y_{t+1}) feasible in relation to the technology at time t. On the other hand, the ratio defined in (4.21) indicates the highest proportional change in output requirements to make (X_t, Y_t) feasible in relation to the technology at time $t+1$.

Finally, the Malmquist productivity index can be expressed as:

$$M_t^{0c}\left(X_t, Y_t, X_{t+1}, Y_{t+1}\right) = \frac{0d}{0f} \Big/ \frac{0a}{0b} \left[\frac{0d/0e}{0d/0f} \frac{0a/0b}{0a/0c}\right]^{1/2} = \frac{0d}{0f} \Big/ \frac{0a}{0b} \left[\frac{0f}{0e} \frac{0c}{0b}\right]^{1/2} \tag{4.22}$$

Moving to the empirical implementation of the Malmquist index — in the general case of multiple outputs

— the following CRTS frontier technology is constructed from the data using DEA:

$$\mathbf{S}_t = \left\{ (\mathbf{X}_t, \mathbf{Y}_t) : Y_t^m \leq \sum_{k=1}^{K} z_{k,t} Y_{k,t}^m; \sum_{k=1}^{K} z_{k,t} X_{k,t}^n \leq X_t^n; z_{k,t} \geq 0 \right\} \tag{4.23}$$

with $k = 1, ..K$ production units using $n = 1, ..., N$ inputs $X_{k,t}$ to produce $m = 1, .., M$ outputs at each t, and the terms $z_{k,t}$ stand for weights on each unit of production. The assumption of CRTS can be relaxed by imposing the restriction $\sum_{k=1}^{K} z_{k,t} \leq 1$, which will give the case of variable returns to scale (VRTS).

In order to calculate the Malmquist productivity index for each producer at each time t, the distance functions similar to those illustrated in (3.6), (3.7), (4.13) and (4.14) have to be evaluated. The distance function in (3.6) is evaluated by solving the following linear programming problem for each producer k':

$$\left[D_t^0 \left(\mathbf{X}_{k',t}, \mathbf{Y}_{k',t} \right) \right]^{-1} = \max \theta_{k'} \tag{4.24}$$

subject to

$$\sum_{k=1}^{K} z_{k,t} Y_{k,t}^m \geq \theta_{k'} Y_{k',t}^m \tag{4.25}$$

$$\sum_{k=1}^{K} z_{k,t} X_{k,t}^n \leq X_{k',t}^n \tag{4.26}$$

$$z_{k,t} \geq 0 \tag{4.27}$$

The evaluation of $D_{t+1}^0(\mathbf{X}_{t+1}, \mathbf{Y}_{t+1})$ will imply solving a linear programming problem such as the one in (4.24)-(4.27), transposing subscripts t with $t + 1$. The mixed distance function in (4.13) is obtained by solving the following problem:

$$\left[D_t^0 \left(\mathbf{X}_{k',t+1}, \mathbf{Y}_{k',t+1} \right) \right]^{-1} = \max \theta_{k'} \tag{4.28}$$

subject to

$$\sum_{k=1}^{K} z^m Y_{k,t}^m \geq \theta_{k'} Y_{k',t+1}^m \tag{4.29}$$

$$\sum_{k=1}^{K} z_{k,t} X_{k,t}^n \leq X_{k',t+1}^n \tag{4.30}$$

$$z_{k,t} \geq 0 \tag{4.31}$$

The solution to the linear programming problem in (4.28)-(4.31) with subscripts t and $t + 1$ transposed will give $D_{t+1}^0(\mathbf{X}_t, \mathbf{Y}_t)$.[26]

5 Estimating TFP (econometric methodologies)

5.1 Estimating TFP from macro data: growth regressions

The *growth regressions approach*[27] originates from the vast empirical literature on growth and convergence that has started in the mid-eighties with the resurgence of the endogenous growth literature. This debate is strictly related to the question of whether TFP convergence is taking place and under what conditions. Indeed, as Bernard and Jones (1996) put it, one of the main controversies in the empirical growth literature is to identify "how much of the convergence that we observe is due to convergence in technology versus convergence in capital-labour ratios" since convergence may be the result of three different mechanisms:

[26] The most widely used specialized computer software for the implementation of DEA models is DEAP 2.1 (Coelli, 1996). Hollingsworth (2004) systematically reviews available DEA computer softwares, providing useful information on their advantages and limitations.

[27] See Bosworth and Collins (2003) and Jorgenson (2005).

convergence due to capital accumulation, convergence due to technology transfer (catch up), and convergence due to both.

Unlike growth accounting methodologies this is a model-based approach to estimate TFP from aggregate data that stem from the seminal Mankiw et al. (1992) contribution (hereafter MRW). One advantage of this approach is that TFP is not estimated as a residual and TFP measures should then be purged from noise. Secondly, this approach does not need to use data on the stocks of physical capital that, as said above, are likely to be characterized by significant measurement error problems.

MRW analysis represents an extension of the standard Solow Swan model. Its main contribution is to identify a structural equation to estimate the hypothesis of cross-countries conditional convergence that states that if preferences, technology or other characteristics differ across countries, then in the long run there should be convergence to the same per capita output growth rate. Countries, however, do not necessarily converge towards the same capital-labor ratio and output per capita level.[28]

In MRW the production function is given by a Harrod neutral technology production function $Y = K^{\alpha}(AL)^{1-\alpha}$ with $0 < \alpha < 1$, where AL is defined as "effectiveness of labour". In this model the technology (or disembodied productivity) is a public good that evolves exogenously: that is, the growth rate of the technology frontier is constant and identified by g. Moreover, the depreciation of capital is proportional at rate δ, while n represents the exogenous growth rate of the labour force. It is also assumed that the economy invests a constant proportion of income, s and that the capital stock of an economy evolves according to:

$$\dot{\widetilde{k}} = sf(\widetilde{k}) - (n + g + \delta)\widetilde{k} \tag{5.1}$$

where $\widetilde{k} = \frac{K}{AL}$ and $\widetilde{y} = \frac{Y}{AL}$. From the transitional dynamics of the Solow model and after standard substitutions, a log-linear approximation of equation (5.1) around the steady state implies that:

$$\ln \widetilde{y}(t) - \ln \widetilde{y}(0) = (1 - e^{-\lambda\tau})(\ln \widetilde{y}^{*} - \ln \widetilde{y}(0)) \tag{5.2}$$

where $\widetilde{y}(0)$ denotes income per effective worker at some initial point of time and the asterisk denotes variables at steady state. This equation indicates that when an economy starts from a level of income in efficiency units lower than its steady state level $\widetilde{y}$, we should observe a positive rate of growth of $\widetilde{y}$ with λ representing the speed of adjustment towards $\widetilde{y}^{*}$. The effect of diminishing returns implies that growth due to capital accumulation vanishes in the long-run. If certain assumptions are satisfied, the process of (absolute) convergence towards the long-run equilibrium may result in a tendency towards convergence in per capita income levels among economies. MRW identify an explicit expression for the steady state of per capita income where:

$$\ln y^{*} = \ln A(0) + gt + \frac{\alpha}{1-\alpha}\ln s - \frac{\alpha}{1-\alpha}\ln(n + g + \delta) \tag{5.3}$$

precisely defines the steady state level of the log of income in per capita terms. In (5.3) the initial level of TFP, $A(0)$, is, together with the saving rate and $(n + g + \delta)$, a determinant of y^{*}. Thus, the convergence equation (5.2) becomes:

$$GRy_i = c + by_i(0)_i + \frac{\alpha}{1-\alpha}\ln s_i - \frac{\alpha}{1-\alpha}\ln(n_i + g + \delta) + \epsilon_i \tag{5.4}$$

where GRy_i is the growth rate of per capita (or per worker) GDP and $b = (1 - e^{-\lambda\tau})$. In (5.4) $A(0)$ represents the unobservable TFP component that differs across countries through $\ln A(0)_i = c + \varepsilon_i$, where c is a constant and country-specific factors are simply considered as part of the error term. Therefore, eq. (5.4) may be conveniently estimated by OLS since differences in TFP levels across countries, ε_i, are assumed to be a purely random phenomenon.

Islam (1995) firstly extends this framework by assuming productivity to vary non randomly across individual economies and introduces the idea that the unobservable differences in TFP are correlated with other

[28] This is different from the so called absolute convergence hypothesis that implies that if a group of countries differs only by their initial capital-labor ratios they will eventually converge to the same per capita output level.

regressors and may be directly estimated applying suitable panel fixed effects methodologies to:

$$y_{it} = \beta y_{it-1} + \sum_{j=1}^{2} \gamma^j x_{it}^j + \eta_t + \mu_i + \upsilon_{it} \tag{5.5}$$

where this is the so called level-specification of eq.(5.4),[29] υ_{it} is the transitory term that varies across countries, and the remaining terms are:

$$x_{it}^1 = \ln(s_{it}) \tag{5.6}$$

$$x_{it}^2 = \ln(n_{it} + g + \delta) \tag{5.7}$$

$$\gamma^1 = (1-\beta)\frac{\alpha}{1-\alpha} \tag{5.8}$$

$$\gamma^2 = -(1-\beta)\frac{\alpha}{1-\alpha} \tag{5.9}$$

$$\mu_i = (1-\beta)\ln A(0)_i \tag{5.10}$$

$$\eta_t = g(t_2 - \beta t_1) \tag{5.11}$$

As before, in equation (5.5) the coefficient $\beta = e^{-\lambda\tau}$ enables to recover the speed of convergence parameter λ, while $\tau = (t_2 - t_1)$, is the time span considered.[30] It is also possible to estimate a restricted version of the model imposing $\gamma^1 = -\gamma^2$:

$$y_{it} = \beta y_{it-1} + \psi x_{it} + \eta_t + \mu_i + \upsilon_{it} \tag{5.12}$$

where $x_{it} = \ln(s_{it}) - \ln(n_{it} + g + \delta)$. In both (5.5) and (5.12), from μ_i it is possible to calculate $A(0)_i$ which is considered a broad measures of the efficiency with which regions/nations transform their factors of production into output as this term should control for technology together with various unobservable factors like institutions or climate. Productivity measures can thus be computed through:

$$TFP_i = \hat{A}(0)_i = \exp\left(\frac{\hat{\mu}_i}{1-\hat{\beta}}\right) \tag{5.13}$$

Different methodologies have been proposed to estimate $\hat{\mu}_i$ and $\hat{\beta}$ from eq. (5.5) and there is no agreement on which estimator suits the case better. In the next section we will describe the main characteristics of these methodologies and discuss arguments for and against the proposed estimators. As we shall see, apart from one, most fixed effects estimators usually transform data in order to eliminate μ_i from eq. (5.5). In this case, estimates of individual intercepts and, through equation (5.13), of TFP_i, may be recovered by:

$$\hat{\mu}_i = \bar{y} - \hat{\beta}\bar{y}_{it-1} - \sum_{j=1}^{2} \hat{\gamma}^j \bar{x}_{it}^j \tag{5.14}$$

where the overbar refers to time averages. This approach to TFP estimates has been criticized since equation (5.5) rules out the hypothesis of technological catching-up assuming instead that differences in TFP are constant and that all economies grow at the same technological rate, g, whatever their initial level of technological knowledge. Conversely, catching-up may be described as a process where the growth rate of technology is proportional to the current gap between the world technology frontier and the technology level currently adopted in an economy. In this case, during the transition, lagging economies would grow faster than g and the technology gap between the leader and a given follower should decrease.

[29] Differently from eq.(5.4) in this specification the dependent variable is the logarithm of the level of per capita GDP and $\beta = b - 1$ where, b is the coefficient of the lagged dependent variable in (5.4).

[30] Most studies use a five year time span to control for the business cycle.

However, if the time dimension of the panel is sufficiently long, equation (5.5) can be separately estimated for different subperiods. This would enable the researcher to obtain estimates of cross-country TFP levels at different points in time and test for the presence of technological convergence comparing the distribution of TFP values obtained over different periods. An important feature of this methodology is that catching up is tested separately from the test for the presence of convergence due to capital accumulation detected by $\hat{\beta}$. Finally, this methodology may be considered as a first step towards the analysis of the determinants of TFP dynamics.[31]

5.1.1 TFP estimates and dynamic panel data problems.

In section 5.2 we describe how and why the use of panel data estimators has been proposed in the micro data framework. Unfortunately, this description does not work for aggregate data analysis since there are significant differences in terms of the estimation strategy between the macro and the micro framework. First of all, in aggregate datasets individual units are countries or regions so that we cannot consider observations to be randomly drawn from a large population as in firms samples.

Secondly, unlike in the previous case, aggregate empirical analysis typically uses panels characterized by a relatively small N and a reasonably sized T sometimes called time series cross-section (TSCS) dataset and this implies significant differences between the micro and the macro approach with respect to the asymptotic properties of fixed effects estimators. In particular, in a typical micro panel T is usually short and assumed as fixed, and this implies that the within group (WG) estimator is a consistent estimator only when regressors are strictly exogenous. This is not the case for TSCS panels since unlike the firm/plant level approach, with TSCS datasets the asymptotic analysis must be considered in the time series dimension of the data, while N (the number of countries) may be considered as fixed. Amemiya (1967) showed that, when the relevant asymptotic is in the direction of $T \to \infty$, the WG estimator of a dynamic panel such as (5.5) is in fact consistent and asymptotically equivalent to Maximum Likelihood. Empirical studies of the aggregate approach exploit this result.

However, while it is true that LSDV-WG estimator is consistent for macro panels, small sample problems may still badly affect these estimates. The finite sample properties of various methodologies for dynamic panel data models as in (5.5) need certainly to be further investigated but, as we shall see, some results by Monte Carlo simulations are already present and used by researchers.

In the previous section we have seen that to estimate cross-country productivity we are directly interested in $\hat{\beta}$, the AR(1) variable coefficient, and the estimated individual intercepts $\hat{\mu}_i$. Since we need our fixed effect estimates to calculate TFP from (5.5), the Least Square with Dummy Variable (LSDV) estimator represents an obvious choice. The LSDV and the WG estimators have the same characteristics and produce exactly the same parameters and standard error estimates. For this reason we will sometimes refer to LSDV-WG estimates without discriminating between the two methodologies. Nevertheless, differently from WG, LSDV includes a dummy variable for each cross-sectional observation along with the explanatory variables, and is based on the direct OLS estimation of equation (5.5). Excluding the time dummies and assuming that x is exogenous it can be shown that the LSDV model of (5.5) can be written as:

$$y = W\delta + (I_n \iota_t)\mu + \varepsilon \tag{5.15}$$

where $W = \left[y^{-1} \vdots X \right]$ is the data matrix that includes the lagged dependent variable, $\delta = \left[\beta, \gamma^{'} \right]$ is a vector of coefficients and $(I_n \iota_t)\mu$ is an expression for unit-level heterogeneity across all time periods. In fact, from (5.5) individual's observations can be rewritten by stacking the T periods for every country in column vectors as:

$$y_i = \begin{bmatrix} y_{i1} \\ \vdots \\ y_{iT} \end{bmatrix}, \; y_i^{-1} = \begin{bmatrix} y_{i0} \\ \vdots \\ y_{iT-1} \end{bmatrix}, \; X_i = \begin{bmatrix} x_{i1}^{'} \\ \vdots \\ x_{iT}^{'} \end{bmatrix}, \; \varepsilon_i = \begin{bmatrix} \varepsilon_{i1} \\ \vdots \\ \varepsilon_{iT} \end{bmatrix}, \; \iota_T = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$$

[31] On this see Islam (2003) and Di Liberto et al. (2008).

and for each country in the panel we may write:

$$y_i = \beta y^{-1} + X_i\gamma + \mu_i\iota_t + \varepsilon_i \tag{5.16}$$

Vectors for all $i = 1, 2, ..., N$ individuals are defined by:

$$y = \left[\begin{array}{c} y_1 \\ \vdots \\ y_N \end{array}\right]_{NTx1}, y^{-1} = \left[\begin{array}{c} y_1^{(-1)} \\ \vdots \\ y_N^{(-1)} \end{array}\right]_{NTx1}, X = \left[\begin{array}{c} x_{i1}^{'} \\ \vdots \\ x_{iN}^{'} \end{array}\right]_{NTxk}, \varepsilon = \left[\begin{array}{c} \varepsilon_{i1} \\ \vdots \\ \varepsilon_{iN} \end{array}\right]_{NTx1}, \mu = \left[\begin{array}{c} \mu_1 \\ \vdots \\ \mu_N \end{array}\right]_{Nx1}$$

The LSDV estimator is not usually introduced in micro data analysis since applying OLS to (5.15) with a large number of cross-sectional observations may result in too many parameters to estimate. In this case the within group (WG) transformation is preferable.

Islam (1995) has been the first to exploit Amemiya (1967) results and provide cross-country aggregate TFP levels estimates from (5.13) within the growth-convergence framework using LSDV-WG. In his study Islam also compares the results obtained by LSDV-WG with that obtained using the Minimum Distance (MD) estimator firstly proposed by Chamberlain (1982). This methodology assumes that individual effects are correlated with the included exogenous variables as in:

$$\mu_i = k_0 + k_1 x_{i1} + k_2 x_{i2} + ... + k_T x_{iT} + \zeta_i \tag{5.17}$$

and

$$y_{i0} = \phi_0 + \phi_1 x_{i1} + ... + \phi_T x_{iT} + \varphi_i \tag{5.18}$$

That is, the fixed effect depends linearly on all leads and lags of the exogenous variable x, $E\left[\zeta_i | x_{i1}...x_{iT}\right] = 0$, and $E\left[\varphi_i | x_{i1}...x_{iT}\right] = 0$. To apply the MD estimator on (5.12) Islam (1995) firstly need to replace μ_i by (5.17) in eq. (5.16) and, secondly, replace by repeated substitutions the lagged dependent variable by its initial value, y_{i0} defined by eq. (5.18). For example, assuming $T = 3$ we would then obtain the following reduced-form equations:[32]

$$\begin{array}{rcl} y_1 & = & \psi x_1 + \beta y_0 + \mu + \upsilon_1 \\ y_2 & = & \beta\psi x_1 + \psi x_2 + \beta^2 y_0 + (\mu + \beta\mu) + (\upsilon_2 + \beta\upsilon_1) \\ y_3 & = & \beta^2\psi x_1 + \beta\psi x_2 + \psi x_3 + \beta^3 y_0 + (\mu + \beta\mu + \beta^2\mu) + (\upsilon_3 + \beta\upsilon_2 + \beta^2\upsilon_2) \end{array} \tag{5.19}$$

In sum, Chamberlain suggests reducing the problem of estimating eq. (5.12), a single equation model involving two-dimensions, into a one-dimensional problem of estimating a T-variate regression model with cross-sectional data, that is combining all equations of a single individual into one system of equations. In order to obtain Chamberlain's MD estimator we must first obtain the unconstrained reduced-form coefficient matrix. In our example this matrix would be defined by:

$$\Pi = \left(\begin{array}{ccc} \psi & 0 & 0 \\ \beta\psi & \psi & 0 \\ \beta^2\psi & \beta\psi & \psi \end{array}\right) + \left(\begin{array}{c} \beta \\ \beta^2 \\ \beta^3 \end{array}\right)\phi^{'} + \left(\begin{array}{c} 1 \\ 1+\beta \\ 1+\beta+\beta^2 \end{array}\right)k^{'} \tag{5.20}$$

Each element of the Π-matrix is a function of the structural-form coefficients that can be summarized by a vector $\vartheta^{'} = [\beta, \psi, k_1, k_2, k_3, \phi_1, \phi_2, \phi_3]$.[33] Islam (1995) suggests to follow Chamberlain and to impose restrictions by using a minimum-distance (MD) estimator:

$$\widehat{\vartheta} = \arg\min \left(vec\Pi - g(\vartheta)\right)^{'} H_N^{-1} \left(vec\Pi - g(\vartheta)\right) \tag{5.21}$$

where $g(\vartheta)$ is the vector value function mapping the elements of into $vec\Pi$ and ${}^{'}H_N^{-1}$ is the weighting matrix. Through (5.21) we may then recover $\widehat{\mu}_i$and $\widehat{\beta}$ and calculate cross-countries productivity from (5.13).

[32]For simplicity, we are suppressing the individual subscript i.

[33]We are ignoring the intercept term, $k_0 = \phi_0 = 0$.

The main criticism against the use of the Chamberlain's estimator is that for consistency it needs to assumes that $x_t's$ are strictly exogenous. As stressed by Caselli et al. (1996) this may not be the case for the convergence equation model. Hence, they suggest to use the Arellano and Bond (1991) estimator to calculate μ_i from (5.5), re-writing such equation as:

$$\tilde{y}_{it} = \beta \tilde{y}_{it-1} + \sum_{j=1}^{M} \gamma^j \tilde{x}_{it}^j + \mu_i + \upsilon_{it} \tag{5.22}$$

where $\tilde{y}_{it} = y_{it} - \bar{y}_t$, and $\bar{y}_t = \frac{\sum y_{it}}{N}$, while the M included variables are additional determinants of the growth rate. The use of demeaned value of per capita output allows time specific constants (and business cycle effects) to be eliminated. Model (5.22) assumes that:

$$E(\mu_i) = 0, E(\upsilon_{it}) = 0, E(\upsilon_{it}\mu_i) = 0, \quad i = 1, ..., N; \quad t = 2, ..., T \tag{5.23}$$

$$E(\upsilon_{it}\upsilon_{is}) = 0, \quad i = 1, ..., N; \quad t \neq s \tag{5.24}$$

As seen above, the idea of Arellano and Bond (1991) is to check for the presence of fixed effects by taking data in first difference and then to use the instrumental variables technique to purge the correlation between the dependent variable and its lag. Unlike other similar estimators[34] Arellano and Bond (1991) suggest exploiting all the orthogonality conditions existing between y_{it} and the disturbances υ_{it}[35] and thus enhance efficiency. Therefore, in this context we have to apply the first-difference transformation in order to eliminate individual effects and subsequently rearrange equation (5.22) to obtain:

$$\Delta\tilde{y}_{it} = \beta \Delta\tilde{y}_{it-1} + \sum_{j=1}^{M} \gamma^j \Delta\tilde{x}_{it}^j + \Delta\upsilon_{it} \tag{5.25}$$

Equation (5.25) cannot be estimated as it stands since $E(\Delta\widetilde{y}_{it}\Delta\upsilon_{it}) \neq 0$; that is, the lagged dependent variable is correlated with the error term through the contemporaneous terms in period $t-1$ (or $t-\tau$, with $\tau = 5$ as in the most studies in this literature). If we consider the simplest AR(1) model with fixed effects, that is, excluding the M additional determinants from (5.22), under (5.23) and (5.24), the Arellano and Bond (1991) estimator identifies $(T-1)(T-2)/2$ linear moment conditions such that:

$$E(\widetilde{y}_{it-s}\Delta\upsilon_{it}) = 0 \tag{5.26}$$

with $t = 3, \ldots, T$ and $s \geq 2$, or $E(Z_i^{'}\Delta\upsilon_i) = 0$, where:

$$Z_i = \begin{bmatrix} \widetilde{y}_{i1} & 0 & 0 & \ldots & 0 & \ldots & 0 \\ 0 & \widetilde{y}_{i1} & \widetilde{y}_{i2} & \ldots & 0 & \ldots & 0 \\ . & . & . & \ldots & . & \ldots & . \\ 0 & 0 & 0 & \ldots & \widetilde{y}_{i1} & \ldots & \widetilde{y}_{iT-2} \end{bmatrix}; \quad \Delta\upsilon_i = \begin{bmatrix} \Delta\upsilon_{i3} \\ \Delta\upsilon_{i4} \\ \vdots \\ \Delta\upsilon_{iT} \end{bmatrix} \tag{5.27}$$

In other words, it is possible to use the lagged levels of y dated $t-2$ and earlier as instruments. The GMM estimator for β will be given by:

$$\widehat{\beta}_{AB} = \frac{\Delta\widetilde{y}_{-1}^{'} Z W_N^{-1} Z^{'} \Delta\widetilde{y}}{\Delta\widetilde{y}_{-1}^{'} Z W_N^{-1} Z^{'} \Delta\widetilde{y}_{-1}} \tag{5.28}$$

where $\widehat{\beta}_{AB}$ is the GMM estimator of the difference equation with W_N a weight matrix determining the efficiency properties of the GMM estimator. With additional regressors the number of moment conditions depends on the assumptions made on x's. The consistency of this estimation procedure crucially depends

[34] As the Anderson-Hsiao who favour the use of Δy_{it-2} or y_{it-2} as instruments.
[35] See Baltagi (2003).

on the identifying assumption that lagged values of both income and other explanatory variables are valid instruments in the growth regression.[36] However, recent studies criticize the use of the GMM-Arellano and Bond (henceforth GMM-AB) estimator in frameworks such as the standard growth-convergence analysis. Blundell and Bond (1998) find that the GMM-AB estimator may perform poorly with datasets that use either a small number of time periods or persistent time series, where these are typical features of aggregate datasets. In particular, they show that when T is small, and a) the autoregressive parameter is close to one or b) the variance of the individual effect is high relative to the variance of the transient shock, the lagged levels of the series will tend to be only weakly correlated with subsequent first differences and the GMM-AB estimator may produce downward biased estimates.

When there is evidence that lagged levels of the explanatory variables provide weak instruments for the model in first difference as in equation (5.25), the inclusion of additional explanatory variables among regressors and the inclusion of additional lags of these regressors among instruments may improve the performance of this estimator. Therefore, Blundell and Bond (1998) suggest specifying a system of equations in both first difference (as described above) and levels where the instruments of the levels equations are the lagged first-differences of the series. In particular, Blundell and Bond (1998) suggest exploiting these additional moment conditions:

$$E(\mu_i \Delta y_{i2}) = 0 \tag{5.29}$$

Assumption (5.29) holds when the process is mean stationary, that is when:

$$y_{i1} = \frac{\mu_i}{1-\beta} + \eta_i \tag{5.30}$$

with $E(\eta_i) = 0$ and $E(\eta_i \mu_i) = 0$. Note that (5.29) implies the exclusion by assumption of the hypothesis technological (or TFP) catching-up across countries. In fact, if the extent of efficiency growth is related to initial efficiency (as in the catching up case), GDP growth rates (the first difference of log output) might be correlated with the individual effect.

As before, in a simple AR(1) model with fixed effects (that is, excluding additional regressors), given these assumptions it is possible to identify the following $(T-1)(T-2)/2$ moment conditions:

$$E(\upsilon_{it} \Delta y_i^{t-1}) = 0, \qquad t = 3, ..., T \tag{5.31}$$

or, using the matrix notation as in (5.32) $E(Z_{li}^{'} \upsilon_i) = 0$ where:

$$Z_{li} = \begin{bmatrix} \Delta y_{i2} & 0 & 0 & \dots & 0 & \dots & 0 \\ 0 & \Delta y_{i2} & \Delta y_{i3} & \dots & 0 & \dots & 0 \\ . & . & . & \dots & . & \dots & . \\ 0 & 0 & 0 & \dots & \Delta y_{i2} & \dots & \Delta y_{iT-1} \end{bmatrix}; \quad \upsilon_i = \begin{bmatrix} \upsilon_{i3} \\ \upsilon_{i4} \\ \vdots \\ \upsilon_{iT} \end{bmatrix} \tag{5.32}$$

The system GMM (or GMM-SYS) methodology exploits the full set of linear moment conditions given by (5.26) and (5.31): the GMM-SYS estimator for β will be thus given by:

$$\widehat{\beta}_{SYS} = \frac{q_{-1}^{'} Z_{SYS} W_N^{-1} Z_{SYS}^{'} q}{q_{-1}^{'} Z_{SYS} W_N^{-1} Z_{SYS}^{'} q_{-1}} \tag{5.33}$$

with $q_i \left(\Delta y_i^{'}, y_i^{'} \right)$ and estimates used to calculate TFP levels.

Finally, to estimate TFP's from (5.5) recent studies recommend the use of a methodology firstly suggested by Kiviet (1995). To avoid the weak instruments problems described above, Kiviet (1995) advocates a more direct approach to the problem of the finite sample bias in dynamic panels by estimating a small sample correction to the LSDV-WG estimator.[37] In particular, he shows that it is possible to take advantage of the efficiency of the LSDV-WG method (with respect to other IV estimators) correcting the downward bias that

[36] Caselli at al. (1996) distinguish between stock variables and flow variables. The former, measured at the beginning of the period, are assumed as predetermined variables, while the latter (usually measured as averages within the time span considered) are not predetermined for υ_{it} but are assumed to be predetermined for $\upsilon_{it+\tau}$.

[37] This methodology does not produce analytical standard errors. The latter may be calculated with bootstrapping.

characterizes $\widehat{\beta}_{LSDV-WG}$ in the small sample estimates of (5.5). In particular, the LSDV-WG estimator for the vector δ defined in (5.15) can be rewritten as:

$$\delta = \left(W^{'}BW\right)^{-1} W'By \tag{5.34}$$

where $B_t = I_t - \frac{1}{T}\iota_t\iota_t^{'}$ and $B = I_N \otimes B_t$. To calculate the Kiviet correction we need to substitute eq. (5.15) in (5.34) and obtain:

$$\begin{aligned}
E(\widehat{\delta} - \delta) &= E\left(W^{'}BW\right)^{-1} W'B\left[W\delta + (I_n\iota_t)\mu + \varepsilon\right] - \delta = E\left(W^{'}BW\right)^{-1} W'B\varepsilon \qquad (5.35)\\
&= E\left(W^{'}BW\right)^{-1}(W'BW\delta) + E\left(W^{'}BW\right)^{-1} W'BW\delta + (I_n\iota_t)\mu + E\left(W^{'}BW\right)^{-1} W'B\varepsilon - \delta\\
&= E\left(W^{'}BW\right)^{-1} W'B\varepsilon\\
&\because \quad E(\left(W^{'}BW\right)^{-1}(W'BW\delta) = \delta\\
&\because \quad B(I_n \otimes \iota_t)\mu = 0
\end{aligned}$$

The main problem arising from the bias defined by eq. (5.35) is that W is stochastic since it contains the lagged dependent variable term which is covariant with the contemporaneous LSDV error term. Thus, to evaluate the expected value defined by eq.(5.35), Kiviet proposes to partition W into its stochastic and non stochastic component:$W = \bar{W} + \widetilde{W}$, with $\bar{W} = \left[\bar{y}^{(-1)} : X\right]$ and $\widetilde{W} = \left[\widetilde{y}^{(-1)} : 0\right]$. Given this partition, $BW = B\bar{W} + B\widetilde{W}$. Kiviet (1995) shows that the stochastic component $B\widetilde{W}$ can be defined as:

$$B\widetilde{W} = (I_N \otimes B_T C)\widehat{\varepsilon}q^{'} \tag{5.36}$$

where $q = (1, 0, ..., 0)^{'}$, a $(K + 1x1)$ vector, and:

$$C = \begin{bmatrix}
0 & . & . & . & . & . & 0\\
1 & 0 & . & . & . & . & .\\
\beta & 1 & 0 & . & . & . & .\\
\beta^2 & \beta & 1 & 0 & . & . & .\\
. & . & . & . & . & . & .\\
. & . & . & . & . & . & .\\
\beta^{T-2} & . & . & . & \beta & 1 & 0
\end{bmatrix} \tag{5.37}$$

Finally, to calculate the bias Kiviet uses (5.36) and obtain:

$$E(\widehat{\delta} - \delta) = -\widehat{\sigma_\varepsilon}^2(\bar{D})^{-1}(g_1 + g_2 + g_3) + O(N^{-1}T^{-\frac{3}{2}}) \tag{5.38}$$

where:

$$\begin{aligned}
\bar{D} &= \bar{W}^{'}B\bar{W} + \widehat{\sigma_\varepsilon}^2 Ntr\left\{C^{'}B_TC\right\}qq^{'}\\
B\bar{W} &= E(BW)\\
g_1 &= \frac{N}{T}(\iota_t^{'}C\iota_t)\left[2q - \bar{W}^{'}B\bar{W}(\bar{D})^{-1}q\right] \qquad (5.39)\\
g_2 &= tr\left\{\bar{W}^{'}(I_N \otimes B_TCB_T)\bar{W}(\bar{D})^{-1}\right\}q\\
g_3 &= \bar{W}^{'}(I_N \otimes B_TCB_T)\bar{W}(\bar{D})^{-1}q + \widehat{\sigma_\varepsilon}^2 Nq^{'}(\bar{D})^{-1}q \times \qquad (5.40)\\
&\quad \times\left[-\frac{N}{T}(\iota_t^{'}C\iota_t)tr\left\{C^{'}B_TC\right\} + 2tr\left\{C^{'}B_TCB_TC\right\}\right]q
\end{aligned}$$

In sum, the Kiviet Correction may be described by the following three steps procedure:

1. Since the bias approximation depends on the unknown parameter β and σ_v^2, firstly estimate Eq.(5.22) using a consistent estimator, such as Anderson-Hsiao or Arellano-Bond and calculate the bias.
2. Estimate the model by LSDV.
3. LSDV estimates may be corrected by subtracting the bias terms described by (5.38).

To sum up, all the estimators used to infer TFP levels from (5.5) have their pros and cons and the answer to the question of "what is the best fixed effects estimator to estimate equation (5.5) and, thus, TFP levels" in the aggregate convergence equation approach is not simple. What Kiviet wrote a few years ago it is probably still true today: "As yet, no technique is available that has shown uniform superiority in finite samples over a wide range of relevant situations as far as the true parameter values and the further properties of the DGP are concerned" (Kiviet, 1995, p. 72). Overall, the existing Monte Carlo analysis that compare the finite sample performance of these reviewed estimators conclude that for TSCS panels but with a relatively small T (as we find in this specific literature) the Kiviet estimator seems more attractive than other available estimators.[38]

5.2 Estimating TFP from micro data

The most common approach to (individual) TFP estimation includes a stochastic disturbance and expresses (1.1) in logs (lowercase letters) as:

$$y_{it} = a_{it} + \mathbf{x}_{it}\beta + e_{it} \tag{5.41}$$

In equation (5.41) a_{it} is individual productivity, $\mathbf{x}_{it}$ is a $(1 \times L)$ vector of inputs, and β is the $(L \times 1)$ vector of the elasticities of output with respect to each input. The error term e_{it} is meant to capture measurement errors and unobserved idiosyncratic shocks, due for instance to environmental or market changes, which are "unanticipated"by the firm and thus uncorrelated with $\mathbf{x}_{it}$.

The value of a_{it} can be recovered by estimating the vector $\hat{\beta}$, computing the fitted value of firm i's output $\hat{y}_{it}$ and deriving $\hat{a}_{it}$ as the (exponential of the) difference between y_{it} and $\hat{y}_{it}$ (i.e. "Solow residual").

However, standard OLS estimation of (5.41) could run in two orders of problems.[39]

The first problem stems from the fact that information on a_{it}, although unknown to the econometrician, is commonly used by the firm in its decision concerning the amount of inputs. This makes the error term e_{it} correlated with $\mathbf{x}_{it}$ and the OLS-estimated β biased. In econometric parlance, a_{it} is said to "transmit"to the explanatory variables, hence the term "transmission bias". Note that such bias cannot be removed by assuming that the productivity component is not observed by the firm, since in any case one has to reckon with the fact that the amount of inputs is jointly determined with y_{it}, which is just an alternative way of saying that the error term is correlated with the explanatory variables. Whether we want to look at this correlation from the former or the latter point of view, y_{it} and $\mathbf{x}_{it}$ must be regarded as the solution of a simultaneous-equations system. Thus, the problem is one of *simultaneity*. Although these two ways of looking at simultaneity are equivalent with respect to the econometric stratagems to which one can resort, it is worth noting how the former poses, more properly, a problem of "omitted variables". This aspect is stressed by the approach described in section 5.2.1.

The second problem originates from the fact that firms' output, which is needed in order to estimate the production function parameters, is commonly unavailable in physical terms. This forces the econometrician to use a proxy that, in the vast majority of cases, consists of sales deflated by an industry-wide price index, given that individual prices are themselves commonly not available. Such circumstance has no relevance under perfect competition as all firms quote the same price. On the contrary, when markets are imperfectly competitive, firm-level estimated productivity is likely to be misstated. Since the problem is caused by omitting the individual price from the estimation, this problem is usually referred to as *omitted price* bias.[40]

[38] See Judson and Owen (1999), Bun and Carree (2005) and Everaert and Pozzi (2007).

[39] Direct estimation of equation (5.41) also suffers, to a greater or lesser extent, from *multicollinearity*, since input demand functions tend to depend on one another, and *heteroskedasticity*, as the variance of the stochastic disturbance might differ across firms. In this survey we are not concerned with these issues.

[40] The following exposition focuses on *output* price dispersion but (see Katayama et al., 2003) a similar problem of *input* price dispersion affects the determination of $\mathbf{x}_{it}$.

Simultaneity and price dispersion form the basis of the following examination. In particular, we are first concerned with simultaneity (section 5.2.1), which affects the estimation both under perfect and imperfect competition, then with price dispersion (section 5.2.2). An issue, the latter, that has to be addressed only if imperfect competition is the framework of reference.

5.2.1 Dealing with simultaneity: proxy variables methodologies

Following Klette and Griliches (1996) in rewriting the estimating version of equation (5.41), the problem of simultaneity can be summarized as follows:

$$\tilde{y}_{it} = x_{it}\beta + u_{it} \tag{5.42}$$

where $u_{it} = a_{it} + e_{it}$. The OLS estimator of β in (5.42) is, in matrix notation:

$$\hat{\beta} = (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\tilde{y} \tag{5.43}$$

where $\mathbf{x}$ is the $(N \times L)$ matrix of factor inputs (with N denoting the number of observations). Assuming orthogonality in the error term e_{it}, the probability limit of $\hat{\beta}$ can be written as

$$plim_{N\to\infty}(\hat{\beta}) = \beta + plim_{N\to\infty}\left[(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'a\right] \tag{5.44}$$

where a is the $(N \times 1)$ vector of individual productivities which are observed by the firm but not by the econometrician.

The second term on the right hand side of equation (5.44) embodies the transmission bias. This can be seen as the OLS estimator of vector ε in the auxiliary regression $a = \mathbf{x}\varepsilon + u^a$, where u^a is an iid error term. Accordingly, we can write:

$$plim_{N\to\infty}(\hat{\beta}) = \beta + \varepsilon \tag{5.45}$$

so that in the limit firm estimated productivity evaluates to

$$plim_{N\to\infty}(\hat{a}_{it}) = y_{it} - \mathbf{x}_{it}plim_{N\to\infty}(\hat{\beta}) = a_{it} - \mathbf{x}_{it}\varepsilon$$

where $\mathbf{x}_{it}\varepsilon$ is the associated transmission bias.

The theoretical stratagems to which one can resort, in order to keep into account the presence of simultaneity, go along with the "anatomy"of the TFP component. Specifically, the unobserved (by the econometrician) TFP term in eq. (5.41) is both firm-specific (the i index), and time-varying (the t index). Traditional cross-section analysis (Douglas, 1948) substitutes a constant for the unobserved TFP ($a_{it} \to a$), so that all its variability is included in the error term and all the simultaneity bias passed on the estimates. With plant/firm panels, a first step towards mitigating the simultaneity bias can be made by reducing a_{it} to a firm-specific (but time-invariant) unobserved effect ($a_{it} \to a_i$). In this case, the TFP component is understood as an unobservable effect in a *fixed-effects estimation*.[41] However, while this approach takes account of firm heterogeneity, it does not keep the temporal dimension into account. A way to keep also the latter into consideration consists of identifying a (proxy) variable that reacts to the changes in the TFP observed by the firm and is therefore a function of it. Insofar as this function proves to be invertible, its inverse can be calculated and plugged in the estimating equation before proceeding to estimate the production function parameters. Summing up, the idea behind this *proxy-variable method* consists of recovering the productivity component by the traces it leaves in the observed behaviour of the firm. This approach, firstly proposed by Olley and Pakes (1996) using investment as a proxy, has recently been extended by Levinsohn and Petrin (2003) to the use of the intermediate inputs.

[41] Three considerations are in order, which make the FE approach not fully satisfactory. First, the within estimator uses only the variation across time, thus leaving conspicuous part of the cross-sectional information unexploited. Second, the assumption that the unobserved TFP is constant over time seems to be a too strong restriction. Third, the FE estimator is consistent (Chamberlain, 1982) only provided that $E(\ddot{e}_{it} \mid \ddot{x}_{i1}, \ldots, \ddot{x}_{it}) = 0 \quad \forall t = 1, \ldots, T$. By entailing that the error term in each time period is uncorrelated with the explanatory variables in each time period (i.e. strict exogeneity), this condition implies that the unobserved TFP not only does not vary over time, but also does not affect the present and future input choices. Such situation of "no delayed transmission" can be hardly the case in the presence of simultaneity.

To illustrate how the current literature relies on proxy variable methods in order to take into account the presence of simultaneity in the estimation process, it is convenient i) to make explicit the vector of inputs by assuming two production factors, capital (k) and labour (l), whose production coefficients are referred to as respectively β_k and β_l; ii) to hypothesize that individual productivity evolves according to a first-order Markov process. (i) and (ii) imply, respectively, that:

$$y_{it} = \beta_k k_{it} + \beta_l l_{it} + a_{it} + e_{it} \tag{5.46}$$

and

$$a_{it} = E[a_{it} \mid \Omega_{it}] = E[a_{it} \mid a_{it-1}] + u_{it} \tag{5.47}$$

where e_{it} is the "untransmitted shock", Ω_{it} is the information set at time t, and u_{it} denotes "innovation"in a_{it}.

The Olley - Pakes (OP) method.
Olley and Pakes (1996) developed a two-stages estimation procedure. The identification of a proxy variable for a_{it} relies on several assumptions.

1. Proxy variable: investment reacts to the observed (by the firm) TFP according to $i_{it} = i(k_{it}, a_{it})$.
2. Strict monotonicity: $i_{it} = i(\cdot)$ is strictly monotonic in a_{it} .
3. Scalar unobservable: a_{it} is the only unobservable in $i_{it} = i(\cdot)$.[42]
4. Dynamic implications of input choices: capital is a state variable, and is the only state variable. The law of motion follows $k_{it} = k(k_{it-1}, i_{it-1})$. Labour is a 'static input' — i.e. labour demand at a given point in time has no dynamic implications on future profits.
5. Timing of input choices: investment and capital (thorough investment) are both decided at time $t-1$. Labour is chosen in t, when firm productivity is observed. This entails that:

$$\begin{array}{ll} l_{it} \in \Omega_t & l_{it} \ni \Omega_{t-1} \\ k_{it}, i_{it} \in \Omega_t & k_{it}, i_{it} \in \Omega_{t-1} \end{array}$$

Following this set of assumptions, investment and capital are orthogonal in t ($E[i_{it}|k_{it}] = 0$), and i_{it} can be inverted, yielding the following proxy for the unobserved TFP:

$$a_{it} = h(i_{it}, k_{it}) \tag{5.48}$$

The unobservable productivity is thus expressed as a function of observables. Substituting (5.48) in (5.46) yields:

$$y_{it} = \beta_l \, l_{it} + \Phi_{it}(i_{it}, k_{it}) + e_{it} \tag{5.49}$$

where

$$\Phi_{it}(i_{it}, k_{it}) = \beta_0 + \beta_k \, k_{it} + h(i_{it}, k_{it}) \tag{5.50}$$

Equation (5.49) is a "partially linear"model identifying β_l. As the regressors are no longer correlated with the error, β_l can be estimated by approximating Φ by a third or fourth order polynomial $\tilde{\Phi}$ in i and k (i.e. FIRST STAGE).

However, β_k is not identified at this stage. In order to yield a consistent estimation of the latter, we have to introduce further structure into the model and use, in a second stage, the estimated coefficient of labour ($\hat{\beta}_l$). To this aim, net from the output in equation (5.46) the estimated contribution of labour and use equation (5.47)[43]

$$y_{it} - \hat{\beta}_l \, l_{it} = \beta_k k_{it} + E[a_{it}|a_{it-1}] + \nu_{it}. \tag{5.51}$$

[42]Under the Markovian specification above, firms' investment functions can be proved to be strictly increasing in a_{it} (see Pakes, 1994).

[43]The OP procedure allows for more general assumptions on the evolution of a_{it}. To see this, write equation (5.52) as

$$y_{it} - \hat{\beta}_l \, l_{it} = \beta_k k_{it} + g(\tilde{\Phi}_{it-1} - \beta_k k_{it-1}) + e_{it}.$$

where the function $g(\cdot)$ simply reduces to $\tilde{\Phi}_{it-1} - \beta_k k_{it-1}$ under the random walk assumption (5.47).

where $\nu_{it} = u_{it} + e_{it}$ is a "composition of pure errors". Given $E[a_{it}|a_{it-1}] = h_{t-i}(i_{it-1}, k_{it-1}) = \Phi_{it-1} - \beta_0 - \beta_k k_{it-i}$, it follows that β_k is identified by the following "net output"equation:

$$y_{it} - \hat{\beta}_l \, l_{it} = \hat{\Phi}_{it-1} + \beta_k (k_{it} - k_{it-1}) + u_{it} \tag{5.52}$$

where, from (5.49)

$$\hat{\Phi}_{it-1} = y_{it-1} - \hat{\beta}_0 - \hat{\beta}_l l_{it-1} - \hat{e}_{it-1}. \tag{5.53}$$

Equation (5.52) can be estimated through non-linear least squares, needed in order to restrict β_k to be the same for k_{it} and k_{it-1} (i.e. SECOND STAGE).[44]

Operationally, one can proceed by constraining the residual of the regression of $(y_{it} - \hat{\beta}_l l_{it} - \beta_k k_{it})$ on $(\hat{\Phi}_{it-1} - \beta_k k_{it-1})$ to be not above an arbitrarily low level, which can be seen as a moment in the residual ν_{it}.[45]

Note how the whole set of assumptions above has been used in the procedure. In particular, all the hypothesis except the last one (and in particular the orthogonality between contemporaneous levels of i and k, implied by assumptions 1 and 4) are required for perfectly invert out a_{it}. The assumption that k_{it} is decided before time t (the time in which productivity is observed), by requiring that $E[u_{it}|k_{it}] = 0$, is key for identifying correctly the capital coefficient in the second stage. However, as l is chosen at t (assumption 5), exactly when firms' productivity is observed, the labour coefficient cannot be identified in the first stage without assuming that $E[u_{it}|l_{it}] = 0$. Under this condition, the information on firms' investment decision in t can be in fact used in the identification of β_l to control for the productivity shock correlated with l_{it}.

The OP method has several advantages over a within estimator. First of all, as pointed out by Levinsohn and Petrin, it is "less costly". OP leaves in fact more variance in the estimation, since it uses also the cross-section information. Second, by looking a the investment decision as the solution of a dynamic optimisation problem, OP introduces an explicit behavioural hypothesis in the estimation procedure. However, the key hypothesis of orthogonality between k and u, with its requirement that observed productivity fully transmit to the investment decision, might be considered too demanding.

Another crucial strength of OP is that, unlike the other methods that we will describe in this section, this procedure provides a relatively easy solution to the potential *selection bias* associated with non-randomness in plants dropping out (s.c. *selectivity*).[46] The remedy consists of incorporating a fitted value for the probability of exiting from the sample in the estimation of equation (5.52), which becomes

$$y_{it} - \hat{\beta}_l \, l_{it} = \beta_k k_{it} + g(\hat{\Phi}_{it-1} - k_{it-1}, \widehat{Pr}_{t-1}) + u_{it} \tag{5.54}$$

where $\widehat{Pr}_{t-1}$ is estimated as the probit of a survival indicator variable on a polynomial in capital and investment, and $g()$ is a high-order series expansion in the three arguments $\hat{\Phi}_{it-1}, k_{it-1}, \widehat{Pr}_{t-1}$, including all cross terms.

---

[44] Wooldridge (2005) suggests an alternative implementation in which the first and second stage are estimated simultaneously. The procedure can be applied to all the proxy-variable methods described in this survey.

[45] In a GMM context, this can be expressed as:

$$Q(\beta_k^*) = \min_{\beta_k^*} \sum_i \sum_{t=T_{i0}}^{T_{i1}} \hat{\nu}_{it} k_{it}$$

where: T_{i0} and T_{i1} are, respectively, the first and last period in which firm i is observed; and

$$\hat{\nu}_{it} = y_{it} - \hat{\beta}_l l_{it} - \beta_k^* k_{it} - E[\widehat{a_{it} \mid a_{it-1}}],$$

with

$$E[\widehat{a_{it} \mid a_{it-1}}] = \hat{\Phi}_{it-1} - \beta_k k_{it-1}$$

Equation (45) is the sample analogue to the moment $E[u_{it} k_{it}] = 0$, on which the whole second stage is based.

[46] A short-cut solution would consist of considering a balanced (sub-)sample. However, this solution is likely to result in biased estimates according to the systematic differences between exiting and non-exiting firms in terms of production factors. Consider (Arnold, 2005) for example the case in which plants with higher capital stock are less likely to drop out of the market (and the sample) if affected by a negative shock. In the remaining sample, there will be a non-zero (say negative) correlation between the realizations of the error term and the capital stocks. In this case, the estimated capital coefficient will suffer from a downward bias.

Two extensions of the OP are worth mentioning.
De Loecker (2007a) extends the OP framework by allowing market structure (factor markets, demand conditions, exit barriers, etc.) to be different for exporting firms by introducing export into the underlying structural model, so that firm's decisions about how much to invest and whether to exit the market or not depends on the export status (exporting versus non-exporting). This modified OP procedure is meant to capture unobserved productivity shocks correlated with export status and to filter out differences in market structures between domestic and exporting firms within a given industry.

Operationally, an export dummy is introduced in equation (5.47), which becomes

$$a_{it} = h_e(i_{it}, k_{it}) \tag{5.55}$$

where e denotes the presence of the export dummy. Under this condition, the first stage estimation now includes the export dummy and all terms interacted with it. Apart from the interacted terms, this is equivalent to introducing the export status as an input in the production function estimation. In particular, the polynomial (5.50) has to be re-written as

$$\Phi_{e,it}(i_{it}, k_{it}) = \beta_0 + \beta_k\, k_{it} + h_e(i_{it}, k_{it}) \tag{5.56}$$

entailing the following second stage estimating equation:[47]

$$y_{it} - \hat{\beta}_l\, l_{it} = \hat{\Phi}_{e,it-1} + \beta_k(k_{it} - k_{it-1}) + u_{it} \tag{5.57}$$

where, $\hat{\Phi}_{e,it-1}$ is defined as in (5.53), with the only difference of the export status dummy. We refer to Appendix B of De Loecker (2007a) for a discussion of the direction of the bias associated with not controlling for the export status. However, note that, compared to the standard OP approach, the labour coefficient is expected to be lower, while the direction of the bias in the capital coefficient cannot be identified univocally.

In the last stage of the estimation procedure suggested by De Loecker (2007a), one is implicitly assuming that the export status only affects the average of the future productivity distribution, entailing that the learning by exporting effects are not firm-specific. Moreover, these effects are time-invariant (i.e. every year, exporting raises output, conditioned on labor and capital, by the coefficient estimated on the export dummy).

Van Biesebroeck (2005) removes this limit by adopting a similar setup in which, however, lagged export status is introduced as a state variable. In particular, in addition to the standard OP procedure, firms have a further state variable $E\mathbf{x}_{t-1}$ and a further control variable, $\Delta E\mathbf{x}_{t-1} = E\mathbf{x}_t - E\mathbf{x}_{t-1}$. The estimation procedure is as in OP - De Loecker (2007a), with the only difference that the policy function for investment becomes $i_{it} = i(k_{it}, a_{it}, ex_{it-1})$.[48] This extension is of particular interest, because the estimated coefficient of the export variable is informative about the presence or absence of learning by doing effects. Van Biesebroeck (2005), for example, finds significative evidence in favor of the presence of such effects.

The Levinsohn - Petrin (LP) method.
Levinsohn and Petrin (2003) rely on intermediate inputs as a proxy variable for a_{it}, rather than on investment. The identification of this proxy relies on the following assumptions.

1. Proxy variable: intermediates react to the observed (by the firm) TFP according to the demand function $m_{it} = m(a_{it}, k_{it})$.
2. Strict monotonicity: $m_{it} = m(\cdot)$ is strictly monotonic in a_{it}.

[47] If one is controlling for the selection bias, also the survival equation will have to include the export dummy and all terms interacted with the export dummy, and the second stage estimating equation will be given by:

$$y_{it} - \hat{\beta}_l\, l_{it} = \beta_k k_{it} + g(\hat{\Phi}_{e,it-1} - k_{it-1}, \widehat{Pr}_{e,t-1}) + u_{it}$$

where $\widehat{Pr}_{e,t-1}$ is estimated as the probit of a survival indicator variable on a polynomial in capital and investment, and $g()$ is a high-order series expansion in the three arguments $\hat{\Phi}_{e,it-1}, k_{it-1}, \widehat{Pr}_{e,t-1}$, including all cross terms.

[48] If one is controlling for the selection bias, also the the survival equation has to be changed, as it is now a function of both current and past export status.

3. Scalar unobservable: a_{it} is the only unobservable in $m_{it} = m(\cdot)$.

4. Dynamic implications of input choices: labour is a "static input"— i.e. input demand at a given point in time has no dynamic implications on future profits.[49]

5. Timing of input choices: capital is decided at time $t-1$, labour and intermediates are chosen in t, when firm productivity is observed. Formally:

$$\begin{array}{cc} l_{it}, m_{it} \in \Omega_t & l_{it}, m_{it} \ni \Omega_{t-1} \\ k_{it} \in \Omega_t & k_{it} \in \Omega_{t-1} \end{array}$$

Let us start describing the LP method by including the intermediates demand function (m_{it}) in equation (5.41):[50]

$$y_{it} = a_{it} + \beta_k\, k_{it} + \beta_l\, l_{it} + \gamma m_{it} + e_{it}. \tag{5.58}$$

In (5.58) $m_{it} = m(a_{it}, k_{it})$ replaces OP's investment function in order to generate, once inverted (the invertibility condition is, as before, that, conditional on capital, intermediate inputs demand is increasing in a), the proxy:

$$a_{it} = h(m_{it}, k_{it}) \tag{5.59}$$

which, plugged into (5.58), yields:

$$y_{it} = \beta_l\, l_{it} + \Phi_{it}(m_{it}, k_{it}) + e_{it} \tag{5.60}$$

where

$$\Phi_{it}(m_{it}, k_{it}) = \beta_0 + \beta_k\, k_{it} + \gamma m_{it} + h(m_{it}, k_{it}) \tag{5.61}$$

As before, only β_l is identified at this stage and, as before, β_k can be estimated in a second stage. In addition, also γ has to be identified in the second stage.

Proceeding as in OP, we end up with:

$$y_{it} - \hat{\beta}_l\, l_{it} = \hat{\Phi}_{it-1} + \beta_k(k_{it} - k_{it-1}) + \nu_{it} \tag{5.62}$$

with $\hat{\Phi}_{it-1} = y_{it-1} - \hat{\beta}_0 - \hat{\beta}_l l_{it-1}$.

However, differently from OP, $\nu_{it} = u_{it} + e_{it}$ is no longer a "composition of pure errors". Intermediates are in fact correlated with the error term, as they react to the innovation u_{it}. Thus, OLS provide inconsistent estimation. Owing to this, $(\hat{\beta}_k, \hat{\gamma})$ this time are obtained by minimising the following GMM criterion function

$$Q(\beta_k^*, \gamma^*) = \min_{(\beta_k^*, \gamma^*)} \sum_h \left(\sum_i \sum_{t=T_{i0}}^{T_{i1}} \hat{\nu}_{it} Z_{iht} \right)^2, \tag{5.63}$$

where: h indexes the elements of $Z_t = (k_t, m_{t-1})$; i indexes firms; T_{i0}, T_{i1} are, respectively, the first and last period in which firm i is observed; and

$$\hat{\nu}_{it} = y_{it} - \hat{\beta}_l l_{it} - \beta_k^* k_{it} - \gamma^* m_{it} - E[\widehat{a_{it} \mid a_{it-1}}]. \tag{5.64}$$

According to (5.64), in order to proceed with the minimisation of (5.63), we need to know β_k^*, γ^*, and $E[\widehat{a_{it} \mid a_{it-1}}]$ ($\hat{\beta}_l$ is known from the first stage).[51]

[49] This is not a necessary condition. Labour could be allowed to have dynamic implications, but in this case l_{t-1} should be included in the intermediate input demand function.

[50] In the following description we largely borrow from Levinsohn et al. (2003), a pdf posted on the authors' web page, supplementing the STATA package "levpet".

[51] The vector

$$E[(e_{it} + u_{it}) \mid Z_t],$$

at the base of the moment conditions, results from two assumptions (which in turn represent the conditions under which intermediate input can be thought of as a "perfect proxy"for a_{it}). The first one is that period t's capital is determined by the investment decisions in the previous period, so that it does not respond to the productivity innovation (e_{it}) in the current period:

$$E\left[(\nu_{it}) \mid k_{it}\right] = 0$$

The second assumption is that last period's intermediate input choice is uncorrelated with the innovation in the current period:

$$E\left[(\nu_{it}) \mid m_{it-1}\right] = 0.$$

We can start from calculating the following residuals:

$$y_{it} - \beta_l l_{it} = \hat{\Phi}_{it} \tag{5.65}$$

then, a_{it} can be obtained using any candidate values β_l^* and γ^* in the following equation:

$$\hat{a}_{it} = \hat{\Phi}_{it} - \beta_k^* k_{it} - \gamma^* m_{it}. \tag{5.66}$$

Using these values, we are able to obtain a consistent approximation to $E[\widehat{a_{it} \mid a_{it-1}}]$ from

$$E[\widehat{a_{it} \mid a_{it-1}}] = \delta_0 + \delta_1 a_{it-1} + \delta_2 a_{it-1}^2 + \delta_3 a_{it-1}^3 + \epsilon_{it} \tag{5.67}$$

Finally, given $\hat{\beta}_{l\,it}$, β_k^*, γ^*, and $E[\widehat{a_{it} \mid a_{it-1}}]$, the solution of problem (5.63) provides the estimation of capital ($\hat{\beta}_k$) and intermediate input ($\hat{\gamma}$) coefficients.

Compared to OP, the LP procedure has two important advantages. The first one is theoretical. As recognised by Levinsohn and Petrin themselves, LP provides "a better link between the estimation strategy and the economic theory, primary because intermediate inputs are not typically state variables". The second advantage stems from a practical problem: balance sheet data are often characterised by a high degree of zero-investment reports. Thus, a conspicuous number of observations fall out of the estimation. This might have strong implications on the estimates. First, through the (im)possibility to invert the investment function, in order to obtain equation (5.47). Second, through the fact that, if the presence of those zeros is due to adjustment costs, the exclusion of the relevant observations leads to significant truncation bias. By contrast, zero-reports for intermediate inputs are rare.

The Ackerberg - Caves - Frazer (ACF) correction.
Ackerberg et al (2006) criticize both OP and LP, claiming that the produced estimates would suffer from collinearity, arising in the first stage of the estimation procedure. The reason is easy to show. Consider the first stage estimating equation

$$y_{it} = \beta_l \, l_{it} + \Phi_{it}(\cdot) + e_{it} \tag{5.68}$$

in which $\Phi_{it}(\cdot)$ is given by equations (5.50) for OP and (5.61) for LP. For β_l to be correctly identified, the labour demand l_{it} has to vary independently of Φ. Now, the underlying model suggests both, since the most obvious hypothesis one can have in mind on the data generating process for labour demand is for it to be, in both cases, a function of capital and productivity $l_{it} = f(a_{it}, k_{it})$. However, once substituted for a_{it}, by using respectively (5.48) and (5.59), one remains with $l_{it} = f(h(i_{it}, k_{it}), k_{it})$ in OP and $l_{it} = f(h(m_{it}, k_{it}), k_{it})$ in LP. Thus, labour demand is only a function of capital plus the variable chosen as proxy (that is: investment, in OP, or intermediates, in LP), entailing that the labour coefficient cannot be identified in the first stage, as one cannot simultaneously estimate a fully non-parametric function of two variables (i, k in OP and m, k in LP) along with a coefficient on a variable (l) that is only a function of those same variables. ACF provide us with an alternative, consisting of a slight modification to the OP/LP timing of input decisions. In particular, the ACF correction has to do with the timing of input choices (hypothesis 5).

Assume that firms' input decision proceed with the following timing. In particular, assume that

1. Proxy variable: intermediates react to the observed (by the firm) TFP according to the demand function $m_{it} = m(a_{it}, k_{it}, l_{it})$.
2. Strict monotonicity: $m_{it} = m(\cdot)$ is strictly monotonic in a_{it}.
3. Scalar unobservable: a_{it} is the only unobservable in $m_{it} = m(\cdot)$.
4. Timing of input choices: labour is decided at time $t - b$, with $(0 < b < 1)$, capital is chosen at $t - 1$, the intermediate input is chosen at t, when firm productivity is observed. Formally:

$$\begin{array}{ll} k_{it} \in \Omega_{t-b} & k_{it} \in \Omega_{t-1} \\ l_{it} \in \Omega_{t-b} & l_{it} \ni \Omega_{t-1} \\ m_{it} \ni \Omega_{t-b} & m_{it} \ni \Omega_{t-1} \end{array}$$

Note how, differently from OP and LP, in which labour is a "perfectly variable" input decided at the time production takes place, the above timing of input decisions implies that labour is a "less variable"input than intermediates, as labour is chosen one subperiod before productivity is observed.

Note also that we are no longer assuming that labour demand has no dynamic implications on future profits (i.e. assumption 4 in OP and LP).

Moreover, suppose that a_{it} still evolves according to a first order markov process between these subperiods, namely:

$$\begin{aligned} a_{it} &= E[a_{it} \mid \Omega_{it-b}] = E[a_{it} \mid a_{it-b}] + u_{it} \\ a_{it-b} &= E[a_{it-b} \mid \Omega_{it-1}] = E[a_{it-b} \mid a_{it-1}] + u_{it-b} \end{aligned}$$

Under these hypothesis one is unable to identify the labour coefficient in the first stage, but it is still possible to use the first stage in order to net output of the untransmitted shock e_{it}. In particular, as before, the intermediate demand function can replace, once inverted, the productivity term in the production function, yielding:

$$y_{it} = \Phi_{it}(m_{it}, k_{it}, l_{it}) + e_{it} \tag{5.69}$$

where

$$\Phi_{it}(m_{it}, k_{it}, l_{it}) = \beta_0 + \beta_k\, k_{it} + \beta_l l_{it} + h(m_{it}, k_{it}, l_{it}) \tag{5.70}$$

Once $\hat{\Phi}$ is obtained, one can proceed as before, with the difference that now both β_l and β_k have to be recovered in the second stage. Thus, two moment conditions have to be used. However, given the new timing assumption, we know that labour in $t-1$ is uncorrelated with unobserved productivity in t, hence $E[u_{it}|l_{it-1}] = 0$. This, together with $E[u_{it}|k_{it}] = 0$, gives rise to the following GMM criterion function:

$$Q(\beta_k^*, \beta_l^*) = \min_{(\beta_k^*, \beta_l^*)} \sum_h \sum_i \sum_{t=T_{i0}}^{T_{i1}} \hat{u}_{it} Z_{iht}, \tag{5.71}$$

where: h indexes the elements of $Z_t = (k_t, l_{t-1})$; i indexes firms; T_{i0}, T_{i1} are, respectively, the first and last period in which firm i is observed.

Operationally, start from calculating the residuals

$$\hat{a}_{it} = \hat{\Phi}_{it} - \beta_k k_{it} - \beta_l l_{it}. \tag{5.72}$$

for any given candidate to (β_k^*, β_l^*) $\forall t$. Then recover the implied residuals u_{it}'s by non-parametrically regressing a_{it} on a_{it-1} (plus a constant term), and proceed by minimising (5.71).

Two considerations are in order.

First, a key feature of the ACF procedure is that it leaves the door open to any generalization of the production function in which all inputs are not perfectly variable. Whenever the demand of one or more inputs does not belong to the information set Φ_t (input demand is assumed to be decided before time t, which is the period in which productivity is observed), a set of moment conditions of the type $E\left(u_{it} \cdot \binom{k_{it}}{Z_{ht-1}}\right) = 0$ can be formed, with Z_{ht-1} representing the vector of inputs decided before time t.

Second, we illustrated the ACF correction using intermediates as a proxy (this follows the preference of the authors). However, the same approach can be applied to OP. In this case, the moment conditions are the same but the procedure becomes inconsistent with the generalization above, as other inputs than labour cannot be admitted to have dynamic effects through entering either the investment or labour decision, as this makes the inversion problematic. Hence, the assumption on the dynamic implications of input choices (assumption 4 in OP and LP) has to be maintained.

5.2.2 Dealing with the Omitted Price Bias: a Price-Dispersion-Corrected (PDC) measure

As mentioned above, when markets are imperfectly competitive, firm sales deflated by an industry-wide price index are no longer a correct proxy for firm's output.

As in section 5.2.1 for the transmission bias, we can summarize the problem by rewriting the estimating version of equation (5.41) as follows:

$$\tilde{r}_{it} = \mathbf{x}_{it}\beta^r + u_{it}^r \tag{5.73}$$

where $u^r_{it} = a^r_{it} + e^r_{it}$ and, due to data availability, physical output has been replaced by deflated sales $\tilde{r}_{it} = r_{it} - p_t = y_{it} + q_{it}$, with r_{it} indicating firm revenues and $q_{it} = p_{it} - p_t$ measuring the difference between (the log of) the firm specific price p_{it} and (the log of) the deflator p_t. The OLS estimator of β^r in matrix notation is:

$$\hat{\beta}^r = (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\widetilde{r}$$

where $\widetilde{r}$ is the $(N \times 1)$ vector of deflated sales and $\mathbf{x}$ is the $(N \times L)$ matrix of factor inputs (with N denoting the number of observations). In the absence of simultaneity, mindful that $\tilde{r}_{it} = y_{it} + q_{it}$, the probability limit of $\hat{\beta}^r$ can be written as:

$$plim_{N\to\infty}(\hat{\beta}^r) = \beta + plim_{N\to\infty}\left[(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'q\right] \tag{5.74}$$

where q is the $(N \times 1)$ vector of the differences between individual prices and the industry deflator, and we assumed orthogonality in the error term e^r_{it}. The second term on the right hand side is the omitted price bias. This can be seen as the OLS estimator of vector ω in the auxiliary regressions $q = \mathbf{x}\omega + u^q$, where u^q is an orthogonal error term. Accordingly, we can write:

$$plim_{N\to\infty}(\hat{\beta}^r) = \beta + \omega \tag{5.75}$$

so that, in the limit, estimated TFP evaluates to

$$plim_{N\to\infty}(\widehat{a}_{it}) = \widetilde{r}_{it} - \mathbf{x}_{it}\, plim_{N\to\infty}(\hat{\beta}^r) = a_{it} - \mathbf{x}_{it}\omega$$

where $\mathbf{x}_{it}\omega$ is the associated omitted price bias.

The inconsistency of the estimator obtained from a production function regression such as (5.73) has been analyzed by Klette and Griliches (1996), who show that using a common, industry-wide, sales deflator results in downward biased estimated returns to scale and, thus, in overstated firm productivity. Klette and Griliches (1996) provide a remedy to this bias but in absence of simultaneity. Melitz (2000) shows that there is a relatively simple way to obtain a consistent estimator for β by adapting the Klette and Griliches approach to one of the Proxy-Variables frameworks described above. The resulting procedure (henceforth PDC method) allows to keep both simultaneity and price dispersion into consideration.

To illustrate Melitz's proposal, let us start by assuming that using Proxy-Variables enables us to purge the estimates from the simultaneity bias, hence $\varepsilon = 0$. This assumption entails that under perfect competition the estimator built on the Proxy-Variables approach is consistent, as $p_{it} = p_t$ implies $\omega = 0$ so that $plim_{N\to\infty}(\hat{\beta}) = \beta$ and $plim_{N\to\infty}\hat{a}_{it} = a_{it}$.

Whenever $p_{it} \neq p_t$, the estimator is however affected by the omitted price bias, since $\omega \neq 0$.

To solve this problem, let us reinterpret the Klette and Griliches (1996) approach by the light of the contribution by Melitz (2000). To this aim, assume a generic CES demand function with common elasticity of substitution σ among any two varieties

$$U = U\left(\left[\sum_{i=1}^{n_t} Y_{it}^{\rho}\right]^{\frac{1}{\rho}}, Z\right) \tag{5.76}$$

where n_t denotes the number of firms (i.e. varieties), Z captures aggregate demand (i.e. preference) shifts, and $\rho = \frac{\sigma-1}{\sigma}$. Assuming that $U(\cdot)$ is differentiable and quasi-concave, the inverse demand function faced by each firm is (in logs):

$$y_{it} = \frac{p_{it} - p_t}{1-\rho} + \tilde{\tilde{r}}_t. \tag{5.77}$$

In (5.77): p_{it} and p_t denote respectively (the log of) the price set by firm i at time t and the industry deflator (i.e. average price index); $\tilde{\tilde{r}}_t = [(r_t - p_t) - n_t]$ denotes average firm deflated sales; r_t indicates average sales at time t. Using (5.77) and (5.41), the ratio of the firm price to the industry deflator can be expressed as

$$q_{it} = p_{it} - p_t = (1-\rho)\left[\tilde{\tilde{r}}_t - \mathbf{x}_{it}\beta - a_{it}\right]. \tag{5.78}$$

Equation (5.78) identifies the sources of the omitted price bias in a monopolistic competition framework. In order to take advantage of this information, we can use equation (5.78) together with equation (5.73) to purge deflated sales from the unobserved output:

$$\tilde{r}_{it} = \rho\ y_{it} + (1-\rho)\ \tilde{\bar{r}}_t \tag{5.79}$$

where y_{it} is defined as in (5.41). Equation (5.79) is an estimable equation and, as we assumed the absence of simultaneity, consistent estimates of β, no longer influenced by σ, can now be obtained through the OP or LP procedure.

Two papers, De Loecker (2007b) and Del Gatto et al. (2008), have independently implemented this correction into, respectively, the OP and the LP framework. Let us base the following exposition on Del Gatto et al. (2008), with the understanding that the changes to the original procedures needed in the two cases are indeed very similar.[52]

Relying on LP only requires a new invertibility condition, needed in order to express a_{it} as a function of intermediate inputs, as in equation (5.59). Under perfect competition this simply required intermediate input use to be increasing in TFP conditional on capital. Melitz (2000) shows that, under monopolistic competition, this 'monotonicity condition' holds whenever more productive firms do not set disproportionately higher markups than less productive firms.[53] The PDC procedure suggested by equation (5.79) entails that, in addition to the usual input vector, a further regressor, average firm deflated sales $\tilde{\bar{r}}$, takes now part of the estimation in the first stage. It is worth noting that this procedure provides the estimated elasticity of substitution $\hat{\sigma}$ as a by-product.

Once β and ρ have been estimated under this specification (let us call $\hat{\beta}^{PDC}$ the PDC estimated vector of production coefficients), the PDC estimated productivity evaluates to:

$$\hat{a}_{it}^{PDC} = -\mathbf{x}_{it}\hat{\beta}^{PDC} + \frac{1}{\hat{\sigma}-1}\left(\tilde{\bar{r}}_t - \hat{\sigma}\tilde{r}_{it}\right). \tag{5.80}$$

Since q_{it} is now correctly identified, the estimated productivity in equation (5.80) can be intended as the 'true' productivity (i.e. $\hat{a}_{it}^{PDC} = a_{it}$).

To isolate the omitted price bias, note that neglecting price dispersion would imply that $\hat{a}_{it}^{LP} = \tilde{r}_{it} - \mathbf{x}_{it}\hat{\beta}^{LP} = \hat{a}_{it}^{PDC} - \mathbf{x}_{it}\omega$. Hence, we have:

$$\hat{a}_{it}^{LP} = a_{it} - \frac{1}{2\left(\hat{\sigma}-1\right)}(r_{it} - \bar{r}_t) \tag{5.81}$$

where $\bar{r}_t = r_t - n_t$ stands for industry average revenues and we have used the fact that $\hat{a}_{it}^{PDC} = a_{it}$ and $\mathbf{x}_{it}(\hat{\beta}^{PDC} - \hat{\beta}^{LP}) = \mathbf{x}_{it}\omega$ by (5.75). Hence, the 'correction factor' to be applied to the standard LP estimate $\hat{a}_{it}^{LP}$ is an increasing function of a firm's revenues relative to the industry average. This correction factor is positive for above average firms and negative for below average ones. In other words, *disregarding price dispersion results in understating the productivity of firms that are more productive than the average and overstating the productivity of firms that are less productive than the average.*[54] Moreover, the magnitude of the bias depends on the estimated elasticity of substitution $\hat{\sigma}$: the lower the elasticity of substitution (i.e. the more differentiated the products), the larger the bias. Finally, if one were interested in average productivity, correcting for the omitted price bias would be uninfluential, as the bias vanishes on average. The same does not apply to the dispersion measures, which are crucially affected by the bias.

5.3 Estimating (and decomposing) TFP from macro or micro data: Stochastic Frontiers Analysis

The econometric models reviewed so far in this section ignore the contribution of efficiency change to productivity change. An alternative way of estimating TFP — in macro as well as micro contexts — is based

[52] Apart from the baseline approach, the only true difference between the two papers is that De Loecker (2007b) follows more strictly the Melitz's proposal by dealing explicitly with the presence of multi-product firms.

[53] Formally, this requires the elasticity of the markup with respect to productivity to be bounded above by ρ.

[54] This is consistent with the evidence reported by Foster et al. (2005), who have the rare chance of comparing, for several industries, the estimated productivity outcomes resulting from either firm output or firm deflated sales. Although they find that the two measures are highly correlated, they show that quantity-based productivity measures exhibit greater dispersion than revenue-based ones.

on stochastic frontier models. Originally proposed by Aigner et al. (1977), Meeusen and van den Broeck (1977), and Battese and Corra (1977), the estimation of stochastic frontiers represents a well established empirical tool, widely employed in the last three decades by scholars interested in efficiency analysis. On the other hand, its application to the study of TFP growth represents a more recent advance. As in the case of DEA, the existence of technical inefficiency (a discrepancy between observed and potential output) is assumed. This assumption allows one to decompose productivity changes into two parts: the change in technical efficiency (movements towards the frontier) and technical progress (the shift of the frontier over time). In contrast to DEA, the analysis is pursued in a stochastic context.

Although many SFA studies estimate productivity change using either cost or profit functions exploiting the duality theory, in order to make the discussion comparable with that of the econometric models reviewed above, we confine the analysis to the case of production frontiers (primal approach). We first introduce a simple cross section stochastic frontier model which gives the flavor of the departures of SFA from *Non-Frontier* models and focuses on the estimation of technical inefficiency. Then, we describe the stochastic frontier approach to the decomposition of TFP in a panel context proposed by Kumbhakar (2000).

Given I producers each using $X \in R_+^N$ inputs to produce a scalar output $Y \in R_+$, a frontier production model takes the following generic form:

$$Y_i = f(X_i;\beta)\exp(v_i - u_i) \tag{5.82}$$

where β is the vector of unknown parameters to be estimated that characterize the structure of the technology, $f(X_i;\beta)$ defines a deterministic production frontier common to all I producers and the random error term $v_i \gtreqless 0$ captures the effect of (producer-specific) external shocks on observed output Y_i. The stochastic production frontier $f(X_i,\beta)\exp(v_i)$ defines maximum feasible output in an environment characterized by the presence of either favorable or unfavorable events beyond the control of producers. The error term $u_i \geqq 0$ is introduced in the model in order to capture shortfall of Y_i from $f(X_i,\beta)\exp(v_i)$, i.e. technical inefficiency.

According to the output-oriented definition of technical efficiency (TE), we can write:

$$TE_i = \frac{Y_i}{f(X_i;\beta)\exp(v_i)} = \exp(-u_i) \leqslant 1 \tag{5.83}$$

that is, producer i achieves maximum feasible output if and only if $TE_i = 1$, otherwise technical inefficiency occurs and $TE_i < 1$ measures the shortfall of Y_i from maximum feasible output in an environment characterized by the presence of noise. The log-linear version of (5.82) to be estimated with the ultimate objective of obtaining an estimate of technical efficiency is:[55]

$$y_i = \alpha + \beta x_i + v_i - u_i \tag{5.84}$$

Estimating technical efficiency defined in (5.83) requires the estimation of (5.84) in order to obtain estimates of the technology parameters β and to separate estimates of v_i and u_i. In turn, this requires to impose distributional and independence assumptions on the two error components. As Fried et al. (2008, p. 37) point out, the price to pay for obtaining separate estimates of the two error components in (5.84) is indeed the imposition of distributional and independence assumptions in the estimated model. The conventional assumption of $v_i \sim N(0,\sigma_v^2)$ holds in frontier models, while variants of them have been developed in order to accommodate for alternative distributional assumptions on u_i. In particular, Battese and Corra (1977) assumed u_i to follow a half-normal distribution — $u_i \sim N^+(0,\sigma_u^2)$ — Meeusen and van den Broeck (1977) an exponential one, while Aigner et al. (1977) considered both assumptions. Later, Stevenson (1980) and Greene (1980a, 1980b) assumed u_i to follow the more flexible truncated normal and gamma distributions respectively. It is worth noting that the fact that the selection of a particular distribution for the u_i term is not grounded on an a-priori justification represents a common criticism to frontier models. Moreover, either distributional assumption implies that the composed error $e_i = v_i - u_i$ in (5.84) is negatively skewed which prevents from OLS estimation and makes it necessary MLE. OLS, as a matter of fact, neither provides consistent estimates of all β nor is able to deliver an estimate of technical efficiency.

If the modal value of inefficiency is close to zero and relatively high efficiency is expected to be more likely than relatively low efficiency, then the half-normal distributional assumption on u_i will be appropriate,

[55] In general, frontier models are often referred to as "composed error models"for the presence of the composite error term $e_i = v_i - u_i$.

and indeed it is the most widely used in empirical applications. Accordingly, model (5.84) is completed by the following assumptions:[56]

(i) $v_i \sim N(0, \sigma_v^2)$;

(ii) $u_i \sim N^+(0, \sigma_u^2)$;

(iii) u_i and v_i are distributed independently of each other and of the regressors.

Given these assumptions, is then possible to define the log-likelihood function to be maximized with respect to parameters $(\beta, \sigma_v^2, \sigma_u^2)$ and to obtain consistent estimates of all parameters.[57]

Two alternative parameterisations of the log-likelihood function have been proposed by Aigner et al. (1977) and Battese and Corra (1977). Aigner et al. (1977) express the log-likelihood function in terms of the two parameters $\sigma^2 \equiv \sigma_u^2 + \sigma_v^2$ and $\lambda \equiv \sigma_u/\sigma_v$. On the other hand, Battese and Corra (1977) provide a parameterisation of the log-likelihood function in terms of the variance parameter $\gamma \equiv \sigma_u^2/\sigma^2$. The latter parameterisation of the log-likelihood function allows an easy way of testing the frontier model (5.84) vs its *Non-Frontier* version (with no inefficiency effects). Indeed, the parameter γ takes values between 0 and 1, with $\gamma = 0$ ($\gamma = 1$) indicating that the deviations from the frontier are entirely due to statistical noise (technical inefficiency). For details on the test of the null hypothesis that $H_0 : \gamma = 0$ (no scope for the frontier model), the reader is referred, for instance, to Coelli et al. (1998, pp. 190-192).[58]

Once one has obtained ML estimates of all parameters, technical efficiency has to be estimated for each of the i's observed production units. Jondrow et al. (1982) were the first to deliver a result. They noticed that the definition of technical efficiency in (5.83) involves the unobservable technical inefficiency component u_i. This implies that "even if the true value of the parameter vector β was known, only the difference $e_i = v_i - u_i$ could be observed"(Coelli et al., 1998, p. 190) and that the best prediction for u_i is the conditional expectation of u_i, given the value of e_i:

$$E[u_i|e_i] = \frac{\sigma\lambda}{(1+\lambda^2)}\left[\frac{\phi(e_i\lambda/\sigma)}{\Phi(-e_i\lambda/\sigma)} - \frac{e_i\lambda}{\sigma}\right] \tag{5.85}$$

where $e_i = v_i - u_i$, $\phi(\cdot)$ is the density of the standard normal distribution, $\Phi(\cdot)$ is the cumulative density function, λ is defined as above and $\sigma = (\sigma_u^2 + \sigma_v^2)^{1/2}$.

Then, since $1 - u_i$ is a first-order approximation to the infinity series $\exp(-u_i) = 1 - u_i + u_i^2/2 + u_i^3/3!..$ they suggested to estimate technical efficiency defined in (5.83) as $TE_i = \exp(-E[u_i|e_i]) = 1 - E[u_i|e_i]$.[59]

Later, Battese and Coelli (1988) proposed the following alternative point estimator for technical efficiency:

$$E[\exp(-u_i)|e_i] = \frac{1 - \Phi[\delta + (\gamma e_i/\delta)]}{1 - \Phi(\gamma e_i/\delta)} \exp[\gamma e_i + (\delta^2/2)] \tag{5.86}$$

where $\delta = \sqrt{\gamma(1-\gamma)\sigma^2}$ and γ is defined as above. Notice here that since $\exp(-E[u_i|e_i]) \neq E[\exp(-u_i)|e_i]$, (5.85) and (5.86) deliver different results. Furthermore, neither is a consistent estimate of technical efficiency, since the variance of $E[u_i|e_i]$ does not go to zero as the size of the cross section increases.

The above cross-sectional production frontier model has been extended to panel data under alternative assumptions on the distribution of the inefficiency term as well as on its behavior over time. Pitt and Lee (1981) specified a panel data version of (5.84) under the assumption of time-invariant half-normal distributed inefficiency effects, while Kumbhakar (1987) and Battese and Coelli (1988) extended Pitt and Lee's model to the case of normal-truncated time invariant u_i. Then, Schmidt and Sickles (1984) were the first to use conventional panel data techniques in a frontier context. The work by Schmidt and Sickles (1984) represents a contribution of particular importance within the frontier literature as it provides a full picture of the advantages associated with the use of panel data versus cross-sectional data in frontier

[56] This model is known in the literature as the normal-half-normal model.

[57] See Greene (2008, pp. 189-190) for a discussion on the estimation tools available in computer software.

[58] Coelli et al. (1988) also provide useful operational information on hypothesis testing of alternative distributional assumptions on u_i and other mis-specification issues of model (5.84).

[59] Jondrow et al. (1982) also delivered the expected value of u_i conditional on the composed error term under the exponential distributional assumption.

models. These advantages can be summarized as follows. First, cross-section models require the imposition of the independence assumption between the u_i's and input variables,[60] while panel data estimations do not. Second, panel data frontier models deliver consistent estimates of the inefficiency term. Third, Schmidt and Sickles (1984) observed that when panel data are available, there is no need for any distributional assumption for the inefficiency effects and all the relevant technological parameters can be obtained by traditional panel data estimation procedures, in both variants of fixed and random-effects.[61]

The assumption of time-invariant inefficiency effects it is not easy to justify in long-term panel data as one would expect producers to observe past inefficient behavior and possibly correct for non price and organizational inefficiency determinants. The first extension of the Schmidt and Sickles (1984) model which accommodates for time-varying inefficiency effects was developed by Cornwell et al. (1990). Since then, alternative specifications of time-varying technical inefficiency terms proposed in the literature have been the following:

(i) Kumbhakar (1990) and Battese and Coelli (1992) assume that inefficiency evolves according to a parametric function of time: $u_{it} = u_i\alpha(t)$. In both works, a non-linear specification is used. In Kumbhakar (1990), technical inefficiency effects vary according to $u_{it} = u_i[1 + \exp(bt + ct^2)]^{-1}$, where $u_i \sim N^+(0, \sigma_u^2)$ and b and c are unknown parameters to be estimated. Battese and Coelli (1992) assume $u_{it} = u_i\exp[-\eta(t - T)]$, where $u_i \sim N^+(\mu, \sigma_u^2)$ and η is unknown and to be estimated;

(ii) Lee and Schmidt (1993) assume $u_{it} = u_i d_t$, where d_t are time effects represented by time dummies and u_i are either fixed or random producer-specific effects and no assumption is imposed on the temporal pattern of inefficiency;

(iii) Cornwell et al. (1990) assume $u_{it} = a_{1i} + a_{2i}t + a_{3i}t^2$.

Nishimizu and Page (1982) firstly worked out a decomposition of TFP change in order to obtain a measure of the contribution of technical efficiency change assuming constant return to scale. Later, Kumbhakar (2000) refined their decomposition of TFP change also accounting for time-varying scale effects and changes of allocative inefficiency over time.

Following Kumbakhar (2000), the Solow residual defined in (4.3) attainable within frontier models can be estimated and decomposed as follows. Consider the following production function:

$$Y_{it} = f(X_{it}, t)\exp(-u_{it}) \tag{5.87}$$

where $i = 1, ..., N$ producers are observed over $t = 1, ..., T$ years, Y, $f(\cdot)$ and $\exp(-u_{it})$ are interpreted as above in this section and time is included as a regressor in the production function in order to capture technical change. Omitting the i and t subscripts and totally differentiating $\ln Y$ with respect to time:

$$\frac{\mathrm{d}\ln Y}{\mathrm{d}t} = \frac{\mathrm{d}\ln f(X,t)}{\mathrm{d}t} - \frac{\partial u}{\partial t} \tag{5.88}$$

Totally differentiating $\ln f(X,t)$ with respect to time:[62]

$$\begin{aligned}\frac{\mathrm{d}\ln f(X,t)}{\mathrm{d}t} &= \frac{\partial \ln f(X,t)}{\partial t} + \sum_j \frac{\partial f(X,t)}{\partial X_j}\cdot\frac{\mathrm{d}Xj}{\mathrm{d}t} \\ &= \frac{\partial \ln f(X,t)}{\partial t} + \sum_j \epsilon_j \cdot \dot{X}_j \end{aligned} \tag{5.89}$$

and replacing (5.89) in (5.88) is then possible to obtain the following decompoition of output growth:

$$\dot{y} = \frac{\partial \ln f(X,t)}{\partial t} + \sum_j \epsilon_j \cdot \dot{X}_j - \frac{\partial u}{\partial t} \tag{5.90}$$

Notice that equation (5.90) distinguishes three sources of output growth:

[60] "The independence assumption is essential to the MLE procedure", Fried et al. (2008, p. 37).

[61] For surveys of panel data production frontier models see Kumbhakar and Lovell (2000) and Greene (2008).

[62] Notice that $\partial \ln f(X,t)/\partial \ln X_j$ defines the output elasticity ϵ_j of input X_j at the frontier.

(i) $TC = \partial \ln f(X,t)/\partial t \Rightarrow$ exogenous Technical Change (TC). That is to say, given a certain inputs use, if $TC > 0$ ($TC < 0$), exogenous TC shifts the production frontier upward (downward);

(ii) $TEC = -\partial u/\partial t \Rightarrow$ Technical Efficiency Change (TEC). TEC represents the rate at which an inefficient producer moves towards the frontier (technical efficiency declines over time if $TEC < 0$);

(iii) $\sum_j \epsilon_j \cdot \dot{X}_j \Rightarrow$ change in input use. It is worth noting that if input quantities do not change over time, then $\dot{y} = TC + TEC$.

The decomposition of output growth defined in (5.90) can be replaced in the Solow residual defined in equation (4.3):[63]

$$T\dot{F}P = TC - \frac{\partial u}{\partial t} + \sum_j (\epsilon_j - s_j)\dot{X}_j \tag{5.91}$$

Using the measure of return to scale $RTS = \sum_j \epsilon_j$ (i.e. the assumption of constant return to scale holds only if $RTS = 1$) and defining $\lambda_j = f_j X_j / \sum_k f_k X_k = \epsilon_j / \sum_k \epsilon_k = \epsilon_j / RTS$, where f_j is the marginal product of the jth input, equation (5.91) can be rewritten as follows:

$$T\dot{F}P = TC - \frac{\partial u}{\partial t} + (RTS - 1)\sum_j \lambda_j \dot{X}_j + \sum_j (\lambda_j - S_j)\dot{X}_j \tag{5.92}$$

The first and second terms at the right-hand side of equation (5.92) are interpreted as above, while the third and forth terms represent scale effects and price effects, respectively. The contribution of scale effects to TFP change depends on both technology and on factor accumulation. In the case of constant return to scale ($RTS = 1$), the third term at the right-hand side of equation (5.92) cancels out. On the other hand, if $RTS \neq 1$, a share of TFP change can be potentially attributed to changes in the scale of production. For instance, in the case of increasing return to scale, an increase in the amount of inputs contribute positively to TFP change, while reducing the amount of inputs will cause a lower TFP change.[64] The price effects component reflects the contribution of changes in allocative efficiency to TFP change. Indeed, the forth term at the right-hand side of equation (5.92) captures either deviations of input prices from the value of their respective marginal products, or departure of the marginal rate of technical substitution from the ratio of input prices.

Kumbhakar (2000) shows how to estimate the four components of TFP change in (5.92) in a translog production frontier model under the two alternative (i) and (ii) assumptions on time-varying inefficiency effects u_{it}'s mentioned at page 36, using the following translog production function:

$$\ln Y_{it} = a_0 + \sum_j a_j \ln X_{jit} + a_t t + \frac{1}{2}\sum_j \sum_k a_{jk} \ln X_{jit} \ln X_{kit} + \frac{1}{2} a_{tt} t^2 + \sum_j a_{jt} \ln X_{jit} t + v_{it} - u_{it} \tag{5.93}$$

The first variant of the model is based on the assumption that the temporal pattern of inefficiency is described by $u_{it} = u_i \alpha(t)$. Given this, provided that $v_{it} \sim i.i.d.N(0, \sigma_v^2)$, $u_i \sim i.i.d.N^+(\mu, \sigma_u^2)$ and that v_{it} are independent of u_i for any i and t, it is possible to derive the log-likelihood function for (5.93) and to obtain ML estimators of the technological parameters, all the parameters in $\alpha(t)$, σ_v^2, σ_u^2 and μ. Then, estimates of u_{it} can be obtained by using either (5.85) or (5.86). Finally, the four components of TFP change for each producer at each point in time can be computed on the basis of the following estimates:

$$RTS = \sum_j \epsilon_j = \sum_j (a_j + \sum_k a_{jk} \ln X_k + a_{jt} t)$$

$$\lambda_j = \epsilon_j / RTS$$

[63] Notice that we have now J inputs.

[64] The inverse reasoning applies to the case of decreasing returns to scale.

$$TC = a_t + a_{tt}t + \sum_j a_{jt} \ln X_j \tag{5.94}$$

$$\frac{\partial u_{it}}{\partial t} = u_i \frac{\partial \alpha(t)}{\partial t}$$

In the second variant of model (5.93), inefficiency varies over time according to the expression $u_{it} = u_i d_t$ (see assumption (ii), page 36). This assumption implies the advantages over the first variant of the model of not imposing any functional form for the temporal pattern of inefficiency. Further, the model can be estimated by non-linear least squares without distributional assumptions on the v error term and — in a fixed effect specification — the u_i's can be calculated from individual dummies. Once obtained all parameters, technical inefficiency terms will be obtained from $\hat{u}_{it} = \max_i\{u_i d_t\} - u_i d_t$ and technical efficiency change will be defined by $\hat{u}_{it} - \hat{u}_{it-1}$. Finally, expressions (5.94) can be implemented for the calculation of all other components of TFP change.

6 Conclusions

There is an extensive and still rapidly evolving literature on productivity estimates and an exhaustive account of it is certainly beyond the scope of this paper. This survey reviews most of the available methodologies for productivity estimation and suggests a scheme to classify the different approaches used to estimate productivity. The first classification criterion discriminates between deterministic and econometric estimation strategies while a second one discriminates between *Frontier Approaches* and *Non-Frontier Approaches*. Moreover, we identify if a specific methodology has been applied only to macro context (using countries/regions or industry data), to micro (firms/plant) datasets or to both.

Recent trends in empirical analysis of TFP show a growing attention away from the study of TFP at the aggregate and industry level of detail and towards the firm/plant level. Despite that, both macro and micro TFP analysis are still investigated. The two strands of literature are developing along different lanes and are rather difficult to compare. While firm analysis enables to investigate TFP patterns at a deeper level controlling for non-competitive markets, increasing returns and heterogeneous firms issues, their results may be hard to generalise, and aggregate analysis still plays an important role in cross-country comparative analysis. In general, the links between the micro and macro levels of TFP analysis need to be further developed and this may be considered as one of the main challenges currently facing this literature.

References

[1] Abramovitz M. 1956. Resource and Output Trends in the United States Since 1870. *American Economic Review*, 46(2): 5-23.

[2] Acemoglu D., Aghion P. and Zilibotti F. 2006. Distance to Frontier, Selection, and Economic Growth. *Journal of the European Economic Association*, 4(1): 37-74.

[3] Ackerberg D., Caves K. and Frazer G. 2006. Structural identification of production functions. Mimeo, UCLA.

[4] Aghion P. and Howitt P. 1992. A model of growth through creative destruction. *Econometrica*, 60: 323-51.

[5] Aghion P., Dewatripont M. and Rey P. 1999. Competition, Financial discipline and Growth. *Review of Economic Studies*, 66: 825-52.

[6] Aigner J., Lovell K., Schmidt, J. 1977. Formulation and estimation of stochastic frontier production function models. *Journal of Econometrics*, 6: 21-37.

[7] Aiyar S. and Dalgaard C. 2005. Total Factor Productivity Revisited: A Dual Approach to Development Accounting IMF Staff Papers, Vol 52.

[8] Aiyar S. and Feyrer J. 2002. A Contribution to the Empirics of Total Factor Productivity. Mimeo, IMF, Washington.

[9] Amemiya T. 1967. A Note on the Estimation of Balestra-Nerlove Models. Institute for Mathematical Studies in Social Sciences Technical Report No. 4, Stanford University, Stanford.

[10] Arellano M. and Bond S. 1991. Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations. *Review of Economic Studies*, 58: 277-297

[11] Arnold J. 2005. Productivity Estimation at the Firm Level. A Practical Guide. Mimeo, Bocconi University, Milano.

[12] Aw B.Y., Chung S., and Roberts M. 2003. Productivity, Output, and Failure: A Comparison of Taiwanese and Korean Manufacturers. *Economic Journal*, 113: 443-705.

[13] Baier S.L., Dwyer G.P., and Tamura R. 2006. How Important are Capital and Total Factor Productivity for Economic Growth? *Economic Inquiry*, 44(1): 23-249.

[14] Baily M.N., Bartelsman E.J., and Haltiwanger J. C. 1996. Downsizing and Productivity Growth: Myth or Reality? *Small Business Economics*, 8(4): 259-278.

[15] Baltagi B.H. 2003. *Econometric Analysis of Panel Data*. John Wiley and Sons, Chichester.

[16] Barro R. and Sala-i-Martin X. 2004. *Economic growth*. MIT Press.

[17] Bartelsman E.J. and Dhrymes P.J. 1998. Productivity dynamics: U.S. manufacturing plants, 1972-1986. *Journal of Productivity Analysis*, 9(1): 5-34.

[18] Bartelsman E. J. and Doms M. 2000. Lessons from Longitudinal Microdata. *Journal of Economic Literature*, 38(3): 569-594.

[19] Battese G. and Corra G. 1977. Estimation of a production frontier model: with application to the pastoral zone of Eastern Australia. *Australian Journal of Agricultural Economics*, 21: 169-179.

[20] Battese G. and Coelli T. 1988. Prediction of firm-level technical efficiencies with a generalized frontier production function and panel data. *Journal of Econometrics*, 38: 387-399.

[21] Battese G. and Coelli T. 1992. Frontier production functions, technical efficiency and panel data: with application to paddy farmers in India. *Journal of Productivity Analysis*, 3: 153-169.

[22] Benhabib J. and Spiegel M.M. 1994. The role of human capital in economic development. Evidence from aggregate cross-country data. *Journal of Monetary Economics*, 34: 143-173.

[23] Benhabib J, Spiegel M.M. 2005. Human capital and technology diffusion. In: Aghion P, Durlauf S. (Eds) *Handbook of Economic Growth*, Volume 1A: 935-966. North-Holland. Amsterdam.

[24] Bernard A, Jones C. 1996. Technology and convergence. *Economic Journal,* 106: 1037-44.

[25] Bernard A., and Jensen B. 1999. Exceptional Exporter Performance: Cause, Effect, or Both? *Journal of International Economics*, 47: 1-25.

[26] Bernard A., Eaton J., Jensen B and Kortum S. 2003. Plants and Productivity in International Trade. *American Economic Review*, 93: 1268-1290.

[27] Bernard A., Jensen B., and Schott P. 2006. Trade costs, firms and productivity. *Journal of Monetary Economics*, 53: 917–937.

[28] Bernard A., Redding S. and Schott P. 2007. Comparative Advantage and Heterogeneous Firms. *Review of Economic Studies*, 74: 31–66.

[29] Blundell R.W. and Bond S.R. 1998. Initial Conditions and Moment Restrictions in Dynamics Panel Data Models. *Journal of Econometrics*, 87: 115-143.

[30] Bosworth B. and Collins S. 2003. The Empirics of Growth: An Update. *Brookings Papers on Economic Activity*, 2003(2): 113-179.

[31] Bun M.J.G. and Carree M.A. 2005. Bias-corrected estimation in dynamic panel data models. *Journal of Business & Economic Statistics*, 23(2): 200-210.

[32] Carlaw K. and Lipsey R. 2003. Productivity, Technology and economic growth: what is the relationship? *Journal of Economic Survey*, 17(3): 457-495.

[33] Caselli F. and Wilson D.J. 2004. Importing Technology. *Journal of Monetary Economics*, 51(1): 1-32.

[34] Caselli F., Esquivel G. and Lefort F. 1996. Reopening the Convergence Debate: a New Look at Cross Country Growth Empirics. *Journal of Economic Growth*, 1: 363-89.

[35] Caselli F. 2005. Accounting for Cross-Country Income Differences. In: Aghion P. and Durlauf S. (Eds.) *Handbook of Economic Growth*, Volume 1A. North-Holland. Amsterdam.

[36] Caves D.W., Christensen L.R. and Diewert W.E. 1982a. The Economic Theory of Index Numbers and the Measurement of Input, Output, and Productivity. *Econometrica*, 50(6): 1393-1414.

[37] Caves D.W., Christensen L.R. and Diewert E.W. 1982b. Multilateral Comparisons of Output, Input, and Productivity using Superlative Index Numbers. *Economic Journal*, 92: 73-86.

[38] Chamberlain G. 1982. Multivariate Regression Models for Panel Data. *Journal of Econometrics*, 18: 5-46.

[39] Chaney, Thomas. 2008. Distorted Gravity: The intensive and Extensive Margins of International Trade. *American Economic Review*, forthcoming.

[40] Charnes A., Cooper W.W. and Rhodes E. 1978. Measuring the Efficiency of Decision Making Units. *European Journal of Operational Research*, 2: 429-444.

[41] Clerides S., Lach S. and Tybout J. 1998. Is Learning By Exporting Important? Micro-Dynamic Evidence From Colombia, Mexico and Morocco. *Quarterly Journal of Economics*, 113: 903-947.

[42] Coelli, T. J. 1996. A guide to DEAP version 2.1: A Data Envelopment Analysis Computer program. *CEPA Working Papers* No. 8/96, University of New England, Armidale, Australia.

[43] Coelli T., Rao D. and Battese E. 1998. *An Introduction to Efficiency and Productivity Analysis*. Kluwer Academic.

[44] Cornwell C., Schmidt P. and Sickles R. 1990. Production frontiers with cross-sectional and time-series variation in efficiency levels. *Journal of Econometrics*, 46: 185-200.

[45] Daraio C. and Simar L. 2007. *Advanced Robust and Nonparametric Methods in Efficiency Analysis: Methodology and Applications*. Springer, New York.

[46] Del Gatto M., Pagnini M., and G.I.P. Ottaviano. 2008. Openness to Trade and Industry Cost Dispersion. *Journal of Regional Science*, 48(1): 97-129.

[47] De Loecker J. 2007a. Do exports generate higher productivity? Evidence from Slovenia. *Journal of International Economics*, 73: 69-98.

[48] De Loecker J. 2007b. Product differentiation, multi-product firms and estimating the impact of trade liberalization on productivity, *NBER Working Paper* No. 13155.

[49] Denison E. 1985. *Trends in American Economic Growth, 1929-1982*. Brookings Institution Press.

[50] Deprins D., Simar L. and Tulkens, H. 1984. Measuring Labor Inefficiency in Post Offices. In: Marchand, M., Pestieau P. and Tulkens H. (Eds.) *The Performance of Public Enterprizes: Concepts and Measurements*: 243-267. North-Holland. Amsterdam.

[51] Diewert W. 1976. Exact and superlative index numbers. *Journal of Econometrics*, 4: 115-145.

[52] Di Liberto A., Pigliaru F. and Mura R. 2008. How to measure the unobservable: a panel technique for the analysis of TFP convergence. *Oxford Economic Papers*, 60: 343-382.

[53] Domar E. 1961. On the measurement of technological change. *Economic Journal*, 71: 586-588.

[54] Douglas P.H. 1948. Are there Laws of Production? *American Economic Review*, 38: 1-41.

[55] Eaton J. and S. Kortum. 2002. Technology, Geography, and Trade. *Econometrica*, 70: 1741-1779.

[56] Everaert G. and Pozzi L. 2007. Bootstrap-based bias correction for dynamic panels. *Journal of Economic Dynamics and Control*, 31: 1160-84.

[57] Fadinger H. and P. Fleiss. 2008. Trade and Sectoral Productivity. Mimeo.

[58] Färe R., Grosskopf S., Lindgren B. and Ross P. 1994a. Productivity Developments in Swedish Hospitals: A Malmquist Output Index Approach. In: Charnes A., Cooper W., Lewin A. and Seiford L. (Eds.) *Data envelopment analysis: theory, methodology and applications*. Kluwer Academic Publishers.

[59] Färe R., Grosskopf S., Norris M., Zhang Z. 1994b. Productivity Growth, Technical Progress, and Efficiency Change in Industrialized Countries. *American Economic Review*, 84(1): 66-83.

[60] Farrell M.J. 1957. The Measurement of Productive Efficiency, *Journal of the Royal Statistical Society, Series A*, 120: 253-90.

[61] Finicelli A., Pagano P., and M. Sbracia. 2008. Trade-Revealed TFP. Mimeo.

[62] Foster, L., J. Haltiwanger, and C. Krizan. 2001. Aggregate productivity growth: Lessons from microeconomic evidence. In: C.R. Hulten, E.R. Dean, and M.J. Harper (Eds.) *New Developments in Productivity Analysis*, Volume 63, NBER Studies in Income and Wealth: 303-63. University of Chicago Press. Chicago.

[63] Foster, L., J. Haltiwanger, and C. Syverson. 2005. Reallocation, Firm Turnover, and Efficiency: Selection on Productivity or Profitability? *Center for Economic Studies Working Paper*, No. 05-11.

[64] Fried H.O., Lovell C.A.K. and Schmidt S.S. (Eds) 2008. *The measurement of productive efficiency and productivity growth*. Oxford University Press.

[65] Greene W. 1980a. Maximum likelihood estimation of econometric frontier functions. *Journal of Econometrics*, 13: 27-56.

[66] Greene W. 1980b. On the estimation of a flexible frontier production model. *Journal of Econometrics*, 13: 101-115.

[67] Greene W. 2008. The econometric approach to efficiency analysis. In: Fried H.O., Lovell C.A.K. and Schmidt S.S. (Eds.) *The measurement of productive efficiency and productivity growth.*: 92-250. Oxford University Press.

[68] Greenwood J. and Krusell P. 2007. Growth accounting with investment-specific technological progress: A discussion of two approaches. *Journal of Monetary Economics*, 54: 1300-1310.

[69] Griliches Z. and J. Mairesse 1995. Production Functions: the Search for Identification. *NBER Working Paper* No. 9617.

[70] Grosskopf S. 1993. Efficiency and Productivity. In: Fried H.O., Lovell C.A.K., Schmidt S.S. (Eds) *The Measurement of Productive Efficiency: Techniques and Applications*: 160-194. Oxford University Press.

[71] Hall R.E. 1988. The relation between price and marginal cost in U.S. industry. *Journal of Political Economy*, 96: 921-947.

[72] Hall R.E. and Jones C.I. 1999. Why do Some Countries Produce so Much More Output per Worker than Others? *Quarterly Journal of Economics*, 114: 83-116.

[73] Hercowitz Z. 1998. The embodiment controversy: A review essay. *Journal of Monetary Economics*, 41: 217-224.

[74] Hollingsworth, B. 2004. Non Parametric Efficiency Measurement. *Economic Journal*, 114: 307-311

[75] Hsieh C. 1999. Factor Prices and Productivity Growth in East Asia. *American Economic Review*, 89(2): 133-138.

[76] Hulten C. 2001. Total Factor Productivity: a short biography. In: Hulten C. Dean E. and Harper M. (Eds) *New Developments in Productivity analysis*. The University of Chicago Press.

[77] Hulten C. 1978. Growth accounting with intermediate inputs. *Review of Economic Studies*, 45: 511-518.

[78] Islam N. 2003. Productivity dynamics in a large sample of countries: a panel study. *Review of Income and Wealth*, 49: 247-72.

[79] Islam N. 1995. Growth Empirics: a Panel data Approach. *Quarterly Journal of Economics*, 110: 1127-70.

[80] Jondrow J., Lovell C.A.K., Materov I.S. and Schmidt P. 1982. On the estimation of technical inefficiency in the stochastic frontier production function model *Journal of Econometrics*, 19: 233-238.

[81] Jorgenson D. 2005. Accounting for Growth in the Information Age. In: Aghion P. and Durlauf S. (Eds.) *Handbook of Economic Growth*: 743-815, Volume 1A. North-Holland, Amsterdam.

[82] Jorgenson D. and Griliches Z. 1967. The explanation of Productivity change. *Review of Economic Studies*, 34(3): 249-83.

[83] Jorgenson D., Ho M. and Stiroh K. 2007. A Retrospective Look at the U.S. Productivity Growth Resurgence, Federal Reserve Bank of New York Staff Reports No. 277.

[84] Judson R. and Owen A. 1999. Estimating dynamic panel data models: a guide for macroeconomists. *Economic Letters*, 65: 9-15.

[85] Katayama H., Lu S. and Tybout J.R. 2003. Why plant-level productivity studies are often misleading, and an alternative approach to inference. *NBER Working Paper* No. 9617.

[86] Kiviet J. 1995. On Bias, Inconsistency, and Efficiency of Various Estimators in Dynamicc Panel Data Models. *Journal of Econometrics*, 68: 53-78.

[87] Klenow P.J. and Rodriguez-Clare A. 1997. The Neoclassical Revival in Growth Economics: Has it Gone too Far? In: Ben S. Bernanke and Julio J Rotemberg (Eds.), NBER Macroeconomics Annual. MIT Press. Cambridge.

[88] Klette T.J. and Griliches Z. 1996. The Inconsistency of Common Scale Estimators when Output Prices are Unobserved and Endogenous. *Journal of Applied Econometrics*, 11(6): 343-361.

[89] Kumar S. and Russell R. 2002. Technological Change, Technological Catch-Up and Capital Deepening: Relative Contributions to Growth and Convergence. *American Economic Review*, 92: 527-48.

[90] Kumbhakar S.C. 1987. The specification of technical and allocative inefficiency of multi-product firms in stochastic production and profit frontiers. *Journal of Quantitative Economics*, 3: 213-223.

[91] Kumbhakar S.C. 1990. Production frontiers and panel data, and time varying technical inefficiency. *Journal of Econometrics*, 46: 201-211.

[92] Kumbhakar S.C. 2000. Estimation and decomposition of productivity change when production is not Efficient. *Econometric Reviews*, 19: 425-60.

[93] Kumbhakar S.C. and Lovell C. 2000. *Stochastic Frontier Analysis*. Cambridge University Press.

[94] Lee Y. and Schmidt P. 1993. A production frontier model with flexible temporal variation in technical efficiency. In: Fried H.O., Lovell C.A.K., Schmidt S.S. (Eds) *The Measurement of Productive Efficiency: Techniques and Applications*. Oxford University Press.

[95] Levinsohn J. and Petrin A. 2003. Estimating Production Functions Using Inputs to Control for Unobservables. *Review of Economic Studies*, 70: 317-341.

[96] Levinsohn J., Petrin A. and Poi B.P. 2003. Production Functions Estimation in Stata Using Inputs to Control for Unobservables. Mimeo.

[97] Maddison A. 1995. *Monitoring the World Economy 1820-1992*. Organisation for Economic Co-Operation and Development, Paris.

[98] Malmquist S. 1953. Index numbers and indifference surfaces. *Trabajos de Estadistica*, 4: 209-242

[99] Mankiw N.G., Romer D. and Weil D.N. 1992. A Contribution to the Empirics of Economic Growth. *Quarterly Journal of Economics*, 107: 407-437.

[100] Marschack J. and Andrews W.H. 1944. Random Simultaneous Equations and the Theory of Production. *Econometrica*, 12: 143-205.

[101] Meeusen W. and Van Den Broeck J. 1977. Efficiency estimation from Cobb-Douglas production functions with composed error. *International Economic Review*, 18: 435-44.

[102] Melitz M. 2000. Estimating Firm-Level Productivity in Differentiated Product Industry. Mimeo.

[103] Melitz M. 2003. The Impact of Trade on Intra-Industry Reallocations and Aggregate Industry Productivity. *Econometrica*, 71: 1695-1725.

[104] Melitz M. and G. Ottaviano. 2005. Market size, trade and productivity. *Review of Economic Studies*, 75: 295–316.

[105] Mincer J. 1974. *Schooling, Earnings and Experience*. Columbia University Press, New York.

[106] Nishimizu M. and Page J.M. 1982. Total factor productivity growth, technical progress and technical efficiency change: dimensions of productivity change in Yugoslavia 1956-78. *Economic Journal*, 92: 930-936.

[107] Olley S. and Pakes A. 1996. The Dynamics of Productivity in the Telecommunications Equipment Industry. *Econometrica*, 64(6): 1263-1297.

[108] Oulton N. 2007. Investment-specific technological change and growth accounting. *Journal of Monetary Economics*, 54: 1290-1299.

[109] Pakes, A. 1994. *The Estimation of Dynamic Structural Models: Problems and Prospects, Part II. Mixed Continuous-Discrete Control Models and Market Interactions*. In: J.J. Laffont and C. Sims. (eds) *Advances in Econometrics: Proceedings of the 6th World Congress of the Econometric Society*, Chapter 5, 171-259.

[110] Pavcnik N. 2002. Trade Liberalization, Exit, and Productivity Improvements: Evidence from Chilean Plants. *Review of Economic Studies*, 69: 245-276.

[111] Parente S.L. and Prescott E.C. 1994. Barriers to Technology Adoption and Development. *Journal of Political Economy* 102: 298-321.

[112] Pitt M. and Lee L. 1981. The measurement and sources of technical inefficiency in the Indonesian weaving industry. *Journal of Development Economics*, 9: 43-64.

[113] Pritchett L. 2000. The Tyranny of Concepts: CUDIE (Cumulated, Depreciated, Investment Effort) Is Not Capital. *Journal of Economic Growth*, 5(4): 361-84.

[114] Roberts M. and Tybout J. 1997. The Decision to Export in Colombia: An Empirical Model of Entry with Sunk Costs. *American Economic Review*, 87: 545-564.

[115] Roeger W. 1995. Can Imperfect Competition Explain the Difference between Primal and Dual Productivity Measures? Estimates for U.S. Manufacturing. *Journal of Political Economy*, 103(2): 316-330.

[116] Schmidt P. and Sickles R. 1984. Production frontiers and panel data. *Journal of Business and Economic Statistics*, 2: 367-374.

[117] Shephard R.W. 1970. *Theory of cost and production functions*. Princeton University Press.

[118] Solow R. 1957. Technical Change and the Aggregate Production Function. *The Review of Economics and Statistics*, 39(3): 312-320.

[119] Stevenson R. 1980. Likelihood functions for generalized stochastic frontier estimation. *Journal of Econometrics*, 13: 58-66.

[120] Syverson, C. 2004. Product substitutability and productivity dispersion. *Review of Economics and Statistics*, 86, 534-550.

[121] Temple J. 2001. Growth Effects of Education and Social Capital in the OECD Countries. *CEPR Discussion Papers* No. 2875.

[122] Van Biesebroeck J. 2005. Exporting raises productivity in sub-Saharan African manufacturing firms. *Journal of International Economics*, 67(2): 373-391.

[123] Van Biesebroeck J. 2007. Robustness of productivity estimates. *The Journal of Industrial Economics*, 55(3): 529-569.

[124] Van Biesebroeck J. 2008. Aggregate and decomposing Productivity. *Review of Business and Economics* LIII(2): 122-146.

[125] Wooldridge J.M. 2005. On Estimating Firm-level Production Functions Using Proxy Variables to Control for Unobservables. Mimeo, Michigan State University.

[126] Young A. 1995. Confronting the Statistical Realities of the East Asian Growth Experience. *Quarterly Journal of Economics*, 110: 641-80.

Ultimi Contributi di Ricerca CRENoS

I Paper sono disponibili in: http://www.crenos.it

08/17 *Edoardo Otranto* "Identifying Financial Time Series with Similar Dynamic Conditional Correlation"

08/16 *Rinaldo Brau, Raffaele Doronzo, Carlo V. Fiorio, Massimo Florio* " Gas Industry Reforms and Consumers' Prices in the European Union: An Empirical Analysis"

08/15 *Oliviero A. Carboni* "The Effect of R&D Subsidies on Private R&D:Evidence from Italian Manufacturing Data"

08/14 *Gerardo Marletto* "Getting out of the car. An institutional/evolutionary approach to sustainable transport policies"

08/13 *Francesco Lisi, Edoardo Otranto,* "Clustering Mutual Funds by Return and Risk Levels"

08/12 *Adriana Di Liberto, Francesco Pigliaru, Piergiorgio Chelucci,* "International TFP Dynamics and Human Capital Stocks: a Panel Data Analysis, 1960-2003"

08/11 *Giorgio Garau, Patrizio Lecca,* "Impact Analysis of Regional Knowledge Subsidy: a CGE Approach"

08/10 *Edoardo Otranto,* "Asset Allocation Using Flexible Dynamic Correlation Models with Regime Switching"

08/09 *Concetta Mendolicchio, Dimitri Paolini, Tito Pietra,* "Investments In Education In A Two-Sector, Random Matching Economy"

08/08 *Stefano Usai,* "Innovative Performance of Oecd Regions"

08/07 *Concetta Mendolicchio, Tito Pietra, Dimitri Paolini,* "Human Capital Policies in a Static, Two-Sector Economy with Imperfect Markets"

08/06 *Vania Statzu, Elisabetta Strazzera,* "A panel data analysis of electric consumptions in the residential sector"

08/05 *Marco Pitzalis, Isabella Sulis, Mariano Porcu,* "Differences of Cultural Capital among Students in Transition to University. Some First Survey Evidences"

08/04 *Isabella Sulis, Mariano Porcu,* "Assessing the Effectiveness of a Stochastic Regression Imputation Method for Ordered Categorical Data"

08/03 *Manuele Bicego, Enrico Grosso, Edoardo Otranto,* "Recognizing and Forecasting the Sign of Financial Local Trends Using Hidden Markov Models"

08/02 *Juan de Dios Tena, Edoardo Otranto,* "A Realistic Model for Official Interest Rates Movements and their Consequences"

08/01 *Edoardo Otranto,* "Clustering Heteroskedastic Time Series by Model-Based Procedures"

07/16 *Sergio Lodde,* "Specialization and Concentration of the Manufacturing Industry in the Italian Local Labor Systems"

07/15 *Giovanni Sulis,* "Gender Wage Differentials in Italy: a Structural Estimation Approach"

07/14 *Fabrizio Adriani, Luca G. Deidda, Silvia Sonderegger,* "Over-Signaling Vs Underpricing: the Role of Financial Intermediaries In Initial Public Offerings"

07/13 *Giovanni Sulis,* "What Can Monopsony Explain of the Gender Wage Differential In Italy?"

07/12 *Gerardo Marletto,* "Crossing the Alps: Three Transport Policy Options"