PRBB 7|26|02
VAN BIESEBROECK
10:15AM

Impact of Measurement Error on Productivity Estimates

Johannes Van Biesebroeck*

University of Toronto

July 11, 2002

[Preliminary - Comments welcome]

Abstract

Researchers interested in estimating productivity can choose from an array of methodologies, each with its strengths and weaknesses. Many methodologies are not very robust to measurement error in inputs. This is particularly troublesome, because fundamentally the objective of productivity measurement is to identify output differences that cannot be explained by input differences. Measurement error complicates the decomposition and the problem is further exacerbated by the endogeneity of productivity in the firm's input choice decisions.

I compare the robustness of five widely used techniques: (a) index numbers, (b) data envelopment analysis, and three parametric methods, (c) instrumental variables estimation, (d) stochastic frontiers, and (e) semi-parametric estimation. First, the sensitivity of the different methods to measurement error are evaluated using a Monte Carlo simulation. I find the parametric methods to be least affected. Second, using a panel of manufacturing plants in Zimbabwe, I show that the different methods generate surprisingly similar results. All methods confirm that exporters and large firms are more productive on average and that firms that invest in new technology improve productivity faster, facts that have previously been established for U.S. plants.

*Comments welcome, please send to jovb@chass.utoronto.ca. I would like to thank Marcel Fafchamps for his help getting me access to the data and Mel Fuss and Frank Wolak for comments. All remaining errors are my own. Financial support from the Connaught Fund is gratefully acknowledged.

1 Motivation

Productivity is a concept that is used and discussed widely. Ever since Solow (1957) decomposed output growth into the contribution of inputs and a residual productivity term, the concept has increased in popularity. Productivity has generated a lot of interest in its own right and is used as a benchmark to rank firms or countries. Such rankings gained credibility once other studies documented that productivity is correlated with other measures of success such as employment growth, export status, or technology adoption. Low productivity has also been found to predict exit, the ultimate performance standard. The importance can also be gauged from the attention it receives as a criterion to evaluate a policy intervention or firms' decisions. In industrial economics, for example, a large literature investigates the effect of R&D on productivity and the resulting impact on industry structure. Efforts to evaluate the impact of trade liberalization has recently turned from estimating changes in price-cost margins to productivity changes.

Accurate measurement is at the heart of productivity. Fundamentally, the objective of productivity measurement is to identify output differences that are cannot be explained by input differences. In order to perform this exercise, one needs to control for input substitution. Differences between firms' input and output choices result from differences in technology —which we will label productivity differences— or differences in factor prices which lead firms to pick different points on the production frontier.[1] The possibility to substitute one input for another is captured by the production function —or any other representation of the production technology— and is naturally not observable. Methodologies to estimate productivity differ by the mix of statistical techniques and economic assumptions they employ to control for input substitution.

Another issue that has received a fair bit of attention is accurate measurement of output and inputs. Mismeasurement can result, among other things, from unobserved quality or price differences, aggregation problems, recall errors in surveys, or incompatibilities in refer-

[1] One might argue that some output shortfall is the result of inefficiency at the firm-level. To be consistent with a profit maximizing model of the firm, I classify such shortfalls as productivity differences as they might be caused by technology differences, unmeasured inputs, or quality differences in outputs, among other possibilities. See Stigler (1976) for a more elaborate motivation.

ence period for output and inputs. The effect on productivity measures will obviously depend on the estimation method. I evaluate five popular methodologies, which fall in three broad classes. The first two, index numbers and data envelopment analysis, are very flexible in the specification of technology, but they do not allow for data errors, making the effect of measurement error completely unpredictable. The other, parametric methods calculate productivity from an estimated production function. In the simplest linear regression model measurement error in the dependent variable leaves least squares estimation unbiased and efficient, while errors in the independent variables biases coefficients downwards. For the production function the effect is not unambiguous, because more complicated estimation methods have been devised to deal with the simultaneity of productivity and input choices. Moreover, the principal interest is in the residual of the production function or in the part of the residual interpreted as productivity. The extent to which the resudual is affected by measurement error is impossible to predict.

First, I evaluate the different methodologies using a Monte Carlo study, where the focus is on the accuracy of productivity estimates when variables are measured with errors. The data for the simulations is constructed, incorporating the various complications the different methodologies are supposed to resolve. Both productivity levels and growth rates are evaluated, as they are often calculated differently. The results are verified using an actual sample of manufacturing plants from Zimbabwe. The different methods yield surprisingly similar productivity estimates and they consistently find support for a number of stylized facts, that have previously been established for U.S. plants.

In the next section, I give some background on productivity measurement and introduce the issues involved. Subsequently, I introduce the different methodologies. An attempt is made to present the general idea and convey the distinctive features of each methodology. Links to the literature for more detailed information are provided in the respective sections. Section 4 contains the Monte Carlo study and Section 5 the empirical example. At the end, I summarize the lessons to take away from these exercises.

2 Issues in productivity measurement

In plain English, one firm is more productive than another if it is able to produce the same amount of outputs with less inputs or if it produces more outputs from the same amount of inputs. Similarly, a firm has experienced positive productivity growth if output has increased more than inputs or inputs decreased more than outputs. The comparison becomes more difficult, of course, if one firm uses more of the first input, while the other firm relies more on a second input.

Formally, two production plans are compared, k and l, which can refer to two different firms or to the production plans of a single firm at two different points in time. The comparison is between two transformation functions evaluated at two different points,

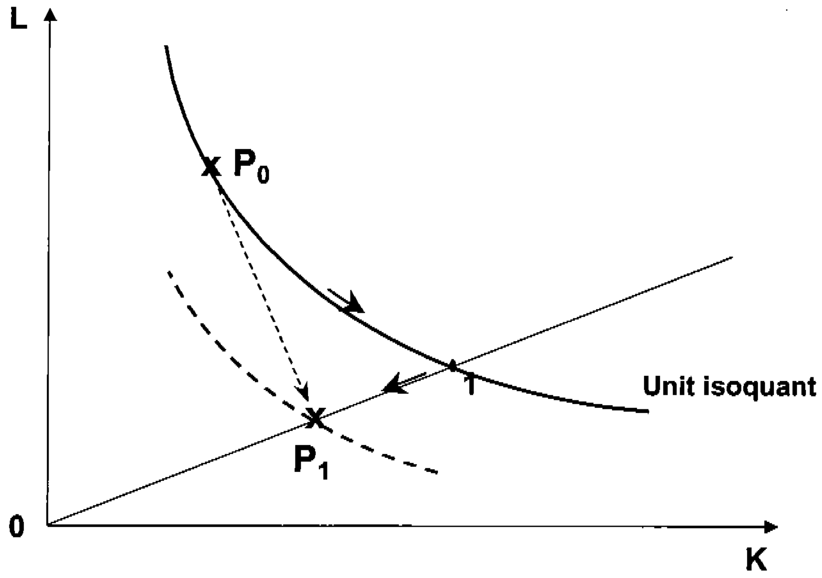
$$\begin{aligned} F^k(q^k, x^k) &= 0 \qquad (1) \\ F^l(q^l, x^l) &= 0, \end{aligned}$$

where q and x are vectors of outputs and inputs. The superscripts on the transformation functions $F(.)$ indicate that the two firms potentially produce with different technologies. Two different productivity comparisons are possible. The first one is output-based and comparing two firms it provides an answer to the question: "How much extra output does a firm produce, relative to another firm, conditional on its (extra) input use?" The second measure is input-based and asks "What is the minimum input requirement for one firm to produce the same output as another firm?" Because the technology, represented by the transformation curves in (1), can differ by firm, each productivity comparison can be carried out using the technology of either firm, yielding a total of four different measures.

It should be stressed that taking more than one input into account makes it impossible to talk about productivity without specifying a transformation function (or production function or any other way to represent the production technology). Since firms' input substitution possibilities are determined by the technology they employ, each productivity measure is only defined with respect to that specific production technology. Measuring productivity necessarily involves decomposing differences in the input-output composition into shifts along a production function and shifts of the function itself.

Figure 2 illustrates the objective of productivity comparisons. It compares two production plans, P_0 and P_1, in input space. Part of the difference, from P_0 to 1, is a shift along the frontier —represented here by the unit isoquant— exploiting the input substitution possibilities of the technology. Part of the difference, from 1 to P_1, is an actual shift of the frontier, which is counted as technical change or productivity growth. If technical change is limited to be Hicks-neutral, the shape of the unit isoquant will be unchanged after the shift. Capital-biased or labor-saving technical change, on the other hand, would entail a larger shift down than left. It leads to a higher capital-labor ratio if factor prices are unchanged. Figure 2 also illustrates that productivity differences can only be measured relative to a specific technology.

Figure 1: Decomposing shifts along the frontier from a shift of the frontier



Most applications limit themselves to a single output, and I'll do likewise. Some methods can easily deal with multiple outputs while it is impossible with others. In practice, most applications rely on the aggregation of different products using prices within the firm and use deflated sales or value added as a single output aggregate. For the production function

$$Q_{it} \quad = \quad A_{it}\, F_{it}(X_{it}), \tag{2}$$

all productivity differences are concentrated in the multiplicative factor A_{it}, which differs between firms and changes over time.[2] The productivity of firm i at time t can be compared to the productivity of the same firm in the previous period, to find productivity growth, in (3). Another comparison is with the productivity level of another firm. In (4), the benchmark is the average productivity level for the industry, $\overline{A_t}$, which facilitates multilateral comparisons.[3]

$$\begin{aligned}
\log \frac{A_{it}}{A_{it-1}} &= \log \frac{Q_{it}}{Q_{it-1}} - \log \frac{F_{it}(X_{it})}{F_{it-1}(X_{it-1})} \quad \text{productivity growth} && (3)\\
\log \frac{A_{it}}{\overline{A_t}} &= \log \frac{Q_{it}}{\overline{Q_t}} - \log \frac{F_{it}(X_{it})}{\overline{F(X_t)}} \quad \text{productivity level.} && (4)
\end{aligned}$$

Both ratios illustrate that productivity is intrinsically a relative concept. The calculation of the last term in (3) and (4) —the ratio of input aggregators— is the primary difference between methods. Three broad classes of methodologies are compared.

Data envelopment analysis (DEA) is a nonparametric approach and is agnostic on the shape of the production function. Conceptually, it lays a piece-wise linear frontier on top of the observed production plans in input-output space. Firms that are dominated by a competitor or a linear combination of competitors are deemed less than 100% efficient. Domination occurs when other firms produce more output using the same amount of inputs (for an output-based measure) or when other firms use less inputs to produce the same output (input-based measure). Aggregation of inputs and outputs is linear, while the weights are allowed to vary for each unit under investigation.

The second approach, index numbers, provides a theoretically motivated aggregation method for inputs and outputs and is also fairly agnostic on the shape of the underlying technology. Research on exact index number, starting with Diewert (1976), provide a basis

[2] Such productivity differences are called Hicks-neutral, because they affect all inputs identically.

[3] In practice, most studies have used $\log A_{it} - \overline{\log A_t}$ as multilateral productivity comparison and for comparability I follow this practice.

for discrete comparisons of production plans that are exact for a family of production functions. Under a number of assumptions, e.g. profit maximization, it is possible to calculate the last terms in (3) and (4) from observables, without having to specify the exact production function. With some limitations, heterogeneity in production technology between firms is still allowed.

The third approach, parametric estimation, assumes that the input tradeoff and returns to scale are the same for all units under study. Functional form assumptions often yield more precise estimates at the expense of concentrating all firm heterogeneity in the productivity term.[4] Parametric estimation poses another problem, first raised by Marschak and Andrews (1944). If plants make input decisions based on their own level of productivity, which is unobservable to the econometrician, it creates an endogeneity problem. Three different techniques to solve this problem are compared. The most straightforward solution is to use instrumental variables. I rely on the method proposed by Blundell and Bond (1998) to generate moment conditions. The use of distributional assumptions on unobserved productivity is the solution adopted in the stochastic frontier literature. A more recent paper, by Olley and Pakes (1996), introduces a nonparametric solution to the problem. It controls for unobserved productivity differences by introducing an observable variable —investment— that has a systematic relation with it.

I evaluate the different methodologies when an additional problem is present, measurement errors in inputs. It is apparent from the definition of productivity —the difference in output after controlling for input differences— that this problem is potentially very serious. The construction of comparable input and output aggregates invariably involves several judgement calls and it is likely that measurement errors are introduced. On the measurement front, there are a lot of issues that researchers have dealt with that I will abstract from.[5]

[4] While it is possible to estimate production functions with random coefficients, allowing for some technology differences between firms, this approach has not been fruitful, see Mairesse and Griliches (1990) for a discussion.

[5] If competition is less than perfect, the use of deflated sales as output measure is problematic. Klette and Griliches (1996) provide a solution if one is willing to make assumptions on the type of competition and functional form of demand. In the first study using a full census of manufacturing plants, Griliches and Ringstad (1971) discuss the relative merits of a value added and gross output production function. The problems associated with the aggregation of inputs and outputs, as well as quality differences in heterogeneous inputs are the subject of an exchange between some of the pioneers of productivity decompositions, see Jorgenson and Griliches (1967, 1972) and Denison (1972). Methods have also been developed to deal with variations in capacity utilization (Berndt and Fuss 1986) and regulated firms (Denny, Fuss, and Waverman 1981). Finally, the choice of functional form for the production function has also received careful analysis

3 Methodologies to measure productivity

3.1 Index numbers

When Solow (1957) performed the first growth accounting exercise, he defined productivity as the time derivative of the production function and approximated it with year-to-year changes. For a production function with two inputs, constant returns to scale, and only Hicks-neutral technical change differentiating and reordering the production function (2) gives total factor productivity (TFP) growth:

$$\frac{\partial A/\partial t}{A} = \frac{\partial Q/\partial t}{Q} - \left(\frac{A\,\frac{\partial F}{\partial K}\,K}{Q}\right)\frac{\partial K/\partial t}{K} - \left(\frac{A\,\frac{\partial F}{\partial L}\,L}{Q}\right)\frac{\partial L/\partial t}{L}.$$

Using the first order conditions for optimal capital and labor input ($A\frac{\partial F}{\partial K} = r$ and $A\frac{\partial F}{\partial L} = w$) and approximating the time derivatives discretely gives

$$\frac{\Delta A}{A} = \frac{\Delta Q}{Q} - s_k\frac{\Delta K}{K} - (1 - s_k)\frac{\Delta L}{L} \tag{5}$$

or as Solow (1957) calculated it

$$\frac{\Delta A}{A} = \frac{\Delta(Q/L)}{Q/L} - s_k\frac{\Delta(K/L)}{K/L},$$

where s_k is the value share of capital in output.

A first generalization was provided by Diewert (1976). He showed that under certain conditions the ratio of two unknown functions evaluated at different points can be calculated exactly with an index number without knowledge of the parameters. In particular, if the production function in (2) is translog, the second term in productivity growth (3) can be calculated with a Törnqvist index number such that

$$\log\frac{A_{it}}{A_{it-1}} = \log\frac{Q_{it}}{Q_{it-1}} - (\tfrac{s_{it}+s_{it-1}}{2})\log\frac{L_{it}}{L_{it-1}} - (1 - \tfrac{s_{it}+s_{it-1}}{2})\log\frac{K_{it}}{K_{it-1}}. \tag{6}$$

(Berndt and Khaled 1979).

s_{it} is the fraction of the wage bill in output or total cost, which should be equal given the constant returns assumption. Because the comparison is between two discrete time periods, the formula is equally applicable to bilateral productivity level comparisons between firms. If there are multiple outputs, the single output ratio is simply replaced by the weighted sum of each output ratio, using revenue shares as weights, similarly to the cost shares for inputs.

The drawback is that (6) is only exact under a host of assumptions. Some are rather innocuous and one would make them anyway, such as the assumption that input and output markets are competitive or that firms are minimizing costs. Other assumptions are more restrictive. In particular, one has to assume that production is according to a constant returns to scale technology and that there is no measurement error, in addition to the translog production function assumption. It might be a consolation that if some conditions do not hold the index number is not exact, but still a valid second-order approximation to the productivity ratio. The Törnqvist index is just one possibility and different technologies require a different index number. It is the most popular one because it conveniently rationalizes Solow's original TFP formula.

Subsequently, Caves, Christensen, and Diewert (1982a) extended (6) further, allowing for technical change that is not Hicks-neutral and variable returns to scale, and giving it a much more general interpretation. A general method to compare two production plans is a Malmquist firm l input-based productivity index:[6]

$$M^l \equiv \frac{D^l(q^k, x^k)}{D^l(q^l, x^l)} = \max_{\delta} \ \{\delta : f^l(q^k_{-1}, \frac{x^k}{\delta}) \geq q^k_1\}. \tag{7}$$

It is the ratio of two input distance functions evaluated at the different production plans. The index is firm l and input-based, because it measures how much to deflate firm k's inputs for its production plan to lay on the transformation frontier of firm l. A firm k based index would use the technology embodied in f^k. An output-based productivity index would make the comparison by inflating or deflating output, keeping inputs constant. Under constant returns

[6] The transformation function is written as $f(q_{-1}, x) = q_1$, but is otherwise similar to the specification in (1). $D()$ is the input distance function, which is yet another way to represent a technology. It measures the amount of input deflation (or inflation) needed for a production plan to lay on the transformation function. By definition, $D^l(q^l, x^l) = 1$.

to scale, the input and output based indices are identical.

In order to do the actual comparison, knowledge of the functional form of the input distance function D^l is required. Alternatively, Caves, Christensen, and Diewert (1982a) prove that under the same assumptions as before (and some additional ones) the geometric mean of firm l and firm k output-based indices, $m^o(x^l, x^k, y^l, y^k)$, *exactly* equals the difference between a Törnqvist output index and the corresponding input index with a scale factor to account for non-constant returns to scale:

$$\begin{aligned} \log m^o(x^l, x^k, q^l, q^k) &= \sum_i^I \tfrac{r_i^l + r_i^k}{2}(\log q_i^l - \log q_i^k) - \sum_n^N \tfrac{s_n^l + s_n^k}{2}(\log x_n^l - \log x_n^k) \\ &\quad + \sum_n^N \tfrac{s_n^l(1-\epsilon^l) + s_n^k(1-\epsilon^k)}{2}(\log x_n^l - \log x_n^k). \end{aligned} \qquad (8)$$

r_i^s is the revenue share of output i and firm s $(i = 1...I, s = k, l)$, s_n^s is the cost share of input n and firm s $(n = 1...N)$, and ϵ^s are the (local) returns to scale for firm s. The corresponding input-based productivity index, $m^i(x^l, x^k, y^l, y^k)$, differs only in the scale factor:

$$\begin{aligned} \log m^i(x^l, x^k, q^l, q^k) &= \sum_i^I \tfrac{r_i^l + r_i^k}{2}(\log q_i^l - \log q_i^k) - \sum_n^N \tfrac{s_n^l + s_n^k}{2}(\log x_n^l - \log x_n^k) \\ &\quad + \sum_i^I \tfrac{r_i^l(1/\epsilon^l - 1) + r_i^k(1/\epsilon^k - 1)}{2}(\log q_i^l - \log q_i^k). \end{aligned} \qquad (9)$$

Note that under constant returns to scale the last term disappears and both indices are equal. Very few applications include the scale factor, adjusting for variable returns to scale, in the calculations. Instead, only the first two terms on the right hand side of (8) and (9) are included, which amounts to lumping the effect of scale economies in the productivity measure. For comparability with the other methodologies, I do include the scale factor.

The formulas hold if firms maximize profits, are price takers on input markets, and if the underlying technology can be characterized by translog output or input distance functions, which are otherwise left unspecified.[7] The only restriction on technology is that the second order terms are the same for the two units compared. The coefficients on all first order terms

[7] In the single output case, only cost minimization is needed.

are allowed to differ, technical change can be non-neutral, and returns to scale can vary, although one needs to know them to implement equations (8) and (9).

Equations (8) and (9) can be used for productivity growth calculations by replacing the l and k superscripts by t and $t-1$. For comparisons between firms, multilateral comparisons are generally preferred over the bilateral ones, because Törnqvist indices are not transitive., Caves, Christensen, and Diewert (1982b) propose an alternative multilateral comparison. Each firm is compared with a hypothetical firm, with output vector $\overline{y}$, input vector $\overline{x}$, revenue share $\overline{r}$, and cost shares $\overline{s}$. This allows for multilateral comparisons, yields bilateral comparisons that are transitive, and still allows for technology that is firm-specific.

The main advantages of the index number approach to productivity measurement is that they are calculated easily, taking by far the least computer and programming time. The specification of technology is flexible, allowing firms to a large extent to produce with different technologies. The method can easily handle multiple outputs and a large number of inputs, without a need for restrictions on the structure of production or separability assumptions. The main disadvantages are the requirements on data quality and the assumptions on firm behavior and market structure. It is impossible to account for measurement errors or to deal with outliers, except for some ad hoc trimming of the data. More complicated extensions exist for regulated firms, non-competitive output markets, and temporary equilibrium but they involve estimating some structural parameters or are more data intensive. Even the calculations under variable returns to scale require data on the local returns to scale for each firm and on the price of capital, which is not easily obtained.[8]

3.2 Data envelopment analysis

A second approach to productivity measurement relies on nonparametric estimation techniques, using linear programming. The basic method dates back to Farrell (1957) and it was operationalized by Charnes, Cooper, and Rhodes (1978).[9] It is most popular for cross-section

[8] To implement the Törnqvist index number in the variable returns to scale case with two inputs, I estimate returns to scale using least squares and calculate the capital share to be consistent with the observed labor share and the estimated scale economies, such that $s_k + s_l = \frac{1}{\epsilon}$.

[9] More information on the method and applications can be found in Seiford and Thrall (1990).

comparisons. The intuition is to lay a piece-wise linear production frontier in input-output space over the most efficient observations. Observations that are not dominated are labeled 100% efficient. Domination occurs when another firm, or a linear combination of other firms, uses less inputs to produce the same outputs (for an input-based measure) or produces more output using the same inputs (for an output-based measure). If there are multiple inputs or outputs, they are aggregated linearly. Weights are chosen optimally for the unit under consideration, with the restriction that the efficiency of all observations can be no more than 100% when the same weights are applied to them.

A linear programming problem is solved separately for each observation. Input and output weights are chosen to maximize efficiency. The number of restrictions equals the number of observations, plus sign restrictions on the weights. For unit 1 the problem amounts to

$$\begin{aligned} \max_{v_j,\,u_k} \quad & \theta_1 = \frac{\sum_{j=1}^{J} v_j q_{1j}}{\sum_{k=1}^{K} u_k x_{1k}} \\ \text{subject to} \quad & \frac{\sum_{j=1}^{J} v_j q_{ij}}{\sum_{k=1}^{K} u_k x_{ik}} \leq 1 \qquad && i = 1 \ldots N \\ & v_j,\ u_k \geq 0 && j = 1 \ldots J,\ k = 1 \ldots K, \end{aligned}$$

i indexes firms, j outputs, and k inputs. The general problem is converted to a linear programming problem by multiplying the objective and restrictions by their denominator. Since the scale of the weights is not defined —multiplying all weights by the same multiplier does not change the problem— a normalization is necessary. Usually the linear combination of inputs is set to unity for the unit under investigation. In the two-input, single output case, the problem for unit 1 becomes

$$\begin{aligned} &\max_{v_q, u_l, u_k} \quad \theta_1 = v_q Q_{1t} \\ \\ &\text{subject to} \quad u_l L_1 + {}_{k}u \; K_1 = 1 \\ &\qquad\qquad\quad v_q Q_i - u_l L_i - u_k K_i \leq 0 \qquad i = 1...N \\ &\qquad\qquad\quad v_q, u_l, u_k \geq 0. \end{aligned} \tag{10}$$

Such problem has to be solved for each observation (firm-year). Often the dual problem is solved, where θ_1 is chosen directly:

$$\begin{aligned} &\min_{\theta_1, \lambda_i} \quad \theta_1 \\ \\ &\text{subject to} \quad \textstyle\sum_{i=1}^{N} \lambda_i Q_i - Q_1 \geq 0 \qquad i = 1...N \\ &\qquad\qquad\quad \theta_1 L_1 - \textstyle\sum_{i=1}^{N} \lambda_i L_i \geq 0 \qquad i = 1...N \\ &\qquad\qquad\quad \theta_1 K_1 - \textstyle\sum_{i=1}^{N} \lambda_i K_i \geq 0 \qquad i = 1...N \\ &\qquad\qquad\quad \theta_1 \text{ free}, \;\; \lambda_i \geq 0. \qquad i = 1...N \end{aligned} \tag{11}$$

The optimized θ_1 provides an input-based efficiency measure for firm 1. It captures the percentage reduction in inputs necessary for unit 1 for its production plan to lie on the production frontier of the most efficient units in the sample.[10] The problem is similar to the Malmquist index in equation (7), but instead of assuming a translog input distance function, inputs are aggregated linearly.

Interchanging the roles of inputs and output in (10) and minimizing the objective function, gives the corresponding output-oriented programming problem. Efficiency is calculated as the inverse of the optimized objective value. A dual problem can be written as well, see Seiford and Thrall (1990) for more details. In the constant returns to scale case, the input- and output-oriented variants both yield the same efficiency estimates, as was the case for the index numbers calculations.

[10] Rather than interpreting the extra input use as inefficiency, I will say that the production function of firm 1 is shifted downwards, relative to best practice.

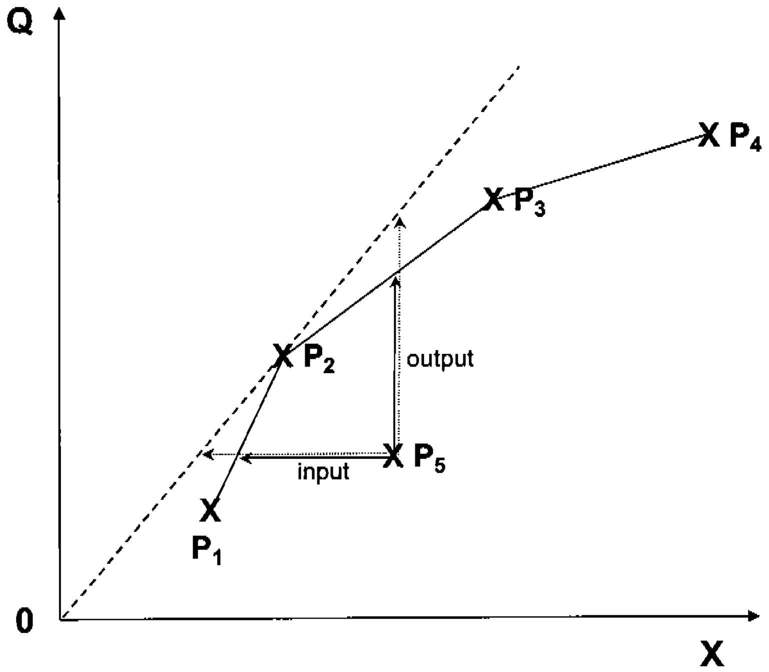
Both formulations of the problem, (10) and (11), implicitly incorporate a constant returns to scale assumption. To relax this, an extra constraint has to be added to (11):[11]

$$\sum_{i=1}^{N} \lambda_i = 1.$$

The restriction rules out the extrapolation of a firm beyond its size ($\lambda_i > 1$) or scaling down observed production plans ($\sum_i \lambda_i < 1$) to make the comparison.

Figure 2 provides some intuition for the DEA methodology. It is drawn for a single input and output, but the intuition is similar for higher dimensional problems as the inputs and outputs are always aggregated linearly.[12]

Figure 2: Nonparametric production frontiers



P_1 to P_5 represent the production plans of different firms that are compared. The solid line represents the frontier if variable returns are allowed. It is obtained by laying a piece-wise

[11] Alternatively, an extra variable can be introduced in the dual formulation (10).

[12] Because the weights to construct the input and output aggregate are chosen optimally for the observation under consideration, a different graph has to be drawn for each observation if there are multiple inputs or outputs.

linear curve over the extreme points. Four of the five observations lie on the frontier and are deemed 100% efficient. If the technology is restricted to constant returns to scale, the frontier is allowed to go through the origin and extrapolated beyond the observed data points, resulting in the dashed line as production frontier. Only the second firm is fully efficient in this case. Imposing constant returns to scale adds an extra constraint to the problem, restricting the weights further. As a result, the maximized objective values —the estimated efficiencies— will be (weakly) lower.

The distance from the production plan of each unit to the frontier represents the estimated efficiency. In an input orientation, one improves efficiency by reducing inputs: a horizontal projecting onto the frontier. In an output orientation, efficiency is increased by increasing output until the unit produces on the frontier, given its observed inputs: a vertical projection. Figure 2 makes clear that under variable returns both orientations yield different results, as the frontier does not go through the origin and the slope of the segments the unit gets projected onto might differ.

The efficiency measures can be interpreted as the productivity difference between unit i and the most productive unit,

$$\theta_i = \frac{TFP_i}{TFP_{\max}}.$$

For a measure comparable to the results obtained using other methodologies, I define

$$\log TFP_i^{DEA} - \overline{\log TFP}^{DEA} = \log\theta_i - \frac{1}{N}\sum_{i=1}^{N}\log\theta_i,$$

as the productivity level. Productivity growth fits less straightforwardly into the DEA framework. Including all the different firm-years as separate observations in the analysis, it is possible to calculate productivity growth as

$$TFPG_{it}^{DEA} = \log TFP_{it}^{DEA} - \log TFP_{it-1}^{DEA} = \log\theta_{it} - \log\theta_{it-1}.$$

While these transformations are arbitrary, they do not change the ranking of firms, only the absolute productivity levels and growth rates.

DEA has the advantage to deal with many outputs in a consistent way. It leaves the underlying technology unspecified and allows for heterogeneity, by allowing different input and output weights for different observations. There is also no need for functional form or behavioral assumptions. While there is no theoretical justification for the linear aggregation, it does come natural if one considers an activities analysis framework. Each firm is considered a separate process that can be combined with others to replicate the production plan of the unit under investigation. The flexibility in weighting is a concern. It has the implication that each firm with the highest output-input ratio for any combination of outputs and inputs will be considered efficient, as it can put maximum weight on these factors. Under variable returns to scale, each firm with the lowest input or highest output level in absolute terms is also fully efficient. The method is not stochastic, which is demanding on the data, and makes the method sensitive to outliers.[13] One might object to the label "100% efficient" for the best practice firms in the sample. In some situations no firm might be efficient, e.g. due to regulation.

3.3 Parametric methods

If one is prepared to make functional form assumptions on the production technology, econometric estimation of the production or cost function provides a flexible alternative that requires few other assumptions. While specification error of the functional form is a potential problem, at least as important is the embedded assumption that all firms operate with the same technology.

I follow most of the literature by using a simple Cobb-Douglas production function,

$$Q_{it} = A_{it} L_{it}^{\alpha_l} K_{it}^{\alpha_k} e^{\epsilon_{it}},$$

or in logarithms

$$q_{it} = \alpha_0 + \alpha_l l_{it} + \alpha_k k_{it} + \omega_{it} + \epsilon_{it}. \tag{12}$$

[13] More recently, stochastic DEA methods have been developed, but they are not universally accepted yet and most application still apply the deterministic variants.

With this functional form, only Hicks-neutral technical change can be identified.

Productivity comparisons are straightforward:

$$\log \frac{A_{it}}{A_{j\tau}} = \omega_{it} - \omega_{j\tau} = \log \frac{Q_{it}}{Q_{j\tau}} - \alpha_l \text{lo g} \frac{L_{it}}{L_{j\tau}} - \alpha_k \log \frac{K_{it}}{K_{j\tau}} + (\epsilon_{it} - \epsilon_{j\tau}). \tag{13}$$

Depending on one's taste one can look at $\log(\frac{A_{it}}{A_{j\tau}})$ as in Baily, Hulten, and Campbell (1992), $\frac{A_{it}}{A_{j\tau}}$ as in Olley and Pakes (1996), or at $\frac{A_{it}-A_{j\tau}}{A_{j\tau}}$ as in Solow (1957). To calculate productivity, two unobservable components have to be separated: the random noise and productivity differences. Alternatively, the last term in (13) is dropped because $E(\epsilon_{it} - \epsilon_{j\tau}) = 0$, in which case the difference in random noise $(\hat{\epsilon}_{it} - \hat{\epsilon}_{j\tau})$ ends up in the productivity term.

Consistent estimation of the input parameters faces an endogeneity problem, first discussed by Marschak and Andrews (1944). Firms choose inputs, knowing their own level of productivity, which is unobservable to the econometrician. A regression of output on inputs will give biased estimates of the production function coefficients. The most straightforward solution is to use instrumental variables that are uncorrelated with productivity. Because instruments are hard to come by in this context, other methods were developed. The stochastic frontier literature makes explicit distributional assumptions about the unobserved productivity factor and estimates the primitives of the distribution. Olley and Pakes (1996) obtain an expression for unobserved productivity by inverting the investment function nonparametrically and substitute the expression in the production function. I will discuss each of the three approaches in turn.

3.3.1 Instrumental variables estimation

Using instrumental variables is the most straightforward solution to an endogeneity problem. In the context of production functions it is, unfortunately, hard to obtain valid or strong instruments. Most methods dictate estimating the production function in first difference form, but the results have generally been unsatisfactory, see for example Mairesse and Griliches (1998). The coefficient on capital is estimated much lower than in the level equation and returns to scale are generally estimated implausibly low. This is what one might expect if the inputs and output are persistent over time and instruments are weak.

A general approach to estimate error component models was developed in Blundell and Bond (1998) and applied to production functions in Blundell and Bond (2000). They propose a new set of moment conditions with a more solid theoretical underpinning and obtain more plausible results. The production function they estimate takes the form

$$\begin{aligned} q_{it} &= \alpha_t + \alpha_l l_{it} + \alpha_k k_{it} + (\omega_i + \omega_{it} + \epsilon_{it}) \\ \omega_{it} &= \rho\omega_{it-1} + \eta_{it} \qquad\qquad |\rho| < 1 \\ &\epsilon_{it},\ \eta_{it} \sim i.i.d. \end{aligned}$$

The three error components in the production function are a firm specific fixed-effect ω_i, an autoregressive component ω_{it}, with η_{it} an idiosyncratic productivity shock, and ϵ_{it} is measurement error. The equation includes year specific intercepts. The goal is to consistently estimate the structural parameters of the model, α_l, α_k, α_t, and ρ, when the number of time periods is fixed.

In its dynamic representation the model becomes

$$\begin{aligned} q_{it} &= \alpha_l l_{it} + \rho\alpha_l l_{it-1} + \alpha_k k_{it} + \rho\alpha_k k_{it-1} + \rho q_{it-1} \\ &\quad + \underbrace{(\alpha_t - \rho\alpha_{t-1})}_{\alpha_t^*} + \underbrace{\omega\,(1-\rho)}_{\omega_i^*} + \underbrace{(\eta_{it} + \epsilon_{it} - \rho\epsilon_{it-1})}_{\varepsilon_{it}}. \end{aligned} \tag{14}$$

All variables on the first line are observable; firm and year dummies will take care of the first two terms on the second line. There is still a need for moment conditions to provide instruments because the inputs and lagged output will be correlated with the composite error ε_{it}.

Standard assumptions on the initial conditions,

$$\begin{aligned} E[l_{i1}\eta_{it}] = E[k_{i1}\eta_{it}] = E[q_{i1}\eta_{it}] = 0 \qquad\qquad t = 2, ..., T \\ E[l_{i1}\epsilon_{it}] = E[k_{i1}\epsilon_{it}] = E[q_{i1}\epsilon_{it}] = 0, \qquad\qquad t = 2, ..., T \end{aligned}$$

yield three times $T - s$ moment conditions

$$\begin{aligned} E[l_{it-s}\Delta\varepsilon_{it}] &= 0 \\ E[k_{it-s}\Delta\varepsilon_{it}] &= 0 \\ E[q_{it-s}\Delta\varepsilon_{it}] &= 0, \end{aligned} \tag{15}$$

for $s \geq 2$ if there is no measurement error ($\sigma_\epsilon = 0$) and $s \geq 3$ otherwise. These moment conditions allow the estimation of (14) in first-differenced form using twice or three times lagged inputs and output as instruments. Blundell and Bond (2000) illustrate theoretically and with a practical application that these instruments can be weak.

If one is willing to make the additional assumptions that

$$E[\Delta l_{it}\omega_i^*] = E[\Delta k_{it}\omega_i^*] = 0 \qquad\qquad t = 2, ..., T$$

and that the initial conditions satisfy

$$E[\Delta q_{i2}\omega_i^*] = 0,$$

one can derive two additional moment conditions

$$\begin{aligned} E[\Delta l_{it-s'}(\omega_i^* + \varepsilon_{it})] &= 0 \\ E[\Delta k_{it-s'}(\omega_i^* + \varepsilon_{it})] &= 0, \end{aligned} \tag{16}$$

for $s' = 1$ if there is no measurement error and $s' = 2$ otherwise. Once or twice lagged first differences of inputs are valid instruments for the production function (14) in levels. Further lagged differences can be shown to be redundant once the moment conditions in (16) have been exploited. Blundell and Bond (2000) show that joint stationarity of the inputs and output, conditional on common year dummies, is sufficient, but not necessary for (16) to hold. Both equations, in first differences and levels, can be estimated as a system, imposing the coefficient restrictions embodied in (14).

Advantages of this method are the flexibility in generating instruments and the possibility of testing for overidentification. It allows for an autoregressive component to productivity, in addition to a fixed-effect and an idiosyncratic component. The major disadvantage is the need for a longer panel. The minimum number of time periods necessary to estimate the model is three, in which case there is no room for measurement error in the dependent variable. The number of overidentifying moment restrictions is equal to the number of independent variables if cross-equation restrictions are enforced. With four years of data, an i.i.d. error component can be incorporated. At least five years of data are needed to generate additional, overidentifying moment conditions.

3.3.2 Stochastic frontier estimation

The stochastic frontier literature adds assumptions on the distribution function of the unobserved productivity component to the specification of the production function. The productivity term is modeled as a negative term, drawn from a known distribution. The method is credited to Aigner, Lovell, and Schmidt (1977) and Meeusen and van den Broeck (1977) who used the negative of an exponential and half-normal distribution for unobserved productivity. Stevenson (1980) introduced a truncated normal distribution that is more flexible on the location of the mode of the distribution. Estimation is usually with maximum likelihood.

In production function (12), the term ω_{it} is now weakly negative and it represents the inefficiency of firm i at time t. The production plan of firm i is said to lie below the best practice production frontier. Alternatively, one can say that firm i produces according to a production function which is shifted down by ω_{it} with respect to best practice. The shift is zero for the most efficient firm, producing at the frontier.

Most distributional assumptions on the unobserved productivity component result in firms being bunched close to the frontier, with a small number of outliers being significantly less efficient. This is consistent with the view that a best practice technology exists and all firms strive to attain it. High productivity outliers are not observed (truncated at zero), because they advance the technological frontier. A market competition view on productivity, one the other hand, would argue that there is a threshold for productivity, below which a firm is forced to exit the industry. Accordingly, it is more likely to find a mass of firms bunched above the threshold, producing at the low end of the productivity distribution, struggling for survival, while there are some outliers with very high productivity. Low productivity outliers are not observed, because it forces a firm into bankruptcy.

The original stochastic frontier models were developed to assess productivity in a cross section of firms.[14] The model was subsequently generalized for panel data in a number of different ways. Battese and Coelli (1992) provide the most straightforward generalization, modeling the inefficiency terms as

$$\omega_{it} \quad = \quad -e^{-\eta(t-T)}\,\omega_i. \tag{17}$$

A firm fixed-effect, ω_i is drawn from a truncated normal distribution and is multiplied by a factor that increases (if η is positive) or decreases (if η is negative) over time. The ranking of firms is unchanged over time and the inefficiency evolves identically and deterministically for all firms.

A less restrictive model, introduced by Huang and Liu (1994), specifies

$$\omega_{it} \quad = \quad -(Z_{it}\delta + Z^*_{it}\delta^* + \nu_{it}), \tag{18}$$

where ν_{it} is drawn from a normal distribution, such that ω_{it} is negative. The variables in Z are exogenous determinants of efficiency and those in Z^* are appropriate interactions between input variables and variables in Z. The model is called non-neutral because inefficiency is allowed to vary by input use. Because the truncation depends on variables that vary by firm, the inefficiency terms are still independently distributed, but not identically.

If one observes firms only once, making strong assumptions is the only possibility to separate the productivity component ω_{it} from the random error component ϵ_{it}. Panel data contains more information on each firm and allows identification under weaker assumptions. Schmidt and Sickles (1984) propose to use the standard fixed-effects panel data estimator to estimate a constant firm-level productivity term. Firm dummies give a direct estimate of ω_i.

[14] The same holds for DEA, which is also called deterministic frontier analysis.

The problematic correlation between inputs and unobserved productivity has been ruled out by assumption. Cornwell, Schmidt, and Sickles (1990) generalize the method by estimating a time-varying component that is firm specific. They adopt a quadratic specification for time and estimate three coefficients per firm. Because my sample is relatively short, I only include a linear time trend:

$$\omega_{it} = \alpha_{i0} + \alpha_{i1}\text{time}. \tag{19}$$

Firm-level productivity increases or decreases deterministically over time at a constant rate, $\alpha_i 1$, which differs between firms.

I implement the most and least restrictive formulations for panel data, the models by Battese and Coelli (1992) and Cornwell, Schmidt, and Sickles (1990). An advantage of stochastic frontiers is their relative simplicity to implement, especially the second formulation. In addition, the parametric specification of the production process can be generalized easily to estimate more sophisticated specifications, e.g. incorporating biased technological change. With a short panel, the results are for the second estimator are not very robust and the estimation used a lot of degrees of freedom. Outliers do not affect the parameter estimates for the coefficients in the production function very much, but productivity estimates for individual firms vary widely.[15] One might also be uncomfortable with the identification coming solely from functional form assumptions, which are especially restrictive for the first estimator.

3.3.3 Semi-parametric estimation

The final method was developed by Olley and Pakes (1996) to estimate the productivity effects of restructuring in the U.S. telecommunications equipment industry. They explicitly address the endogeneity problem caused by the correlation of inputs and unobserved productivity. Least squares estimation will lead, ceteris paribus, to upwardly biased estimates and coefficients of more easily adjustable inputs are expected to be more affected. They argue that an

[15] Only firms that have data for at least three years of data contribute to the estimation. If a firm is only observed twice, both productivity growth and level can be estimated, but the observation does not contribute to the estimation of the technology parameters.

additional sample selection problem exists if exit is correlated with inputs. If firms exit when productivity falls below a threshold and the exit threshold is decreasing in capital, selection will bias the least squares estimate of the capital coefficient downwards.[16]

Olley and Pakes propose a three step estimator to remedy both problems.[17] They estimate the production function in (12), distinguishing the productivity term from the idiosyncratic error. They rely on a theoretical model developed by Ericson and Pakes (1995), where it is shown that investment is a function of the state variables, capital and productivity. Under some weak conditions, investment is a monotonically increasing function of productivity. The relationship can be inverted, expressing productivity as an unknown function of capital and investment. Substituting that expression for productivity in equation (12) gives the estimating equation for the first step

$$q_{it} = \alpha_0 + \alpha_l l_{it} + \phi_t(i_{it}, k_{it}) + \epsilon^1_{it}. \tag{20}$$

The unknown function $\phi_t(.)$ is approximated nonparametrically by a fourth order polynomial. In the first step, $\hat{\alpha}_l$ is estimated and $\hat{\phi}_{it}$ can be calculated, which is needed later[18].

The second step takes care of the exit decision. The intuition is that exit is conditional on the realization of productivity and the exit threshold for productivity. Both are different, unknown functions of investment and capital, approximated with a fourth order polynomial, and included on the right-hand side of a probit regression for exit. After the second step, the exit probability $\hat{P}_{it}$ is predicted, which is needed in the last step.

Finally, in the third step, the coefficient for capital is estimated. Details on the estimation are in Olley and Pakes (1996), but the intuition is straightforward. From the production function (12), one can write the conditional expectation of $q_{it} - \alpha_l l_{it}$ as $\alpha_0 + \alpha_k k_{it}$ plus the con-

[16] One mechanism that creates such dependency is a profit function that is increasing in capital. Firms with a higher capital stock expect a higher future profitability for a given level of productivity and will support larger drops in productivity before exiting the industry. An alternative mechanism that generates the same result, occurs when the market for capital is imperfect. If a bankrupt firm will only be able to sell its capital at a discount, the loss realized in liquidation will be proportional to the capital stock.

[17] An alternative method was proposed by Levinsohn and Petrin (1998), inverting the material inputs instead of investment equation. An advantage is that firms with zero investment do not have to be dropped from the estimation.

[18] The coefficients of all variable inputs in the production function are estimated in the first step.

ditional expectation of productivity in period t. Assuming that productivity evolves according to a stochastic Markov process, it is a function of its value in the previous period and the exit threshold. Productivity in the previous period can be calculated from the results in the first step as $\hat{\phi}_{it-1} - \alpha_k k_{it-1}$. An expression for the exit threshold can be obtained from the second step, because the continuation probability —the probability a firm remains in the industry— is a monotonically increasing function of the exit threshold, again an invertible relationship. The continuation probability calculated in the second step $(\hat{P}_{it})$ is used as second term in the approximation. The estimation equation for the third step is given by

$$q_{it} - \hat{\alpha}_l l_{it} = \alpha_k k_{it} + \psi_t(\hat{\phi}_{it-1} - \alpha_k k_{it-1}, \hat{P}_{it-1}) + \epsilon_{it}^2. \tag{21}$$

Only the capital coefficient is left to estimate at this stage.[19]

The main advantage of this approach is the flexible characterization of productivity. The only assumption is that productivity evolves according to a Markov process. Inputs are allowed to depend on productivity without restrictions. The disadvantage is the use of a nonparametric approximation. The investment function to be inverted is likely to be a very complicated mapping from states to actions since it has to hold for all firms regardless of their size or competitive position. Since the capital stock and productivity of all firms are part of the state vector, firms can take them into account when deciding on investment. It is unsure how good an approximation the fourth order polynomial can provide. The accuracy of the method depends on the extent to which interactions of investment, capital, the continuation probabilities, and the implicitly estimated Markov process capture variations in productivity.

[19] The unknown functions $\phi_t(.)$ and $\psi_t(.)$ are approximated by

$$\phi_t(a, b) = \sum_{j=0}^{4-m} \sum_{m=0}^{4} \phi_{mj(t)} a^m b^j$$

and similarly for ψ_t. If the sample period is sufficiently long enough, the coefficients of these approximations $(\phi_{mj(t)})$ can be made time-variant.

3.3.4 Calculating productivity

Once the coefficients in the production function are estimated, it is possible for each of the three approaches to calculate productivity from (13), dropping the last term.[20] For the instrumental variables approach, this is the only productivity measure possible. The stochastic frontier and semi-parametric methods, on the other hand, yield a direct estimate of $\hat{\omega}_{it}$, purged from random noise $\hat{\epsilon}_{it}$.

For the first stochastic frontier method it is customary to estimate firm specific technical (in)efficiency (TE) as

$$TE = E(e^{\omega_{it}} | \omega_{it} + \epsilon_{it}),$$

which is involved because of the nonlinear transformation. For our purposes, we are looking for a productivity comparison in logarithms and we can stick with the calculations in equation (13) because the best estimate of $E(\omega_{it}|\hat{\omega}_{it} + \hat{\epsilon}_{it})$ is $\hat{\omega}_{it} + \hat{\epsilon}_{it}$, if ω_{it} is independent of ϵ_{it}.

For the second stochastic frontier estimator, productivity level and growth can be calculated as

$$\begin{aligned} \log \frac{\widehat{A_{it}}}{\overline{A_t}} &= (\hat{\alpha}_{i0} + \hat{\alpha}_{i1}\ t) - (\overline{\hat{\alpha}_0} + \overline{\hat{\alpha}_1}\ t) = (\hat{\alpha}_{i0} - \overline{\hat{\alpha}_0}) + (\hat{\alpha}_{i1} - \overline{\hat{\alpha}_1})\ t && (22) \\ \log \frac{\widehat{A_{it}}}{A_{it-1}} &= \hat{\alpha}_{i1}, && (23) \end{aligned}$$

where the overlined variables denote the average over all firms active in year t.

For the semiparametric approach, the productivity level and growth can be calculated as

$$\begin{aligned} \log \frac{\widehat{A_{it}}}{\overline{A_t}} &= (\hat{\phi}_{it} - \hat{\alpha}_k k_{it}) - (\overline{\hat{\phi}_t} - \hat{\alpha}_k \overline{k_t}) && (24) \\ \log \frac{\widehat{A_{it}}}{A_{it-1}} &= (\hat{\phi}_{it} - \hat{\alpha}_k k_{it}) - (\hat{\phi}_{it-1} - \hat{\alpha}_k k_{it-1}). && (25) \end{aligned}$$

[20] This really amounts to calculating $(\hat{\omega}_{it} + \hat{\epsilon}_{it}) - (\hat{\omega}_{j\tau} + \hat{\epsilon}_{j\tau})$.

These measures can only be calculated for firms with positive investment, i.e. the firms included in the estimation procedure, while the calculations in equation (13) also apply to firms with zero investment.

Another consideration is the choice between input- and output-based productivity measures, which produce different results if the returns to scale are estimated to differ from unity. A relation is easily demonstrated if the production function is homogeneous. From the definition of the production function,

$$\frac{Q}{A_O} = F(K, L),$$

it is clear that A_O will form the basis of output-based productivity comparisons. "How much does output have to be increased (or decreased), given inputs, for the production plan to attain the production function of a comparison unit?"

Alternatively, it is possible to define an input-based productivity measure as the amount inputs have to be reduced (or can be increased), given output, for the production plan to attain the production function of another firm. The basis for such input-based comparisons, A_I, can be defined as

$$Q = F(A_I K, A_I L).$$

By definition, returns to scale are θ if the production function satisfies

$$F(\lambda K, \lambda L) = \lambda^\theta F(K, L).$$

Putting the three previous equations together, gives a one-to-one relationship between input- and output-based productivity measures.

$$\log A_O = \theta \log A_I. \tag{26}$$

Dividing the estimated ω's, which form the basis of output-based comparisons, by the estimated returns to scale gives input-based productivity comparisons.

For example, if firm k produces only 80% of the output of firm l, using the same inputs, its output-based productivity $(\frac{A^k_O}{A^l_O})$ is 0.8 or in logarithms -0.22. If returns to scale are increasing and equal to 1.5, this corresponds to an input-based productivity of 0.86 or -0.15 in logarithms. The scale economies embodied in the technology make it easier to replicate another unit's performance by reducing inputs than by increasing output.

4 A Monte Carlo simulation

4.1 Data generation

The different methodologies are compared first with a Monte Carlo simulation. The variables used in the simulations are constructed from an economic model consistent with the assumptions underlying each estimation methodology.[21]

At the core of the data generating process is a firm that chooses labor input and investment to maximize profits, subject to a production function and a capital accumulation equation:

$$\begin{aligned} &\max_{L_t, I_t} \quad Q_t - w_t L_t - r_t K_t - p_t^I I_t - g(K_t - K_{t-1}) \\ &\text{subject to} \quad Q_t = A_t L_t^{\alpha_l} K_t^{\alpha_k} e^{\alpha_t t} \\ &\qquad\qquad\quad K_t = K_{t-1} + I_t. \end{aligned} \tag{27}$$

Q_t is the value of output, w_t and r_t are (user) costs for the input factors, and p_t^I is the purchase price for new investment. The technology is Cobb-Douglas, as has been assumed before. The $g(.)$ function is a (convex) adjustment cost for capital. As a result, the capital stock is less easily adjustable than labor and the factor shares will not equal the production function parameters in the short run. For simplicity, there is no depreciation. Note that there is no source of randomness, the firm knows all variables, including the productivity term A.[22]

[21] Evidently, every method relies on only a subset of the assumptions to derive an estimator. Throughout I assume that no mistakes are made, apart from the neglect of measurement error. The functional form assumption for the production function holds and returns to scale is only restricted to unity if it is warranted.

[22] A more realistic model would have forwardlooking rather than myopic firms and the idiosyncratic shock to productivity would be realized only after some or all of the inputs are chosen. Still, the current model generates most of the correlations observed in the actual data set, used later.

The adjustment cost for capital makes it impossible to solve the two first order conditions explicitly to find an expression for the optimal values of labor and investment. Under constant or increasing returns to scale it is further necessary to pin down the size of the firm exogenously. From the first order conditions (28) and (29), it is clear, however, that both the labor and investment function will depend on productivity and both factor prices.

$$A_t L_t^{\alpha_l}(K_{t-1}+I_t)^{\alpha_k} e^{\alpha_t t} = \frac{1}{\alpha_l}(w_t L_t) \tag{28}$$

$$A_t L_t^{\alpha_l}(K_{t-1}+I_t)^{\alpha_k} e^{\alpha_t t} = \frac{1}{\alpha_k}(r_t + p_t^I + g'(I_t))(K_{t-1}+I_t). \tag{29}$$

In a more general model, Ericson and Pakes (1995) demonstrate that the investment function is monotonically increasing in productivity. I rely on their result to generate the investment series directly as

$$\log I_t = \log w_{it} - \log(r_t + p_t^I) + \phi_t^1 \log A_t, \tag{30}$$

where $\log w_{it}$, $\log(r_t + p_t^I)$, and ϕ_t^1 are generated as random variables, uniformly distributed on the unit interval for the last variable and with (1,2) as the domain for the two factor prices. To match the large incidence of zero investment in actual data sets, I_t is set to zero if $\log I_t$ is below a threshold (arbitrarily set at 5% of the capital stock). The capital series is initialized by drawing capital stock variables in period 0 directly from a Chi-squared distribution with 3 degrees of freedom.[23]

Dividing the two first order conditions for investment and labor gives

$$\frac{K_t}{L_t} = \frac{\alpha_k}{\alpha_l}\frac{w_t}{r_t + p_t^I + g'(I_t)}. \tag{31}$$

Capital and investment are generated independently and rewriting (31) provides an expression

[23] The underlying distributions for the building blocks were chosen such that the generated variables mimic the empirical distribution of the variables in the Zimbabwean data set, which is introduced later.

to generate the labor input series

$$\log L_t \;=\; \log\frac{\alpha_l}{\alpha_k}+\;\log K_t-\log w_t+\log(r_t+p_t^I)+\phi_t^2\log I_t. \tag{32}$$

The dependence of labor on investment is captured by the last term. ϕ_t^2 is generated in the same way as the coefficient multiplying productivity in the investment equation (30). The wage bill, needed for the index number calculations, is obtained by multiplying labor input by the wage rate.[24]

To generate the output variable from the inputs and the production function, the only remaining ingredient is observations on the productivity term A_t. To allow for the various effects that researchers have worried about I define

$$\log A_{it}=\omega_{it}^1+\omega_i^2+\omega_i^3 t+\rho\log A_{it-1}. \tag{33}$$

Productivity contains an idiosyncratic component ω^1 variable between firms and over time, a firm fixed-effect ω^2 and time trend ω^3 that are constant over time, and productivity follows an autoregressive process with $0<\rho<1$. The idiosyncratic component and firm fixed-effect are truncated from above, as is assumed in most of the stochastic frontier literature.[25] After truncation, the variables are recentered to have mean zero. The three variables, ω^1, ω^2, and ω^3, are generated as normally distributed random variables with mean zero. The final complication is that firms exit the sample if productivity falls below an exit threshold that depends on the capital stock. The exit threshold is modeled as a linearly decreasing function of the capital stock.

To summarize the data generating process, I group all equations:

$$\begin{aligned}\log Q_{it} &= \alpha_0+\alpha_l\log L_{it}+\alpha_k\log K_{it}+\alpha_t t+\log A_{it} && \text{if } \log A_{it}\geq a-bK_{it}\\ &= \text{missing}\quad \text{(firm exits)} && \text{otherwise}\end{aligned}$$

[24] The absolute level of the wage rate is normalized such that the average wage share in revenue matches the observed value for Zimbabwean firms. It will differ from the Cobb-Douglas coefficient on labor input because the adjustment cost for investment makes the inputs deviate from their optimal level.

[25] The assumption is that -ω_{it} in (12) is truncated from below at zero.

$$
\begin{aligned}
\log A_{it} &= \omega_{it}^1 + \omega_i^2 + \omega_i^3 t + \rho \log A_{it-1} \\
&\text{with} \quad \omega^z \sim \text{ i.i.d. } N(0,\sigma_z) \quad \text{over} \quad [-\infty, \mu_z] \qquad\qquad \text{for } z = 1,2 \\
&\text{and} \quad \omega^3 \sim \text{ i.i.d. } N(0,\sigma_3) \\
K_{it} &= K_{t-1} + I_{it} \\
&\text{with} \quad \log I_{it} \;=\; \log w_{it} - \log(r_{it} + p_{it}^I) + \phi_{it}^1 \log A_t \qquad \text{if } I_{it} \geq 0.05 \; K_{it} \\
&\qquad\qquad\qquad = \text{ missing} \quad (I_{it} = 0) \qquad\qquad\qquad \text{otherwise} \\
&\text{and} \quad \log K_{i0} \;\sim\; \chi_3^2 \\
\log L_{it} &= \log \frac{\alpha_l}{\alpha_k} + \log K_{it} - \log w_{it} + \log(r_{it} + p_{it}^I) + \phi_{it}^2 \log I_{it} \\
&\text{with} \quad w_{it}, \; r_{it} + p_{it}^I \;\sim\; \text{i.i.d. } U(1,2) \\
&\text{and} \quad \phi^1 \;, \; \phi^2 \;\sim\; \text{i.i.d. } U(0,1)
\end{aligned}
$$

The final catch is that a researcher does not observe output and inputs accurately, but with measurement error:

$$
\begin{aligned}
\hat{X} &= X + \eta_x \qquad\qquad \text{for } X = Q_{it}, L_{it}, K_{it} \\
\eta_x &\sim \text{ i.i.d. } N(0,\sigma_x)
\end{aligned}
\tag{34}
$$

The only variables a researcher observes are $\hat{Q}_{it}$, $\hat{K}_{it}$, $\hat{L}_{it}$, I_{it}, $(wL)_{it}$.[26]

To verify that the properties of the generated sample are plausible, I compare some summary statistics with those from a sample of Zimbabwean manufacturing firms. The statistics on the simulated samples are the average over 50 simulations, where three years of data has been generated for 200 firms. Table 1 confirms that many characteristics of a real sample of firms are replicated rather well.

The first rows confirms that least squares estimation of the production function will lead to an upwardly biased labor coefficient and downward bias in the capital coefficient. This effect is more pronounced if no measurement error is present. A small negative time trend on average productivity growth, combined with an autoregressive process for productivity

[26] For output, η_y is the usual random error appended to the production function for estimation. The measurement error for inputs is not controlled for in any of the methodologies.

and adjustment costs to capital produce a very low estimate for the estimated time trend. Standard errors for the productivity components and thresholds for zero investment and exit are chosen such that the ratios calculated in the second part resemble the values for the actual sample.

Table 1: Comparison of Monte Carlo samples with actual data

	CRS			VRS		
	sample	$\alpha_l = .6,\ \alpha_k = .4,\ \alpha_t = -.02$		sample	$\alpha_l = \alpha_k = .6,\ \alpha_t = -.02$	
		$\sigma_x = 0$	$\sigma_x = 0.5$		$\sigma_x = 0$	$\sigma_x = 0.5$
OLS regression of output on labor, capital and time:						
$\hat{\alpha}_l$	0.499	0.869	0.751	0.809	1.001	0.926
(s.e.)	(0.0)	(0.0)	(0.0)	(.055)	(0.021)	(0.026)
$\hat{\alpha}_k$	0.501	0.131	0.249	0.392	0.191	0.254
(s.e.)	(.030)	(0.035)	(0.036)	(.033)	(0.024)	(0.030)
$\hat{\alpha}_t$	-0.134	-0.088	-0.091	-0.089	-0.064	-0.064
(s.e.)	(.062)	(0.046)	(.057)	(.059)	(0.037)	(0.052)
R^2	0.415	0.034	0.085	0.881	0.953	0.911
I/K	0.386	0.165	0.187	0.386	0.160	0.182
W/Q	0.492	0.524	0.596	0.492	0.480	0.458
exit (% at t=1)	0.032	0.034	0.034	0.032	0.038	0.038
I=0 (%)	0.015	0.059	0.059	0.015	0.126	0.126
Correlations:						
q-k	0.901	0.876	0.843	0.901	0.871	0.843
q-l	0.915	0.942	0.909	0.915	0.973	0.948
q-w	-0.578	-0.298	-0.309	-0.578	-0.620	-0.547
q-i	0.845	0.875	0.859	0.845	0.911	0.900
k-l	0.876	0.913	0.877	0.876	0.856	0.826
k-w	-0.409	0.093	0.071	-0.409	-0.356	-0.284
k-i	0.859	0.968	0.947	0.859	0.959	0.941
l-w	-0.347	-0.132	-0.106	-0.347	-0.594	-0.457
l-i	0.841	0.892	0.875	0.841	0.874	0.861
w-i	-0.389	0.037	0.030	-0.389	-0.502	-0.388

The most important finding from Table 1 is the large degree of similarity between the correlations of all observable variables for the simulated and actual samples. Especially the correlations under variable returns to scale when output and inputs are measured with error are very similar (compare the bottom part of the fourth and sixth column). Correlations between output and inputs are large and positive and the correlation of output with labor exceeds that with capital. Investment is positively correlated with output and inputs, especially with capital. The wage bill is strongly negative correlated with all other variables. The magnitudes of all correlations are surprisingly similar and invariable of the correct sign.

The most noticeable difference for the samples generated assuming constant returns to scale is the lower correlation of the wage bill with other variables, which sometimes even turns, albeit very small.

4.2 Comparison of methodologies

Using the simulated samples, productivity levels ($\log TFP_{it} = \log \frac{A_{it}}{A_t}$) and growth rates ($TFPG_{it} = \log \frac{A_{it}}{A_{it-1}}$) are estimated using all the previously discussed methodologies. As a benchmark, productivity measures are also calculated using the least squares estimates of the production function parameters in equation (13).

The principal criterion to evaluate the results is the mean squared error between the actual and estimated productivity factors:

$$\begin{aligned} \text{MSE}^{\text{P}} &= \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} (\log TFP_{it} - \log \widehat{TFP}_{it})^2 && (35) \\ \text{MSE}^{\text{PG}} &= \frac{1}{N(T-1)}\sum_{i=1}^{N}\sum_{t=2}^{T} (TFPG_{it} - \widehat{TFPG}_{it})^2. && (36) \end{aligned}$$

Because the size of the estimates might be off even though the relative position of firms in the productivity rankings are accurate, I also calculate the correlation between estimated and actual productivity measures.[27]

The results for the output-based productivity measures allowing variable returns to scale are in Table 2. The three panels contain the results for increasing degrees of measurement error on output and inputs, σ_x in (34). In the top panel, the researcher observes output and inputs accurately. In the middle panel, measurement error amounting to around 20% of the

[27] The Spearman rank-correlation did not produce substantially different results from the partial correlations, except in some rare occasions, which are noted in the text.

original variation in the variables is added. In the bottom panel, variables are observed with a lot of measurement error, increasing the variation of the variables by around 75%.

The mean squared errors (MSE) increase substantially with measurement error, but the correlations are less affected. The large MSE for some of the parametric methods in the first panel are the result of unstable coefficient estimates when even output is measured accurately and the residual in the production function is superfluous. The relative performance of the different methodologies is hardly affected by the presence of small amounts of measurement error, but the ranking undergoes some change in the bottom panel, when the error become large. The drop in accuracy is larger for the Törnqvist index and DEA that do not accommodate errors well, but a large amount of measurement error is required before it becomes desirable to purge the regression error from productivity estimates as in the second semiparametric approach and stochastic frontier 2.

The parametric methods that do not attempt to purge the measurement error from the productivity estimates perform consistently the best. For the productivity level calculations, the GMM, Stochastic Frontier 1 (where the ranking of firms is constant over time), the semiparametric, and even the least squares approaches perform equally well. To estimate productivity growth levels, the semiparametric method turns out to be the best of the bunch. Comparing the different panels in Table 3 reveals to what extent these methods are sensitive to measurement error. The correlations between actual and estimated productivity for GMM, SF1, and OLS fall by 20% if moderate measurement error is introduced and a further 50% if measurement error becomes very large. The same pattern holds for the level and growth comparisons. The semiparametric method is significantly more sensitive to large amounts of measurement error lowering the correlation for productivity level by two thirds if measurement error becomes important. For growth comparisons it also sees a slightly larger reduction in the correlation, but is remains ahead of the other methods.[28]

At the other side of the spectrum are two methods that consistently underperform the alternatives. The second stochastic frontier approach (with fixed-effects and firm specific

[28] It is important to realize that the results for the parametric estimators are certain to be a best-case scenario. The problem of specification error for the functional form of the production function has been ruled out by assumption.

time trends) and the semiparametric approach where estimates are purged from measurement error. Both methods attempt to remove random measurement error from the productivity estimation, but without much success. Only when the measurement error becomes really large do they outperform some of their competitors. The semiparametric estimator of productivity growth even becomes the best method overall for growth comparisons. The correlations with true productivity growth hardly decrease when measurement error is increased for this method. It remains puzzling, however, why the two semiparametric calculations differ so

Table 2: Monte Carlo results with variable returns to scale

	productivity level		productivity growth	
	MSE	Correlation	MSE	Correlation
Without measurement error on output and inputs ($\sigma_x = 0$):				
OLS	0.295	0.790	0.474	0.682
Törnqvist Index	0.515	0.641	1.005	0.469
DEA	0.718	0.606	0.865	0.553
GMM	0.246	0.829	0.387	0.737
Stochastic frontier 1	0.252	0.824	0.409	0.722
Stochastic frontier 2	53.387	0.115	0.798	0.371
Semiparametric ($\omega + \epsilon$)	6.229	0.813	0.114	0.934
Semiparametric (ω)	6.631	0.496	0.370	0.514
With some measurement error on output and inputs ($\sigma_x = 0.5$):				
OLS	0.660	0.633	1.254	0.512
Törnqvist Index	0.947	0.495	1.755	0.360
DEA	1.113	0.552	1.651	0.373
GMM	0.640	0.646	1.213	0.526
Stochastic frontier 1	0.633	0.651	1.216	0.525
Stochastic frontier 2	1.551	0.481	0.762	0.407
Semiparametric ($\omega + \epsilon$)	0.686	0.743	0.810	0.704
Semiparametric (ω)	0.929	0.407	0.368	0.509
With a lot of measurement error on output and inputs ($\sigma_x = 2$):				
OLS	6.147	0.334	11.217	0.274
Törnqvist Index	34.882	0.215	146.483	0.138
DEA	1.701	0.056	3.460	0.018
GMM	6.227	0.335	11.836	0.266
Stochastic frontier 1	6.215	0.332	11.438	0.270
Stochastic frontier 2	9.806	0.221	2.925	0.219
Semiparametric ($\omega + \epsilon$)	10.367	0.259	8.485	0.347
Semiparametric (ω)	6.368	0.127	1.063	0.406

widely. Removing the random noise is apparently not done very randomly. In defense of the stochastic frontier estimator, it should be pointed out that it suffers from the short length of the panel and would likely perform better if more years were available. It performs best for productivity level comparisons, as the firm specific time trends are estimated very imprecise, and finished mid-pack if measurement error becomes large.

Not surprisingly the DEA and index numbers approaches are also very sensitive to measurement error as they do not allow for any randomness in the calculations.[29] For productivity level estimations both methods perform mid-pack as long as measurement error does not become too large. Especially DEA turns out to be highly sensitive to large amounts of measurement error, and the results become completely unreliable in the bottom panel. It is surprising that it produces better results than the Törnqvist index to estimate productivity growth, even though it is not intended for that purpose. The index numbers, on the other hand, outperform DEA for level comparisons, for which they are less often used.

Table 3 provides some insight on what is driving the results. It contains the coefficient estimates for the parametric methods and comparable statistics for the index numbers. The MSE column contains the sum of the mean squared error for both input coefficient estimates.

Table 3: Coefficient estimates allowing for variable returns to scale (average over 50 runs)

	no measurement error				some measurement error				much measurement error			
	$\hat{\alpha}_l$	$\hat{\alpha}_k$	RTS	MSE	$\hat{\alpha}_l$	$\hat{\alpha}_k$	RTS	MSE	$\hat{\alpha}_l$	$\hat{\alpha}_k$	RTS	MSE
actual	0.60	0.60	1.20	0.00	0.60	0.60	1.20	0.00	0.60	0.60	1.20	0.00
OLS	0.96	0.22	1.19	0.28	0.89	0.28	1.17	0.19	0.49	0.42	0.91	0.05
Index*	0.68	0.51	1.19	0.27	0.64	0.54	1.18	0.26	0.43	0.48	0.90	0.08
GMM	0.93	0.25	1.18	0.23	0.88	0.30	1.17	0.17	0.52	0.47	0.99	0.03
SF1	0.94	0.27	1.20	0.23	0.87	0.32	1.19	0.16	0.53	0.40	0.93	0.05
SF2	0.76	4.30	5.06	7.48	0.71	0.08	0.79	0.30	0.14	0.02	0.16	0.56
SP	0.72	1.15	1.87	2.29	0.61	0.44	1.06	0.05	0.17	0.04	0.21	0.51

* The reported statistic is the wage share in revenue as a percentage of the inverse of RTS

The OLS estimator results in an upwardly biased labor coefficient and downward bias

[29] For the Törnqvist index number calculations, I obtained an estimate of scale economics using least squares. Table 3 shows that the overestimation of the labor coefficient cancels out with the underestimation of capital, yielding a rather precise estimate of scale economics.

for the capital coefficient, relative to the true values of 0.6. Both biases are diminished by the more complicated methods, but with varying degrees of success. The results for the instrumental variable estimator, implemented as a system using GMM, and the first stochastic frontier implementation, hardly differs from the OLS results. The changes are in the right direction though. With large amounts of measurement error, they become more distinct from OLS and also more accurate.

The second stochastic frontier estimator and the semi-parametric approach of Olley and Pakes (1996) correct more for the underestimation of the labor coefficient, but they strongly overestimate the capital coefficient if there is no measurement error. With moderate measurement error the semiparametric approach does a remarkably good job for both coefficients. It is not very reliable though as it overcorrects for potential bias if no measurement error is present and it yields to very large downward bias if measurement error becomes large. The stochastic frontier with dummies produces even more unreliable results, but imposing constant returns to scale it becomes the most consistent and accurate method (results in the Appendix).

The returns to scale are in most cases rather accurately estimated, as the bias in both input coefficients cancels out. With no or moderate measurement error most methods find increasing returns to scale and the point estimates are much more alike than for the individual coefficients. With large amounts of measurement error, both input coefficients are significantly underestimated and no method finds evidence for increasing returns anymore.

The index number calculations rely on returns to scale estimated using OLS. The statistics in the table are average wage and capital shares expressed as a percentage of the inverse of the returns to scale estimate. These numbers are not affected by the endogeneity of productivity that caused parametric approached to be biased, but they vary across firms.

Restricting the returns to scale to unity, if the underlying technology of the data generating process is indeed constant returns, vastly improves the accuracy of the parameter estimates. The productivity estimates are also much closer to the true values for all methods without measurement errors. Moderate measurement error seems to have a more significant impact than in the variable returns to scale case. The different methodologies produce very similar results for productivity levels and only the stochastic frontier 2 is really out of line for productivity growth estimates. Overall, the GMM method generates the best results under constant returns to scale. Tables with results are in the Appendix.

5 An empirical application: Manufacturing plants in Zimbabwe

5.1 Data collection

Next, I evaluate the different methodologies when applied to a sample of actual plants. The data comes from firm surveys carried out between 1993 and 1995 in Zimbabwe. Approximately 200 firms were interviewed in three consecutive years. Only firms with nonmissing data on output, inputs, costs, wages and investment are retained. Firms come from four different manufacturing sectors: food, textile, wood, and metal, corresponding roughly to the ISIC classification 31, 32, 33, and 38. Some firms exit the sample each year and new firms were added in later rounds to maintain the sample size.[30] More information about the survey and sample construction can be found in Van Biesebroeck (2001).

As before, I estimate value added production functions, using sales minus indirect costs and material input as output, total employment as labor input, and the reported replacement value of the plant and equipment as capital input. Value added and capital are deflated using the manufacturing deflator from the IMF Financial Tables.[31] Table 4 contains some summary statistics on the sample.

5.2 Comparison of methodologies, revisited

Productivity levels and growth rates are estimated using each methodology, both for the constant and variable returns to scale case. The coefficient estimates for the production

[30] Exit from the sample is not only caused by bankruptcy or relocation. In half of the exits, firms refused to cooperate in later rounds or the appropriate person to fill out the questionnaire was not present on three visits.

[31] In absence of more detailed indices both variables are transformed using the GDP deflator.

Table 4: Averages for the sample of manufacturing plants in Zimbabwe

	1993	1994	1995
Number of firms	126	150	116
Output	179064	159925	154523
(standard deviation)	(386266)	(352256)	(362440)
Labor	365	348	293
(standard deviation)	(661)	(754)	(495)
Capital	162439	368952	199765
(standard deviation)	(536477)	(1482130)	(639587)
Output growth		0.04	0.12
Labor growth		0.04	0.04
Capital growth		0.72	-0.09
Wage share (% of revenue)	0.47	0.51	0.50
Investment (% of capital stock)	0.62	0.31	0.26
Zero investment (% of firms)	0.02	0.01	0.02
Exiting firms (proportion of total)	0.03	0.39	
Entering firms (proportion of total)		0.27	0.11

function parameters and comparable statistics for the Törnqvist index are in Table 5.[32]

If constant returns to scale are enforced the parameters are about the same for the first three methods. Only the stochastic frontier estimators give different and opposite results. Recall, that the simulation results in Table A.2 in the Appendix demonstrated that the second stochastic frontier approach was the most accurate of all methods, if the constraint on scale economies is enforced. This suggest that the labor coefficient is likely to be larger than the capital coefficient, but the magnitude of the difference can be doubted.

If returns to scale are left free, all but one method finds positive and significant scale economies of around 1.2. This is not implausible for a relatively underdeveloped manufacturing sector in a small African country. The coefficient estimates with the second stochastic frontier approach look altogether implausible, as was the case in the simulations. A panel of only three years is clearly not sufficient to obtain accurate results.

[32] In the variable returns to scale case, the wage share necessary for the index number calculations is observed. Returns to scale are estimated using least squares and a capital share is calculated to be consistent with the other two statistics.

Table 5: Coefficient estimates for the production function

	CRS		VRS		
	α_l	α_k	α_l	α_k	RTS
OLS	0.50	0.50	0.82	0.39	1.21
Törnqvist Index*	0.49	0.51	0.72	0.49	1.21
GMM	0.50	0.50	0.70	0.43	1.13
Stochastic frontier 1	0.09	0.91	0.70	0.64	1.34
Stochastic frontier 2	0.89	0.11	0.38	0.10	0.48
Semiparametric	.	.	0.75	0.36	1.11

* The statistic is the average revenue share as a percentage of the inverse of RTS

As expected, the least squares estimation which does not correct for any potential sources of bias, produces the largest labor and smallest capital coefficient estimate, consistent with the previous discussion. The average estimates for the four trustworthy methods are 0.72 for the labor coefficient and 0.48 for capital, which look plausible.

Table 6 lists the corresponding productivity estimates for the different methods. The interquartile ranges are very similar, which is surprising because the methods rely on very different calculations and assumptions. Only the second semiparametric approach produces much smaller ranges, while the second stochastic frontier produces very wide ranges, especially with variable returns to scale. These are the only two methods that attempted to separate random measurement error from productivity. The semiparametric approach suggests that only a third of the measures produced by the other methods should be identified as productivity, while two thirds is error. The stochastic frontier results, on the other hand, are inconsistent with the results obtained by other methods.

The estimated ranges are relatively wide. Only half of the firms have a productivity level between 50% below and 50% above the average firm in the sample. Productivity dispersion in the U.S. manufacturing sector is estimated to be lower, but not by much, see for example Baily, Hulten, and Campbell (1992).[33] Most of the intervals are wider above the average, which is by definition zero. Productivity turns out to be right-skewed, which favors the assumption of the semiparametric approach —firms with a productivity level below

[33] No trimming of the data to remove outliers has been performed.

Table 6: Interquartile range for the productivity level

	CRS		VRS output-based		VRS input-based	
	25th %	75th %	25th %	75th %	25th %	75th %
OLS	-0.54	0.60	-0.55	0.55	-0.46	0.45
Törnqvist Index	-0.45	0.73	-0.30	0.78	-0.25	0.65
DEA	-0.52	0.63	-0.54	0.61	-0.57	0.56
GMM	-0.54	0.60	-0.51	0.55	-0.45	0.48
Stochastic frontier 1	-0.71	0.75	-0.73	0.64	-0.55	0.48
Stochastic frontier 2	-0.46	0.72	-0.90	1.23	-1.86	2.55
Semiparametric ($\omega + \epsilon$)	.	.	-0.51	0.55	-0.46	0.50
Semiparametric (ω)	.	.	-0.11	0.22	-0.10	0.20

the exit-threshold leave the industry— over the assumption of the semiparametric approach —positive outliers are not observed because they shift out the technology frontier.

Unlike the usual practice for index number estimate, the estimated presence of increasing returns to scale does not automatically raises productivity estimates if constant returns to scale are enforced. The formulas in (9) and (8) deduct the scale effect from productivity.[34] It is therefore unpredictable which way the effect will go. It turns out that the output-based estimates are very similar to the constant returns to scale results, albeit the ranges are slightly narrower. A small part of the productivity advance of larger firms (see later) is absorbed by the technology, when allowed. The increasing returns lead unambiguously to narrower ranges for the input-based measures than the (more frequently used) output-based measures. The reverse results apply to the second stochastic frontier approach, that estimates decreasing scale economies.

The average productivity growth estimates in Table 7 show that the period from 1993-1995 was not a very successful period for Zimbabwean manufacturing. The large increase

[34] The unobservability of capital prices forced me to obtain an outside estimate for returns to scale ($\hat{\epsilon}$)and calculate:

$$\begin{aligned} TFPG_{\text{output-based}} &= \dot{q} - (s^l\hat{\epsilon})\dot{l} - (1 - s^l\hat{\epsilon})\dot{k} \\ TFPG_{\text{input-based}} &= \frac{1}{\hat{\epsilon}}\dot{q} - (s^l)\dot{l} - (\frac{1}{\hat{\epsilon}} - s^l)\dot{k} \end{aligned}$$

in capital without corresponding decrease in labor input or output increase makes this result inevitable. Almost all statistics are negative and often large in absolute value. On average, productivity declined by 6% per year, assuming variable returns to scale. The two methods that take out measurement error, the second stochastic frontier and semiparametric approaches, both produce less negative results. Those results, if trusted, suggest that measurement error is declining, as $\hat{\epsilon}_{it} - \hat{\epsilon}_{it-1} < 0$, and productivity growth is higher than suggested by the other methods.

The results also reveal that the weighted average of productivity growth is even more negative than the unweighted averages. This finding is consistent across methods, does not depend on the returns to scale assumption, and it even holds using current output weights, which favors firms with positive productivity growth. It implies that output is relatively relocated from enterprises with above average productivity to less productive enterprises. This relocation diminishes aggregate productivity. In most countries, the reverse effects have been found. For more details on similar findings in several African countries, see Van Biesebroeck (2002).

Table 7: Weighted and unweighted productivity growth averages

	CRS		VRS output-based		VRS input-based	
	mean	weighted	mean	weighted	mean	weighted
OLS	-0.083	-0.117	-0.063	-0.111	-0.052	-0.092
Törnqvist Index	-0.072	-0.132	-0.079	-0.130	-0.066	-0.108
DEA	-0.056	-0.046	-0.022	-0.040	-0.029	-0.040
GMM	-0.082	-0.118	-0.072	-0.103	-0.064	-0.092
Stochastic frontier 1	-0.166	-0.140	-0.104	-0.126	-0.078	-0.094
Stochastic frontier 2	0.010	-0.098	0.019	-0.064	0.039	-0.132
Semiparametric($\omega + \epsilon$)	.	.	-0.059	-0.108	-0.053	-0.098
Semiparametric (ω)	.	.	-0.010	-0.016	-0.009	-0.014

The first column contains the simple average; in the second column, output weights are used

The previous section with simulation results showed a very different performance for the different methodologies if measurement error is present. The correlations in Tables 8, 9, and 10 show that the methods produce surprisingly similar productivity estimates. Especially imposing constant returns to scale, in Table 8, all methods produce results that are highly positive correlated. With variable return, the correlations are still positive, see Table 9, but substantially lower.

Table 8: Correlations between productivity level estimates under constant returns to scale

	OLS	Index	DEA	GMM	SF 1
OLS	1				
Törnqvist Index	0.78	1			
DEA	0.95	0.72	1		
GMM	0.99	0.77	0.95	1	
Stochastic frontier 1	0.82	0.54	0.67	0.82	1
Stochastic frontier 2	0.83	0.74	0.90	0.83	0.37

Trying to group methods that produce the most similar results is straightforward. The OLS, GMM, and semiparametric estimates are very much alike. Even DEA produces results that are similar, especially under constant returns to scale. The two stochastic frontier approaches and the index numbers are less closely related to any particular method, but correlations are still highly positive. The second semiparametric estimator is most highly correlated with the second stochastic frontier, the only other method that also attempts to take out measurement error. It is interesting to note that all methods except for OLS and SF1 have the highest correlation with the semiparametric method. It makes that method attractive as it is highly correlated with every single methodology and it performed very well in the previous section. For completeness, I should mention that Spearman-rank correlations between the second stochastic frontier approach and all other methods were substantially lower, except with the DEA method. These two methods place the least restrictions on productivity. The index numbers generate much larger Spearman rank-correlations with the parametric methods (not reported).

Correlations for different productivity growth calculations are in table 10. The OLS, GMM, and semiparametric methods go from being much alike to being almost indistinguishable. All their cross-correlations exceed 0.99. Both stochastic frontier approaches yield similar results as well, especially with the GMM and semiparametric approaches, which again seem to be the best compromise. The relatively small differences in coefficient estimates for the parametric estimators are swamped by the huge differences in output and input growth rates

Table 9: Correlations between output-based productivity level estimates under variable returns to scale

	OLS	Index	DEA	GMM	SF 1	SF 2	SP $(\omega+\epsilon)$
OLS	1						
Törnqvist Index	0.68	1					
DEA	0.81	0.60	1				
GMM	0.99	0.70	0.85	1			
Stochastic Frontier 1	0.85	0.40	0.60	0.81	1		
Stochastic Frontier 2	0.50	0.63	0.67	0.57	0.01	1	
Semiparametric $(\omega+\epsilon)$	0.97	0.74	0.86	0.99	0.73	0.67	1
Semiparametric (ω)	0.20	0.49	0.20	0.26	-0.13	0.65	0.33

across firms. As a result, all parametric methods yield very similar results.[35]

In fact, all correlations are reassuringly high except for the semiparametric estimator where measurement error is taken out. That method has produces results that have hardly anything in common with any other methodology, save for the second stochastic frontier approach. The Törnqvist index produces results that also differ from the parametric alternatives. The DEA results, which only have a natural cross-section interpretation and had to be transformed to carry out the growth comparisons, are still highly correlated with the parametric results.

Table 10: Correlations between output-based productivity growth estimates under variable returns)

	OLS	Index	DEA	GMM	SF1	SF2	SP $(\omega+\epsilon)$
OLS	1						
Törnqvist Index	0.65	1					
DEA	0.92	0.55	1				
GMM	0.998	0.65	0.92	1			
Stochastic Frontier 1	0.96	0.63	0.90	0.98	1		
Stochastic Frontier 2	0.73	0.60	0.65	0.72	0.63	1	
Semiparametric $(\omega+\epsilon)$	0.999	0.65	0.92	0.996	0.95	0.75	1
Semiparametric (ω)	0.03	0.21	-0.02	0.03	0.02	0.11	0.04

[35] Under constant returns to scale one of the few differences is the greater similarity between the index numbers and parametric methods. Under the same assumption, the first stochastic frontier generates results most alike the DEA method, while it resembles the parametric methods under variable returns to scale.

5.3 Validity check

It is impossible to know what the true productivity is, but it is possible to compare the results with findings that have received strong support in the US and other developed countries. Three stylized facts are checked. First, firms that export are usually found to have higher productivity, necessary to overcome entry barriers or cost associated with export markets. Second, larger firms are often found to have higher productivity. This has been strongly confirmed in many developing economies. Often it is conjectured that inefficient capital markets make it harder for firms to realize scale economies. Third, firms that invest in and adopt new technology improve productivity at a faster pace. I use a survey question that asked firms whether they invested in foreign technology in the previous year.

Table 11: Reasonableness test for the results

	exporter?			firm-size			technology efforts		
	yes	no		large	small		some	none	
Under constant returns to scale									
OLS	0.19	-0.24	**	0.14	-0.06	*	-0.09	-0.19	
Törnqvist Index	0.20	-0.25	**	0.24	-0.10	*	-0.11	-0.21	
DEA	0.25	-0.31	**	0.25	-0.11	**	-0.05	-0.22	
GMM	0.19	-0.24	**	0.14	-0.06	*	-0.09	-0.20	
Stochastic frontier 1	0.02	-0.03		0.07	-0.17	**	-0.20	-0.35	
Stochastic frontier 2	0.36	-0.44	**	0.44	-0.19	**	0.18	0.00	**
Under variable returns to scale									
OLS	0.04	-0.05		-0.18	0.08	**	-0.07	-0.15	
Törnqvist Index	0.26	-0.31	**	0.33	-0.14	**	-0.09	-0.18	
DEA	0.24	-0.29	**	0.37	-0.16	**	-0.08	-0.17	
GMM	0.10	-0.12	**	-0.06	0.02		-0.07	-0.16	
Stochastic frontier 1	-0.20	0.23	**	-0.63	0.27	**	-0.14	-0.24	
Stochastic frontier 2	0.87	-1.05	**	1.46	-0.62	**	0.05	0.28	**
Semiparametric ($\omega + \epsilon$)	0.15	-0.18	**	0.04	-0.02		-0.05	-0.14	
Semiparametric (ω)	0.11	-0.14	**	0.14	-0.06	**	0.01	-0.13	*

The first four columns contain the average productivity level by category, the last two the average productivity growth.
**=significant at 1% level, *=at 5% level

The first fact receives near universal support. The productivity advantage for exporters is more pronounced if the technology is limited to constant returns to scale. Exporters are on average larger and the next columns demonstrate that large firms are more productive. If variable returns to scale are allowed, some of the productivity advance is attributed to the technology and not to export status. The reverse happens for the stochastic frontier estimates because returns to scale were estimated to be decreasing.

The second fact gets support under constant returns to scale, but many methods find opposite results when variable returns are allowed. The third and fourth columns in Table 11 are one of the few occasions where the various parametric methods produce different results. The least squares and GMM estimators seem to have overestimated returns to scale, as they both show large firms to have a substantially lower productivity level. The semiparametric method generates the most plausible results of the group. Even the alternative semiparametric method, where measurement error is taken out, reaches similar conclusions. The differences between large and small firms are of the same order of magnitude as the difference by export status. Not surprisingly, because both criteria are likely to split the sample similarly.

Finally, the third fact receives unqualified support using each method. Firms that invest in foreign technology improve productivity on average 12% faster than other firms, or in the case of Zimbabwe they experience less negative productivity growth. The magnitude of this result is large, but still reasonable and very consistent. It does not depend on the returns to scale assumption nor the methodology.

6 Lessons

[Still to add]

References

Aigner, D., C. K. Lovell, and P. Schmidt (1977). Formulation and Estimation of Stochastic Frontier Production function Models. *Journal of Econometrics 6*, 21–37.

Baily, M. N., C. Hulten, and D. Campbell (1992). Productivity Dynamics in Manufacturing Plants. *Brookings Papers: Microeconomics 4*(1), 187–267.

Battese, G. E. and T. J. Coelli (1992). Frontier Production Functions, Technical Efficiency and Panel Data: With Application to Paddy Farmers in India. *Journal of Productivity Analysis 3*, 153–69.

Berndt, E. R. and M. A. Fuss (1986, Oct./Nov.). Productivity Measurement with Adjsutments for Variations in Capacity Utilization and Other Forms of Temporary Equilibrium. *Journal of Econometrics 33*(1/2), 7–29.

Berndt, E. R. and M. S. Khaled (1979). Parametric Productivity Measurement and Choice among Flexible Functional Forms. *Journal of Political Economy 87*(6), 1220–45.

Blundell, R. W. and S. R. Bond (1998). Initial Conditions and Moment Restrictions in Dynamic Panel Data Models. *Journal of Econometrics 87*, 115–43.

Blundell, R. W. and S. R. Bond (2000). GMM Estimation with Persistent Panel Data: An Application to Production Functions. *Econometric Reviews 19*(3), 321–340.

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

Caves, D. W., L. R. Christensen, and E. W. Diewert (1982b). Multilateral Comparisons of Output, Input, and Productivity using Superlative Index Numbers. *Economic Journal 92*, 73–86.

Charnes, A., W. W. Cooper, and E. Rhodes (1978). Measuring the Efficiency of Decision Making Units. *European Journal of Operational Research 2*, 429–444.

Cornwell, C., P. Schmidt, and R. C. Sickles (1990). Productivity Frontiers with Cross-sectional and Time Series Variation in Efficiency Levels. *Journal of Econometrics 46*, 185–200.

Denison, E. F. (1972). Some Major Issues in Productivity Analysis: an Examination of the Estimates by Jorgenson and Griliches. *Survey of Current Business 49*(5, Part II), 1–27.

Denny, M., M. Fuss, and L. Waverman (1981). The Measurement and Interpretation of Total Factor Productivity in Regulated Industries, with an Application to Canadian Telecommunications. In T. Cowing and R. Stevenson (Eds.), *Productivity Measurement in Regulated Industries*, pp. 179–218. New York: Academic Press.

Diewert, W. E. (1976). Exact and Superlative Index Numbers. *Journal of Econometrics 4*, 115–145.

Ericson, R. and A. Pakes (1995, January). Markov Perfect Industry Dynamics: A Framework for Empirical Work. *Review of Economic Studies 62*(1), 53–85.

Farrell, M. J. (1957). The Measurement of Productive Efficiency. *Journal of the Royal Statistical Society, Series A. 120*, 253–290.

Griliches, Z. and V. Ringstad (1971). *Economies of Scale and the Form of the Production Function.* Amsterdam: North Holland.

Huang, C. J. and J.-T. Liu (1994, June). Estimation of a Non-neutral Stochastic Frontier Production Function. *Journal of Productivity Analysis 5*(2), 171–80.

Jorgenson, D. W. and Z. Griliches (1967, July). The Explanation of Productivity Change. *Review of Economic Studies 34*, 349–83.

Jorgenson, D. W. and Z. Griliches (1972). Issues in Growth Accounting: A Reply to Edward F. Denison. *Survey of Current Business 52*, 65–94.

Klette, T. J. and Z. Griliches (1996). The Inconsistency of Common Scale Estimators When Output Prices Are Unobserved and Endogenous. *Journal of Applied Econometrics 11*(4), 343–361.

Levinsohn, J. and A. Petrin (1998). *When Industries Become more Productive, Do Firms?: Investigating Productivity Dynamics.* NBER Working Paper No. 6893.

Mairesse, J. and Z. Griliches (1990). Heterogeneity in Panel Data: Are There Stable Production Functions. In P. C. et al. (Ed.), *Essays in Honor of Edmond Malinvaud*, Volume 3, pp. 125–147. Cambridge: MIT Press.

Mairesse, J. and Z. Griliches (1998). Production Functions: The Search for Identification. In S. Strom (Ed.), *Econometrics and Economic Theory in the Twentieth Century: The Ragnar Frisch Centennial Symposium*, pp. 169–203. Cambridge: Cambridge University Press.

Marschak, J. and W. H. Andrews (1944). Random Simultaneous Equations and the Theory of Production. *Econometrica 12*, 143–205.

Meeusen, W. and J. van den Broeck (1977). Efficiency Estimation from Cobb-Douglas Production functions with Composed Error. *International Economic Review 8*, 435–44.

Olley, G. S. and A. Pakes (1996). "The Dynamics of Productivity in the Telecommunications Equipment Industry". *Econometrica 64*(6), 1263–97.

Schmidt, P. and R. C. Sickles (1984). Productivity Frontiers and Panel Data. *Journal of Business and Economic Statistics 2*, 367–374.

Seiford, L. L. and R. M. Thrall (1990). Recent Developments in DEA. The Mathematical Programming Approach to Frontier Analysis. *Journal of Econometrics 46*, 7–38.

Solow, R. M. (1957). Technical Change and the Aggregate Production Function. *Review of Economics and Statistics 39*, 312–320.

Stevenson, R. E. (1980). Likelihood Functions for Generalized Stochastic Frontier Estimation. *Journal of Econometrics 13*, 57–66.

Stigler, G. J. (1976, March). Xistence of X-efficiency. *American Economic Review 66*(1), 213–16.

Van Biesebroeck, J. (2001, July). Exporting Raises Productivity, At Least in Sub-Saharan African Manufacturing Firms. mimeo. Stanford University.

Van Biesebroeck, J. (2002, April). Comparing the Size and Productivity Distribution of Manufacturing Plants in sub-Saharan Africa and the United States. mimeo University of Toronto.

A Monte Carlo results imposing CRS

I did not perform the semiparametric estimator with constant returns to scale because the three step procedure is not adjusted easily to impose the coefficient restriction if there are only two inputs.

Table A. 1: Monte Carlo results with constant returns to scale

	productivity level		productivity growth	
	MSE	Correlation	MSE	Correlation
Without measurement error on output and inputs:				
OLS	0.073	0.953	0.137	0.914
Törnqvist Index	0.161	0.921	0.406	0.806
DEA	0.070	0.960	0.133	0.925
GMM	0.048	0.969	0.089	0.944
Stochastic frontier 1	0.078	0.950	0.147	0.908
Stochastic frontier 2	0.130	0.914	0.543	0.566
With measurement error on output and inputs:				
OLS	0.425	0.788	0.862	0.675
Törnqvist Index	0.612	0.735	1.689	0.542
DEA	0.483	0.791	0.956	0.667
GMM	0.408	0.799	0.823	0.690
Stochastic frontier 1	0.430	0.786	0.871	0.673
Stochastic frontier 2	0.384	0.773	0.764	0.419

Table A. 2: Coefficient estimates imposing constant returns to scale (average over 50 runs)

	no measurement error			with measurement error		
	$\hat{\alpha}_l$	$\hat{\alpha}_k$	MSE	$\hat{\alpha}_l$	$\hat{\alpha}_k$	MSE
actual	0.60	0.40	0.00	0.60	0.40	0.00
OLS	0.85	0.14	0.13	0.75	0.25	0.05
Törnqvist Index	0.52	0.47	0.01	0.59	0.41	0.01
GMM estimator	0.80	0.21	0.09	0.70	0.30	0.03
Stochastic frontier 1	0.86	0.14	0.14	0.75	0.25	0.05
Stochastic frontier 2	0.73	0.27	0.04	0.65	0.35	0.01