Spatial Economic Analysis
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/rsea20

Agglomeration Economies, Regional Growth, and the Aggregate Production Function: A Caveat Emptor for Regional Scientists
Jesus Felipe & John McCombie


To cite this article: Jesus Felipe & John McCombie (2012): Agglomeration Economies, Regional Growth, and the Aggregate Production Function: A Caveat Emptor for Regional Scientists, Spatial Economic Analysis, 7:4, 461-484

To link to this article: http://dx.doi.org/10.1080/17421772.2012.722668

PLEASE SCROLL DOWN FOR ARTICLE

Full terms and conditions of use: http://www.tandfonline.com/page/terms-and-conditions

This article may be used for research, teaching, and private study purposes. Any substantial or systematic reproduction, redistribution, reselling, loan, sub-licensing, systematic supply, or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The accuracy of any instructions, formulae, and drug doses should be independently verified with primary sources. The publisher shall not be liable for any loss, actions, claims, proceedings, demand, or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.
Agglomeration Economies, Regional Growth, and the Aggregate Production Function: A Caveat Emptor for Regional Scientists

JESUS FELIPE & JOHN McCOMBIE

(Received January 2012; accepted May 2012)

ABSTRACT Over the last 20 years or so, mainstream economists have become more interested in spatial economics and have introduced largely neoclassical economic concepts and tools to explain phenomena that were previously the preserve of economic geographers. One of these concepts is the aggregate production function, which is also central to much of regional growth theory. However, the estimates obtained cannot be regarded as providing evidence of the underlying technological structure of the spatial economy, including the aggregate elasticity of substitution, the degree of returns to scale, or the rate of technical progress.

Économies d’agglomération, expansion régionale et la fonction de production cumulée : un Caveat Emptor pour scientifiques régionaux

RÉSUMÉ Depuis une vingtaine d’années, les économistes traditionnels font preuve d’un intérêt supérieur pour l’économie spatiale, et ont introduit des concepts et des outils en grande partie néoclassiques pour expliquer des phénomènes qui étaient, jusqu’alors, réservés aux spécialistes de l’économie géographique. Un de ces concepts est la fonction de production globale, qui est également au cœur même d’une grande partie de la théorie de croissance régionale. Toutefois, on ne peut considérer que les évaluations obtenues démontrent la structure technologique sous-jacente de l’économie spatiale, y compris l’élasticité globale de la substitution, le degré de rendement à l’échelle, ou le taux de progrès technique.

Economías de aglomeración, crecimiento regional y la función de producción agregada: un Caveat Emptor para científicos regionales

EXTRACTO Durante los últimos veinte años, los economistas establecidos han desarrollado mayor interés en la economía espacial y han introducido conceptos y herramientas económicos mayormente...
1. Introduction

The last 20 years or so have seen analytical tools of mainstream neoclassical economics being increasingly applied to analysing economic geography. There has been an ongoing lively debate between the ‘new’ geographical economists and the economic geographers about the most appropriate methodology for modelling the space economy (Martin, 1999; Garretson & Martin, 2010). Economic geographers, such as Storper (2011), while criticizing the assumption of general equilibrium inherent in the New Economic Geography, see the way forward as being a rapprochement between the economic geographers and the New Economic Geography. This is reflected, for example, in the 2006 symposium in the *Journal of Economic Geography*, reflecting ‘agglomeration and growth: a dialogue between economists and geographers’ (Duranton & Storper, 2006). Nevertheless, the regional political economy approach rejects the whole methodological stance of neoclassical economics from the perspective of Marx, Kalecki, and Sraffa (Sheppard & Barnes, 1990; Sheppard, 2000).

The purpose of this paper is not to discuss the deep methodological issues that still remain, but to present a warning about the uncritical acceptance of a central concept in neoclassical economics that is being widely used in spatial economics, namely, the aggregate production function. It is being used for diverse purposes ranging from estimating the size of agglomeration economies and spatial economic spillovers to determining the rate of regional productivity convergence using Solow’s growth model. For example, Cohen & Morrison Paul (2009, p. 106) in their survey of agglomeration, productivity, and regional growth argue that ‘recognizing and exploiting the potential of production theory models is thus an important step in the empirical literature on agglomeration economies’ (see also the surveys on agglomeration economies by Rosenthal & Strange (2004, 2006) and the World Bank (2009, Chapter 4)). The rate of regional productivity convergence is considered by Barro & Sala-i-Martin (1991, 2004, especially Chapter 11),
Armstrong (1995), Eckey et al. (2007), Maier & Tripl (2009), and Tselios (2009). The aggregate production function is central to all these applied studies and used widely in theoretical spatial modelling.

The production function is essentially a micro-economic concept and should be theoretically specified in terms of output and capital measured in homogeneous physical units. There is a substantial technical literature on the ‘aggregation problem’, which shows that the aggregate production function cannot be derived from micro-production functions, except under the most implausible assumptions (Fisher, 1992). Yet, paradoxically, statistical estimations of aggregate production functions give remarkably good fits with plausible estimates. The answer to this conundrum is that aggregate production functions have to be estimated using data for output and capital measured in constant-price monetary units. This is not an innocuous procedure because of a definitional relationship between output, capital, and labour via an underlying accounting identity. It is this that is responsible for the surprisingly good statistical fits of aggregate production functions. A serious consequence is that it is not possible to interpret the estimates of putative aggregate production functions as technological parameters, such as the elasticity of substitution, the degree of returns to scale (including agglomeration economies) or the rate of exogenous technical progress.

In this paper, we draw together and extend this criticism, showing the importance of the critique for spatial modelling with the intention of drawing the attention of economic geographers and regional scientists to its limitations. For reasons of space, we only sketch the outline of the critique; the reader is referred to out related articles for a discussion of the more complex associated issues, important though these may be.1

We shall demonstrate a paradox. The best statistical fit given by estimating putative regional aggregate production functions must give estimates of ‘constant returns to scale’ with the ‘output elasticities’ equal to their factor shares. This result has been taken as confirming (or rather not refuting) the neoclassical marginal productivity theory of factor pricing (Douglas (1976), inter alios). Regressions that find increasing returns to scale and any differences between the values of the output elasticities and the factor shares do so by virtue of being misspecified. This is irrespective of whether or not increasing returns to scale and agglomeration economies actually exist in reality. It also means that in many cases, with the help of some stylized facts such as a constant capital–output ratio and constant factor shares, which are not dependent upon the existence of an aggregate production function, it is possible to predict the regression estimates before running a single regression.

We reference some regional and urban papers that are subject to our critique. It is not possible, given space considerations, to undertake a comprehensive review of the many individual studies that potentially suffer from this problem. Following Solow (2000) in another context, we leave it to the reader to find Waldo2 (or Wally) in other applied studies that use the (regional) aggregate production function. Waldo in this case is the role of the underlying accounting identity in determining the empirical results. Moreover, we do not argue that it is not possible to test for, and find evidence of, agglomeration economies (see, for example, Delgado et al. (2010, 2011), who do not rely on a neoclassical production function), but that any such results that are derived from the neoclassical aggregate production function are problematic in the extreme.
2. Production Functions and External Economies to Scale: Some Theoretical Issues

One of the central tenets of spatial economics is the importance of, especially, agglomeration economies. These have been incorporated into the neoclassical production function in a number of ways. Most approaches usually commence with a firm’s micro-production function of the form:

$$Q_{ij} = A_{ij}F(K_{ij}, L_{ij})$$

(1)

where $i$ is the $i$th firm in industry $I$ which is located in city or region $j$. (We shall generally use the terms city and region interchangeably when discussing the production function.) $Q$ is the volume of output, $K$ is the capital stock and $L$ is employment. In terms of neoclassical production theory, $Q$ and $K$ are measured in homogeneous physical units.

The production function $F(\bullet)$ is assumed to exhibit constant returns to scale. For expositional ease, and because it generally gives a robust fit to the data, we shall assume that the production function is a Cobb–Douglas, but other more flexible forms such as the CES and the translog production function have also been used. (We discuss the translog below.) The use of other production functions does not affect the argument. Localization economies in region $j$ are sometimes assumed to be a function of total output $Q_{ij}$, of industry $I$ in city $j$. Consequently, this variable is made an argument of the production function:

$$Q_{ij} = A_{ij}K_{ij}^{a_{ij}}L_{ij}^{1-a_{ij}}Q_{ij}^c$$

(2)

where $0 \leq c < 1$.

As agglomeration economies are external to the firm, it remains possible to continue to assume that the individual firm’s production function is subject to constant returns to scale, and perfect competition. Consequently, wages and the rate of profit are assumed to be determined by the neoclassical marginal productivity theory of factor pricing. The marginal productivity conditions are controversial because they assume that wages and profits are determined solely by the technological conditions of production and the utility maximizing supply of labour function. While this does not mean that the distribution of income is necessarily optimal, it does follow that each factor is paid what it contributes to production, with all the consequent normative implications (Clark, 1891). Consequently, in both theoretical and applied modelling this assumption removes the need to consider bargaining models or sociological explanations of wage determination, which are not normally amenable to mathematical modelling within this framework.

As the data for individual firms are often unavailable, it is erroneously assumed (see the next section) that because there are constant returns to scale (see Nakamura, 1985, p. 122) Equation (2) may be summed across firms. After rearranging Equation (2), the following is obtained for the output of industry $I$ in city or region $j$:

$$Q_{ij} = A_{ij}K_{ij}^{a_{ij}}L_{ij}^{1-a_{ij}}Q_{ij}^c$$

(3)

where $1/(1-c)$ is the degree of returns to scale. (Anticipating the argument below, the best statistical fit, however, will always give constant returns to scale.)
We could specify more complex relationships where a firm’s productivity is affected by the output or productivity of firms in surrounding regions, subject to a distance decay function. These relationships can be picked by spatial econometric regression techniques.

Another approach is simply to assume that \( A_t = A_{t0} \exp(\Sigma Z) \) where \( Z \) is a vector of variables such as concentration/specialization ratios or an index of the diversity of the industry or total output in city \( j \) or its population, where subscript 0 denotes the base city (see Glaeser, 2000). The estimation often uses pooled data from the various industries. The absence of data for capital stocks at the urban level presents a problem. One way to overcome this is to simply assume the stylized fact that capital and output grow at roughly the same rate or that the output-capital ratio is constant. An alternative approach is to simply specify the production function in terms of labour only, which is the procedure that Glaeser et al. (1992) follow. For a more recent study along these lines, see Bishop & gripaios (2010).

This methodology is subject to a number of serious aggregation problems, as the focus of attention moves from the plant or firm to the industry. Let us make the perhaps unrealistic assumption that there exist well-defined micro-production functions, with output and capital measured in physical terms. The question arises, under what conditions can these functions be aggregated to give a ‘well-behaved’ aggregate production function? There is a large technical literature on this subject, but the conclusion is that the conditions are so stringent as to preclude any meaningful representation of an aggregate production function (Felipe & Fisher, 2003).

There are essentially two problems. The first arises at the firm (plant) level from the necessity to sum over the individual factors of production to give aggregate measures of output, capital and labour. The second is the necessity of aggregating firms’ (and industries’) production functions. Fisher (2005, pp. 489–490), who has done more work than most on these problems within a quintessentially neoclassical approach (see his collection of essays in Fisher (1992)), came to the conclusion that

except under constant returns, aggregate production functions are unlikely to exist at all. … Even under constant returns, the conditions for aggregation are so very stringent as to make the existence of aggregate production functions in real economies a non-event. … One cannot escape the force of these results by arguing that aggregate production functions are only approximations.

Nearly all the work on aggregation has been done on aggregating across firms or industries. To the previous two problems, we must add a third, or spatial aggregation, problem. This results from summing the individual firms’ output to give a total for the spatial unit (such as the state or city) under consideration (see McCombie & Roberts, 2007).

Given these theoretical difficulties, why is it that the aggregate production function continues to be used in both theoretical and applied work? The reason why the aggregation criticisms have been largely ignored by the economics profession is related to the instrumental methodology that pervades much of neoclassical economics. Ever since Friedman’s (1953) famous ‘Essay on positive economics’, the hallmark of an acceptable theory has been its predictive ability. The unrealism of the assumptions does not matter, as the hallmark of a good model is to ‘explain a lot from a little’. Ever since Cobb & Douglas’s (1928) first empirical study, estimates of the aggregate production function have generally given a close
statistical fit with the sum of estimated output elasticities not significantly different from unity and the individual elasticities close to the factor shares. We shall next explain why this is the case.

3. The Empirical Estimation of Regional Aggregate Production Functions: Too Good to be True?

In this section, we discuss some approaches to the estimation of region and city production functions. In empirical work, physical measures of $Q$ and $K$ are almost never available (except in engineering production functions) and so constant-price monetary values of $Q$ and $K$, the former henceforth denoted either by $Y$ (the value of gross output) or $V$ (value added) and the latter by $J$, must be used. The neoclassical approach assumes (erroneously) that all the theoretical results above follow through seamlessly to the aggregate level. This is illustrated by the frequent use of the term ‘volume of output’ for $Y$ and $V$. It is, of course, not a volume, but a constant-price value measure. In other words,

$$ P_t Y(0)_t = P_t \sum_i p_{io} Y_{it} $$

(4)

where $Y(0)_t$ is the monetary value of gross output at time $t$ measured at base year prices, $t=0$, $p_{io}$ are the individual prices of output $i$ in the base year $t$, and $Y_{it}$ is the physical quantity of $i$ at time $t$. $P_t$ is the price deflator, not the price, although the latter term is often misused in the literature. Value added is gross output minus the value of material inputs, i.e. their price ($m_{io}$) multiplied by the quantity $M_{it}$, or:

$$ P_t V(0)_t = P_t \sum_i (p_{io} Y_{it} - m_{io} M_{it}) $$

(5)

The picture is further complicated by the fact that for a good deal of the economy, most noticeably in the service sector, there is no measure of the output independent of the inputs. In this case, the constant-price value of output is often some function of the deflated wage bill, with an arbitrary allowance made for productivity. $J$ is usually calculated by the perpetual inventory method cumulating the constant-price depreciated value of investment. As we shall see, the fact that only constant-price value data and not physical data are available to the researcher becomes a crucial limitation of estimates of aggregate production functions. We define an aggregate production function as one where prices have been used to construct the measures of output and capital. In other words, it applies to not only the whole economy, but even, for example, to the four-digit SIC level.

The remarkable goodness of fit that the Cobb–Douglas provides to regional data is demonstrated from the following regressions results. Table 1 reports the ordinary least squares estimation of the Cobb–Douglas production function for the two-digit manufacturing industries for the US states for 1963 (as we are making a general point, the date is not important). Instrumental variable estimation did not make any significant difference to the results.

Two specifications of the Cobb–Douglas production function were estimated. The first used the aggregate industry values for the states. However, as it could be argued that these values are determined by arbitrary (non-economic) administrative areas, the Cobb–Douglas was also estimated using state industry per establishment data. As may be seen from Table 1, there was not a great deal of difference between
the results of these two specifications. The statistical fits were very close, with \( R^2 \)’s often over 0.950 (the lowest value is 0.854). The estimated output elasticities are close to their respective factor shares and constant returns to scale generally prevail. In fact, in only one industry (SIC 20, Food and Beverages) do both the total and the per establishment data show statistically significant increasing returns to scale. Moreover, given the close statistical fit of the regressions it can be seen that there is very little variation in the intercept (\( \ln A \)) for agglomeration economies to explain.

Similar cross-sectional results have been found in a number of other estimations using regional data including Hildebrand & Liu (1965) and Moroney (1972) for the USA and Griliches & Ringstad (1971), the latter using firm data for Norway. Moreover, Douglas’s numerous estimations in the early 1930s using cross-industry data found equally close statistical fits with the sum of the estimated coefficients not significantly different from zero and the coefficients very close to their respective factor shares.

Constant returns to scale are also found when the production function is estimated for total manufacturing, which could be (erroneously as it will be shown) interpreted as implying that there are no external economies of scale at this higher level of aggregation. If increasing returns were found for total manufacturing, but not at the individual industry, this would be evidence of agglomeration economies. For example, Garofalo & Yamarik (2002) also found, using panel-data estimation techniques for total US state manufacturing over the period 1947–1996, that the degree of returns to scale (1.008) was not significantly different from unity at the 95% confidence level: ‘The point estimates [of the output elasticities] furthermore, are close to the standard rule of thumb wherein labor is paid two-thirds of national income’ (p. 320). Results such as this have been taken as implying that the marginal productivity theory of factor pricing holds at the aggregate level (Douglas, 1976).

Two observations are in order here. The first is that, notwithstanding the likely complex relationships between inputs and outputs at the plant level and the severe aggregation problems, it is usually argued by neoclassical economists that the estimation of aggregate production functions, such as those reported here, does not refute the hypothesis that there are ‘laws of production’, which take a remarkably parsimonious form. Secondly, the results suggest the prevalence of constant returns to scale, which not all economists find implausible. (It is the standard assumption in much of neoclassical aspatial macroeconomics, notwithstanding the development of endogenous growth models (see, for example, Romer, 2006, p. 11; Hoover, 2012, Chapter 9, especially p. 330)).

### 4. On Identities and Behavioural Equations

Neoclassical production theory with the assumptions of homogeneous output and capital, both measured in physical units, constant returns to scale, perfectly competitive markets, no aggregation problems, and factors that are paid their marginal products will give the result that the total value of output will equal the sum of the remuneration of the factors of production. With the use of physical data, it is thus possible to estimate the parameters underlying the technology of the production function and to test, for example, the marginal productivity theory of factor pricing. This will not be refuted if the estimated output elasticities equal the factor shares.
However, this all amounts to putting the cart before the horse, as we shall show that with the use of constant-price value data, which nearly all estimations of production functions use, this result must always hold, whether or not the above assumptions are satisfied. This is because of the existence of the underlying identity

$$\ln V_i = \alpha + \ln y_i + \beta \ln L_i + \varepsilon_i$$

### Table 1. Estimates of a Cobb–Douglas production function, US manufacturing, state data, 1963

<table>
<thead>
<tr>
<th>Ind. code</th>
<th>$\alpha$</th>
<th>$t$</th>
<th>$\beta$</th>
<th>$t$</th>
<th>$\nu = \alpha + \beta$</th>
<th>$r^*$</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>20 (i)</td>
<td>0.600</td>
<td>6.32</td>
<td>0.450</td>
<td>4.53</td>
<td>1.050</td>
<td>3.11</td>
<td>0.990</td>
</tr>
<tr>
<td></td>
<td>0.661</td>
<td>7.30</td>
<td>0.537</td>
<td>5.08</td>
<td>1.198</td>
<td>3.55</td>
<td>0.921</td>
</tr>
<tr>
<td>22 (i)</td>
<td>0.038</td>
<td>0.15</td>
<td>0.923</td>
<td>3.68</td>
<td>0.651</td>
<td>1.32</td>
<td>0.983</td>
</tr>
<tr>
<td></td>
<td>0.010</td>
<td>0.05</td>
<td>0.875</td>
<td>3.85</td>
<td>0.885</td>
<td>2.38</td>
<td>0.946</td>
</tr>
<tr>
<td>23 (i)</td>
<td>0.372</td>
<td>4.96</td>
<td>0.625</td>
<td>7.24</td>
<td>0.997</td>
<td>0.11</td>
<td>0.982</td>
</tr>
<tr>
<td></td>
<td>0.312</td>
<td>3.60</td>
<td>0.619</td>
<td>8.66</td>
<td>0.931</td>
<td>1.17</td>
<td>0.918</td>
</tr>
<tr>
<td>24 (i)</td>
<td>0.576</td>
<td>5.91</td>
<td>0.345</td>
<td>2.87</td>
<td>0.921</td>
<td>2.03</td>
<td>0.969</td>
</tr>
<tr>
<td></td>
<td>0.444</td>
<td>4.60</td>
<td>0.683</td>
<td>4.24</td>
<td>1.127</td>
<td>1.27</td>
<td>0.878</td>
</tr>
<tr>
<td>25 (i)</td>
<td>0.195</td>
<td>2.59</td>
<td>0.842</td>
<td>10.65</td>
<td>1.037</td>
<td>1.36</td>
<td>0.981</td>
</tr>
<tr>
<td></td>
<td>0.189</td>
<td>2.12</td>
<td>0.808</td>
<td>10.25</td>
<td>0.997</td>
<td>0.05</td>
<td>0.921</td>
</tr>
<tr>
<td>26 (i)</td>
<td>0.499</td>
<td>9.97</td>
<td>0.535</td>
<td>10.18</td>
<td>1.034</td>
<td>1.82</td>
<td>0.990</td>
</tr>
<tr>
<td></td>
<td>0.492</td>
<td>7.72</td>
<td>0.512</td>
<td>5.19</td>
<td>1.004</td>
<td>0.08</td>
<td>0.970</td>
</tr>
<tr>
<td>27 (i)</td>
<td>0.511</td>
<td>3.56</td>
<td>0.523</td>
<td>3.36</td>
<td>1.034</td>
<td>1.77</td>
<td>0.993</td>
</tr>
<tr>
<td></td>
<td>0.579</td>
<td>4.42</td>
<td>0.512</td>
<td>3.07</td>
<td>1.091</td>
<td>1.45</td>
<td>0.936</td>
</tr>
<tr>
<td>28 (i)</td>
<td>0.223</td>
<td>2.70</td>
<td>0.794</td>
<td>9.14</td>
<td>1.017</td>
<td>0.93</td>
<td>0.974</td>
</tr>
<tr>
<td></td>
<td>0.217</td>
<td>2.52</td>
<td>0.806</td>
<td>8.40</td>
<td>1.023</td>
<td>0.63</td>
<td>0.962</td>
</tr>
<tr>
<td>29 (i)</td>
<td>0.626</td>
<td>5.66</td>
<td>0.282</td>
<td>2.05</td>
<td>0.908</td>
<td>2.20</td>
<td>0.985</td>
</tr>
<tr>
<td></td>
<td>0.619</td>
<td>4.93</td>
<td>0.256</td>
<td>1.41</td>
<td>0.875</td>
<td>1.65</td>
<td>0.961</td>
</tr>
<tr>
<td>30 (i)</td>
<td>0.620</td>
<td>4.64</td>
<td>0.425</td>
<td>3.23</td>
<td>1.045</td>
<td>1.26</td>
<td>0.974</td>
</tr>
<tr>
<td></td>
<td>0.590</td>
<td>4.45</td>
<td>0.474</td>
<td>3.27</td>
<td>1.064</td>
<td>1.06</td>
<td>0.937</td>
</tr>
<tr>
<td>31 (i)</td>
<td>0.224</td>
<td>1.43</td>
<td>0.774</td>
<td>4.55</td>
<td>0.998</td>
<td>0.06</td>
<td>0.899</td>
</tr>
<tr>
<td></td>
<td>0.225</td>
<td>1.82</td>
<td>0.667</td>
<td>4.12</td>
<td>0.892</td>
<td>1.32</td>
<td>0.960</td>
</tr>
<tr>
<td>32 (i)</td>
<td>0.304</td>
<td>3.52</td>
<td>0.703</td>
<td>8.78</td>
<td>1.067</td>
<td>0.28</td>
<td>0.983</td>
</tr>
<tr>
<td></td>
<td>0.260</td>
<td>2.86</td>
<td>0.697</td>
<td>8.74</td>
<td>0.957</td>
<td>0.73</td>
<td>0.909</td>
</tr>
<tr>
<td>33 (i)</td>
<td>0.408</td>
<td>6.79</td>
<td>0.576</td>
<td>8.67</td>
<td>0.984</td>
<td>0.94</td>
<td>0.993</td>
</tr>
<tr>
<td></td>
<td>0.409</td>
<td>5.39</td>
<td>0.578</td>
<td>5.49</td>
<td>0.987</td>
<td>0.33</td>
<td>0.980</td>
</tr>
<tr>
<td>34 (i)</td>
<td>0.363</td>
<td>3.60</td>
<td>0.637</td>
<td>6.24</td>
<td>1.000</td>
<td>0.07</td>
<td>0.993</td>
</tr>
<tr>
<td></td>
<td>0.356</td>
<td>3.47</td>
<td>0.628</td>
<td>5.99</td>
<td>0.984</td>
<td>0.32</td>
<td>0.910</td>
</tr>
<tr>
<td>35 (i)</td>
<td>0.200</td>
<td>1.69</td>
<td>0.828</td>
<td>6.66</td>
<td>1.028</td>
<td>1.74</td>
<td>0.992</td>
</tr>
<tr>
<td></td>
<td>0.263</td>
<td>2.19</td>
<td>0.731</td>
<td>5.36</td>
<td>0.994</td>
<td>0.13</td>
<td>0.925</td>
</tr>
<tr>
<td>36 (i)</td>
<td>0.337</td>
<td>2.93</td>
<td>0.708</td>
<td>6.00</td>
<td>1.045</td>
<td>2.02</td>
<td>0.986</td>
</tr>
<tr>
<td></td>
<td>0.334</td>
<td>2.70</td>
<td>0.704</td>
<td>4.76</td>
<td>1.038</td>
<td>0.62</td>
<td>0.908</td>
</tr>
<tr>
<td>37 (i)</td>
<td>0.225</td>
<td>1.49</td>
<td>0.831</td>
<td>4.48</td>
<td>1.056</td>
<td>1.06</td>
<td>0.952</td>
</tr>
<tr>
<td></td>
<td>0.242</td>
<td>1.88</td>
<td>0.892</td>
<td>5.31</td>
<td>1.134</td>
<td>1.78</td>
<td>0.899</td>
</tr>
<tr>
<td>38 (i)</td>
<td>0.015</td>
<td>0.07</td>
<td>1.091</td>
<td>4.14</td>
<td>1.106</td>
<td>2.03</td>
<td>0.977</td>
</tr>
<tr>
<td></td>
<td>0.192</td>
<td>0.79</td>
<td>0.913</td>
<td>2.80</td>
<td>1.105</td>
<td>0.76</td>
<td>0.854</td>
</tr>
</tbody>
</table>

Notes: SIC: 20 Food & Beverages; 22 Textiles; 23 Apparel; 24 Lumber; 25 Furniture; 26 Pulp & Paper; 27 Printing; 28 Chemicals; 29 Petroleum & Coal; 30 Rubber & Plastics; 31 Leather; 32 Stone Clay & Glass; 33 Primary Metals; 34 Fabricated Metals; 35 Non-Electrical Machinery; 36 Electrical Machinery; 37 Transportation Equipment; 38 Instruments.

(i) State total data; (ii) state per establishment data.

$t$ are the $t$-values, $r^*$ is based on the null hypothesis that the coefficient is unity.

Source: US Annual Survey of Manufactures, Area Statistics.

However, this all amounts to putting the cart before the horse, as we shall show that with the use of constant-price value data, which nearly all estimations of production functions use, this result must always hold, whether or not the above assumptions are satisfied. This is because of the existence of the underlying identity.
derivable from the product and income accounts. The following equation must always hold by virtue of being an identity:

\[ V_i = w_i L_i + r_i J_i \]  

(6)

where \( w \) is the wage rate and \( r \) the ex post rate of profit (which may include economic rents). \( J \) is the constant-price value of the capital stock. As Equation (6) is an identity, an increase in any one of the four variables on the right-hand side, holding the other three constant, must by definition lead to an increase in \( V \). There is no error term in Equation (6). For example, partially differentiating Equation (6) with respect to \( L \) gives \( \partial V_i / \partial L_i = w_i \), a result which led Phelps Brown (1957) to conclude that it was solely the accounting identity that generated the observed result that the marginal product of labour equals the wage rate. He also argued that this would occur, as Equation (6) is an accounting identity, whether or not the usual neoclassical assumptions held. Equation (6) holds for all states of competition and even when no true aggregate production function holds.

The implications are far more serious than merely this, as it may be shown that all that the estimations of putative production functions are accomplishing is a regression of a mathematical transformation of Equation (6) with no economic content. It will be shown that the surprisingly good statistical results of estimating the Cobb–Douglas production function are not surprising at all, but inevitable. There are a number of ways of demonstrating this proposition, but perhaps the simplest is to totally differentiate Equation (6) at any point in time to give:

\[ \frac{dV_i}{V_i} = a_i \frac{dJ_i}{J_i} + (1 - a_i) \frac{dw_i}{w_i} + a_i \frac{dL_i}{L_i} + (1 - a_i) \frac{dL_i}{L_i} \]  

(7)

Integrating Equation (7) gives the following result:

\[ V_i = a_i^{-a_i} (1 - a_i)^{-1} r_i w_i^{1-a_i} J_i^{a_i} L_i^{1-a_i} \]  

(8)

or, equivalently,

\[ V_i = A_i J_i^{a_i} L_i^{1-a_i} \]  

(9)

where \( A_i \equiv a_i^{-a_i} (1 - a_i)^{-1} r_i w_i^{1-a_i} \).

Equation (9) resembles a Cobb–Douglas form, but note that there is no error term as it is an identity. Equation (9) is exactly true, i.e. it is not an approximation of Equation (6) as, given we are dealing with infinitesimal changes, the factor shares are constant. Estimation of Equation (9) using cross-regional data, provided factor shares are roughly the same for all regions, will show that factor elasticities are close to the factor shares. Therefore, they will add up to unity, and the fit will be (almost) perfect. If, however, the factor shares vary, then the Cobb–Douglas production function will not necessarily give the best statistical fit. What is required is a more flexible functional form such as the CES (Felipe & McCombie, 2001) or the translog function (Felipe & McCombie, 2003). For example, the data, \( V, J, \) and \( L \) from the accounting identity could be used in estimating a Box–Cox transformation, which gives the linear relationship, the Cobb–Douglas or a functional form akin to the CES depending upon the variability of the factor shares. It is an error to assume that this critique only applies to the case where factor shares are constant and that, if they vary, somehow the estimates do refer to an actual aggregate production function.
The supposed aggregate production function (which is erroneously assumed to exist) is given in value terms by \[ V_i = A_i f(J_i, L_i) \]. Again, assuming a continuum of firms and differentiating, we obtain

\[
\frac{dV_i}{V_i} = \frac{dA_i}{A_i} + \alpha_i \frac{dJ_i}{J_i} + \beta_i \frac{dL_i}{L_i}
\]

(10)

where \( \alpha \) and \( \beta \) are the output elasticities.

Comparing this with Equation (7), which is definitionally true, the following relationships must always hold:

\[
\frac{dA_i}{A_i} \equiv a_i \frac{dr_i}{r_i} + (1 - a_i) \frac{dw_i}{w_i}
\]

(11.1)

\[
\alpha_i \equiv a_i
\]

(11.2)

\[
\beta_i \equiv (1 - a_i)
\]

(11.3)

Consequently, from the identity, it follows that

\[
\frac{dV_i}{V_i} \equiv \frac{dA_i}{A_i} + a_i \frac{dJ_i}{J_i} + (1 - a_i) \frac{dL_i}{L_i}
\]

(12)

which is the same result that would be obtained if all the usual neoclassical assumptions held, but, of course, we have not invoked them in deriving Equation (12). It can be simply shown that if there is a neoclassical aggregate production function and all the neoclassical assumptions are met, then from the ‘dual’ of the production function, Equation (11.1) will hold. Of course, what we have shown is that it must hold regardless of whether or not a well-behaved aggregate production function, or its dual, exists.

If we were able to use physical measures where \( Q, L, \) and \( K \) are not definitionally related, then we have a behavioural relationship. In this case, it is theoretically possible for changes in, say, \( K \) to have no significantly statistical effect on \( Q \), holding \( A \) and \( L \) constant, the effect being merely captured by the error term. Alternatively, it is theoretically possible to find the coefficients of both \( K \) and \( L \) statistically insignificant. Consequently, under these circumstances, the existence of a production function is capable of being refuted.

However, as we have seen above, this is not the case if we use value data. If we regard Equation (9) as relating to cross-section data for, say, regions or cities, then it will only hold exactly if the wage rate, the rate of profit, and the factor shares are constant. If they are not, the goodness of fit obtained by estimating the Cobb–Douglas will now depend upon the degree of the variation of the variables \( a, (1-a), w, \) and \( r \) across the firms or regions. For reasons that are discussed below, \( a \) and \( (1-a) \) tend to be relatively constant between firms in the same industry. Moreover, \( w \) and \( r \) in a regional context are unlikely to vary markedly over space, at least compared with the variation in \( J \) and \( L \). Thus, we are likely to get a very good fit to the Cobb–Douglas relationship. Simon (1979b) and McCombie (1998) show that even considerable variations in factor shares will still give a good fit to the data, with \( t \)-ratios large in comparison with many other cross-sectional regressions of behavioural relationships. Thus, in the light of this, the estimates of the regional production functions discussed above are hardly surprising.
Ideally, in the estimation of the Cobb–Douglas above, spatial econometric techniques should be used. Nevertheless, as all we are estimating are identities, we know what the estimated values of the coefficients should be, namely the values of the factor shares. There are no economic spillover effects between states as the model consists merely of an identity for each state. There is, however, an omitted variable resulting from the assumption in the model that \( A_i \) is spatially constant and this misspecification may lead to spatial autocorrelation. Consequently, any spatial autocorrelation will be due, especially, to those states with above average wages being located next to other states with above average wages, and vice versa. It seems generally that the weighted sum of logarithm of wage rate and the rate of profit is orthogonal to the other regressors, as the estimates of the output elasticities tend to be close to their factor shares. Consequently, the spatial error model could be used to estimate the Cobb–Douglas relationship (see Florax & Folmer (1992) and, for a general discussion, Corrado & Fingleton (2011)). But this econometric problem is very much a secondary issue.

The accounting identity critique also applies to the use of time-series data. The accounting identity can be expressed in growth rates (denoted by \(^\wedge\)) as:

\[
\hat{V}_t \equiv a_t \hat{r}_t + (1 - a_t) \hat{w}_t + a_t \hat{J}_t + (1 - a_t) \hat{L}_t
\]

(13)

Compare this with any neoclassical production function expressed in growth rate form using value data:

\[
\hat{V}_t \equiv \hat{A}_t + \hat{z}_t \hat{J}_t + \hat{\beta}_t \hat{L}_t
\]

(14)

A comparison of Equations (13) and (14) leads to the conclusion that \( \hat{A}_t \equiv a_t \hat{r}_t + (1 - a_t) \hat{w}_t \), \( \hat{z}_t \equiv a_t \) and \( \hat{\beta}_t \equiv (1 - a_t) \).

If there is a ‘true’ aggregate production function with an elasticity of substitution of unity (a Cobb–Douglas) and the usual neoclassical assumptions hold, then this will always give constant factor shares. But, as we have seen, it is extremely unlikely that the aggregate production function does exist. However, if for some other reason factor shares are constant, e.g. firms employ a constant mark-up pricing policy, this will ensure that the data give a good statistical fit to the Cobb–Douglas (see Fisher, 1971).

To illustrate how the critique applies to more general production functions than the Cobb–Douglas, consider the translog production function:

\[
\ln V_t = \epsilon + \lambda t + \gamma J \ln J_t + \gamma L \ln L_t + 0.5 \gamma_{JL} \ln J_t \ln L_t + 0.5 \gamma_{JL} (\ln J_t)^2 + 0.5 \gamma_{LL} (\ln L_t)^2
\]

(15)

The expressions for the output elasticities may be obtained by differentiating \( \ln V \) with respect to \( \ln J \) and \( \ln L \) respectively. If the assumptions of profit maximization and perfect competition are justified, the output elasticities of capital and labour will equal their respective shares.

\[
\frac{\partial \ln V}{\partial \ln J} = \gamma_J = \gamma_J + 0.5 \gamma_{JL} \ln L_t + \gamma_{JJ} \ln J_t = a_t
\]

(16)

\[
\frac{\partial \ln V}{\partial \ln L} = \gamma_L = \gamma_L + 0.5 \gamma_{JL} \ln J_t + \gamma_{LL} \ln L_t = (1 - a_t)
\]

(17)
Furthermore, differentiating Equation (15) with respect to time gives:

\[ \dot{V}_t = \lambda + \alpha \dot{J}_t + (1 - \alpha) \dot{L}_t \] (18)

which, using the marginal productivity conditions, can be expressed as:

\[ \dot{V}_t = \lambda + a \dot{J}_t + (1 - a) \dot{L}_t \] (19)

which is identical to Equation (14). Consequently, given the observed path of factor shares, the translog ‘production function’ will give a good approximation to the identity.5

The heart of the problem is the fact that the production function is a technological relationship, but is estimated using constant-price monetary or value data. Estimations of aggregate regional costs functions (which in neoclassical production theory are derived as the ‘dual’ of the aggregate production function) and labour demand functions are also subject to this critique (see Felipe & McCombie (2009), who also consider the related important critique by Anyadike-Danes & Godley (1989)).

4.1. Estimates of ‘Regional Production Functions’ and ‘Elasticities of Substitution’

Nevertheless, there have been numerous estimations of aggregate regional production functions, most notably the CES, that have attempted to provide an estimate of the regional elasticity of substitution. The relevance of this is that these estimates have an important implication for the efficacy of regional capital and/or labour subsidies (Armstrong & Taylor, 2000, pp. 241–246).6 Much ink has been spilt in refining the estimation techniques using different specifications. Early estimates, such as those of Tooze (1968), Buck & Atkins (1976), and O’Donnell & Swales (1979), used the marginal productivity of labour side relation in the absence of the capital stock, whereas later studies used explicit measures of the capital stock (Harris, 1991). The problems that aggregation pose for the CES production function have been examined by the simulations of Fisher et al. (1977). (Ironically, one of the authors was Solow.) They postulated true underlying micro CES production functions, but where aggregation problems meant that no true aggregate production function existed. The simulations generated estimates of the elasticity of substitution that were statistically significant. Nevertheless, as Fisher et al. point out, ‘the elasticity of substitution in these production functions is an “estimate” of nothing; there is no “true” aggregate parameter to which it corresponds’ (Fisher et al., 1977, p. 312). Indeed, there was a case where the estimated aggregate elasticity of substitution lay outside of the range of the individual micro-elasticity of substitutions.

Another representative example is Graham (2000) who used an aggregate production function to estimate the factors that affected the productivity levels of the UK counties. This further illustrates the generic problems posed by the accounting identity. One specification he estimated was the Cobb–Douglas in the form:

\[ \ln(V_{it}/L_{it}) = \ln A_i + \alpha \ln(J_{it}/L_{it}) + (\alpha + \beta - 1) \ln L_{it} + \ln h(s_i) \] (20)

where, if the coefficient of \( \ln L_{it} > 0 \), there are increasing returns to scale. \( \ln h(s_i) \) is a vector of variables supposedly capturing spatial and other externalities. From the
identity, given by Equation (13), we know that if factor shares are constant the following specification will give a perfect fit to the data:

\[
\ln \left( \frac{V_t}{L_t} \right) = c + (1 - a) \ln w_t + a \ln r_t + a \ln \left( \frac{J_t}{L_t} \right) + (a + (1 - a) - 1) \ln L_t
\]

(21)

The coefficient of \(\ln L\) will equal zero. Of course, the term \((1 - a) \ln w + a \ln r\), which is assumed to be spatially invariant in Equation (20), is likely to show some variation at the county level. However, provided the term is orthogonal with the other regressors, it is not likely to bias the values of the other coefficients.

For manufacturing, the share of capital is about 0.4 and that of labour around 0.6 (for the whole economy the former figure is generally between 0.25 and 0.3). Estimating Equation (21) for each of the two years 1984 and 1991 gives, not surprisingly, a very close statistical fit with the two \(R^2\)s of over 90%. The coefficients of \(\ln \left( \frac{J_t}{L_t} \right)\) are 0.426 (\(t\)-value of 17) and 0.489 (\(t\)-value of 21) for 1984 and 1991, respectively, and those of \(\ln L\) were only statistically significant for 1984, but, even so, the magnitude was small (−0.075, \(t\)-value of −2.6).

In Equation (20), \(\ln h(s)\) included variables as a skills proxy, the proportion of part-time workers and a dummy for London, all of which were statistically significant. How is this to be explained? If we were estimating the identity, these would be merely irrelevant regressors. However, the fact that some of them are statistically significant is because, especially, \(\ln w\) shows some variation across counties. Consequently, it is not surprising that these ‘externality’ variables are significant in the regression as they are likely to be correlated with the logarithm of the wage rate. However, they cannot be interpreted as externality effects within a production function framework.

Graham (2000) also repeated the exercise using the linear approximation to the CES production function (which he calls the translog). Not surprisingly, the statistical fit is better (the \(R^2\)s are 0.96 and 0.95) as this functional form is more flexible than the Cobb–Douglas in that it allows for inter-county variations in factor shares. However, not surprisingly, the elasticities of capital at the mean are 0.413 and 0.453, again merely reflecting the value of the capital’s share.

4.2. The Regional Growth Accounting Approach

In a simulation exercise, Felipe & McCombie (2006) also found that the hypothetical ‘true’ rate of growth of technical change or the growth of total factor productivity differed from the rate calculated using value data. In fact, the growth of total factor productivity calculated using value data is, by definition, the weighted growth of the real wage rate and the rate of profit (see Equations (13) and (14)) (Denison, 1967). This has obvious implications for the regional estimates of total factor productivity growth and the regional growth accounting approach.

The growth accounting approach is based on the neoclassical production function and the neoclassical theory of factor pricing. This allows the output elasticities to be measured by the factor shares and the growth of total factor productivity is given by:

\[
T\ddot{FP}_t \equiv \dot{V}_t - \alpha_t \ddot{J}_t - (1 - \alpha_t) \ddot{L}_t \equiv \dot{V}_t - \dddot{a}_t \dddot{J}_t - (1 - \dddot{a}_t) \dddot{L}_t
\]

(22)
Hulten & Schwab (1984) used the growth accounting approach to calculate the growth of total factor productivity for the US regions. They found that output grew twice as fast in what they termed the Sun Belt states compared with the older Snow Belt regions. This could not be explained by differences in the growth of total factor productivity, but all this tells us is that the growth of the real wage rate and the rate of profit, each weighted by its factor share, did not greatly differ between the regions, a result hardly surprising given the relative homogeneity of states compared with nations. Using their data, the growth of total factor productivity (which is sometimes misleadingly interpreted as technical progress) accounts for 75% of productivity growth in the Snow Belt and 68% in the Sun Belt, leaving relatively little to be explained by the growth in the capital–labour ratio. This is similar to the magnitudes found by Solow (1957) for the US private sector from 1909 to 1948.

However, if the rate of profit is roughly constant over time, then the measured growth of total factor productivity will definitionally be equal to $\left(1 - a\right)\dot{w}$ and, if factor shares are roughly constant, this will equal $\left(1 - a\right)\left(V - \dot{L}\right)$. Consequently, total factor productivity growth will equal 60–75% of productivity growth, depending upon labour’s share—a result Solow (1988), for some reason, found ‘startling’.

Other examples include Moomaw & Williams (1991), who calculated total factor productivity for the 48 US states for 1954–1976. Harris & Trainor (1997) estimated total factor productivity for regional manufacturing for the UK regions over the period 1968–1991. A twist in their procedure was to use Hall’s (1988) method to estimate the mark-up (price-cost margin or $\mu$) due to imperfect product markets and to use $\mu\left(1 - a\right)$ as labour’s true output elasticity where the estimates of $\mu$ are generally greater than unity. However, without going into details, Hall’s model is also a misspecification of the accounting identity, where the misspecification biases the estimate of $\mu$ upwards from the ‘true’ value of unity (Felipe & McCombie, 2002). The underlying problems caused by the identity remain.


5. Regional Convergence and the Aggregate Production Function

As an example of the misleading interpretation that the use of the aggregate production function can entail, we next consider the Solow growth model and its use in estimating the speed of absolute and conditional ‘beta’ regional convergence (see Mankiw et al., 1992). The popularity of this model with some regional economists is surprising given the underlying assumptions, which include perfectly competitive markets, constant returns to scale, a common technology, and a rate of exogenous technological progress that is the same for all regions (i.e. technological progress is a pure public good). These would seem to be most implausible both in a regional and an international context. There are now numerous studies that have estimated absolute and conditional beta convergence for the regions, with the more recent ones mainly concentrating on conditional beta convergence (see, for example, Barro & Sala-i-Martin, 1991, 2004; Armstrong, 1995; Evans & Karras, 1996; Rey & Montouri, 1997; Garofalo & Yamarik, 2002; Badinger et al., 2004; Egger & Pfaffermay, 2006; Benos &

The steady-state productivity growth rate in the Solow growth model is equal to the rate of technical change. The reason why some regions have faster growth than others is that their capital–labour ratios are below the steady-state level. There is thus a transitional disequilibrium component to their growth rate as these regions catch up and converge with their steady-state growth rates. Conversely, regions where the capital–labour ratio is above their steady-state value will have temporarily slower growth than the steady-state rate. (A major problem with this analysis is that the Solow model predicts convergence, but for this to have occurred there needs to have been divergent growth in the first place. Indeed, world growth since the industrial revolution has been, as Pritchett (1997) puts it, ‘divergence, big time’. Apart from resorting to an explanation in terms of substantial ad hoc shocks, the theory is silent on this point.)

Testing this is done by log-linearizing the aggregate production function around its steady-state growth path to give:

\[ \frac{1}{\tau} \ln \left( \frac{v_{it}}{v_{i0}} \right) = \epsilon - \left( \frac{1 - e^{-\eta \tau}}{\tau} \right) \ln v_{i0} \]  

(23)

where \( v_{it} \) is the level of labour productivity in region \( i \) at time \( t \), and \( \eta \) is the speed of convergence. \( \tau \) is the length of period over which the growth rate is calculated, i.e. \( \tau = t - 0 \) where 0 denotes the initial year. For many regional data sets, the estimate of \( \eta \) is statistically significant and, consequently, the rate of convergence takes a value of around 2% per annum (Barro & Sala-i-Martin, 2004).

The problem with this specification, remaining within the neoclassical framework, is that it ignores regional variations in the investment rate, which, while not affecting the common steady-state rate of productivity growth, will affect the steady-state level of productivity. In a seminal paper Mankiw et al. (1992) overcame this shortcoming. Starting with a conventional Cobb–Douglas production function of the form \( Q = (A(t) L)^{(1-\alpha)} K^\alpha \), Mankiw, Romer, and Weil (MRW) showed that the steady-state Solow growth model can be expressed in level form as:

\[ \ln v_{it} = \ln A_0 + \lambda t + \frac{\alpha}{1 - \alpha} \ln s_{Kit} - \frac{\alpha}{1 - \alpha} (n_{it} + \delta + \lambda) \]  

(24)

\( A_0 \) is again the level of total factor productivity, which is assumed to be the same for all countries (and \( \ln A_i = \ln A_0 + \lambda t \)), \( s_K \) is the ratio of investment in physical (as opposed to human) capital to total output, \( n \) is the growth of employment (which is also sometimes assumed to be constant across regions), \( \delta \) is the rate of depreciation and \( \lambda \) is the common rate of technical progress. The parameters \( \alpha \) and \( 1 - \alpha \) are the output elasticities of physical capital and labour. It is assumed that \( \ln A_i = \ln A_0 + \lambda t \) can be proxied by a constant when cross-country (or cross-regional) data are used. In other words, both the level and rate of growth of technology are assumed to be the same for all the regions.

This equation was estimated by MRW using the Penn-World Tables and cross-country value data for advanced and developing countries. While the results were reasonable, except for the OECD countries, they found that the estimate of \( \alpha \) took a value of about two-thirds, higher than its factor share of approximately one-third.
This led them to augment the Solow growth model by introducing the logarithm of the share of human capital in total output as an explanatory variable. The estimate of the coefficient of physical capital now became about one-third, roughly equal to its factor share.

However, if the stylized facts that factor shares and the capital–output ratio are constant hold, then, solely from the accounting identity, we can derive the following equation (see Felipe & McCombie, 2005b):

\[
\ln v_t \equiv c + 1.0 \ln w_t + \frac{a}{1 - a} \ln r_t - \frac{a}{1 - a} \ln \left( n_t + \delta + \frac{(1 - a) \dot{w}_t + a \dot{r}_t}{1 - a} \right)
\]

(25)

It can be seen that this is identical to Equation (24) where, from Equations (25) and (24):

\[
\ln A_i = \ln A_i + \lambda t = c + \ln w_{i0} + \frac{a}{1 - a} \ln r_{i0} + \dot{w}_t t + \frac{a}{1 - a} \dot{r}_t t
\]

(26)

if the wage rate and the rate of profit grow at a constant rate over time. Note that empirically the initial level of the (log) of the wage rate varies considerably across countries, whereas \( \ln r_{i0} \) shows little systematic variation. We can see from Equation (25) that assuming that \( \ln A_i = \ln A_{i0} + \lambda t \) is constant across countries/regions will cause omitted variable bias if one estimates Equation (26). This is because, as we have noted, \( \ln w_t \) varies significantly across countries (\( \lambda \) and the growth rates of the wage rate and the rate of profit will also differ between countries). It is solely for this reason that MRW failed to find a perfect statistical fit to the underlying accounting idea and were compelled to introduce human capital.

Support for our argument is provided by Easterly & Levine (2001) who generalized the MRW model by using dummy variables to allow \( \ln A \) to differ between the major regions of the world. They found that

South Asia and Sub-Saharan Africa have significantly lower productivity than the other regions (income differences that are not explained by the MRW term). The OECD has higher productivity than the rest of the world by a factor of 3… Once the productivity level is allowed to vary, the coefficient on MRW implies a capital share of 0.31—which is in line with most estimates from national income accounting. (Easterly & Levine, 2001, p. 190)

This is without including the human capital variable. Including the human capital variable in Equation (24) brings the estimate of \( \alpha \) closer to \( a \) because it is closely correlated with \( \ln w \) and hence brings the estimating equation closer to the identity.

But we can see immediately why MRW get this result. Because of the underlying accounting identity, the result is inevitable as the dummies are closely proxying the variation of the \( \ln w_{i0} \) term in the identity (or \( \ln w_t \) as the two will be highly correlated), bringing the estimated regression coefficients closer to those of the identity. If we were also to allow the growth rates of ‘total factor productivity’ (\( \lambda \), or \( \dot{w} + a/(1 - a) \dot{r} \)) to vary between countries (as the neoclassical growth accounting approach shows should be the case) rather than imposing a constant growth rate, we would end up merely estimating the identity. (MRW’s assumptions are inconsistent with the standard neoclassical growth accounting results on this point, and \( \text{vice versa} \).)
The rate of conditional beta convergence is estimated by subtracting the logarithm of initial level of productivity from both the right-hand and the left-hand side of Equation (25). It can been seen that if Equation (26) is used (which under a neoclassical interpretation implies that countries have different levels and growth rates of total factor productivity) then the coefficient on ln $v_0$ will equal minus unity. This implies an infinitely fast rate of convergence. The fact that ln $w$ is highly correlated with ln $v$ means that excluding the former will bias the absolute value of the coefficient of ln $v$ downwards, giving the estimates generally found empirically. Nevertheless, the MRW model has been used at the regional level.

It is interesting that when Lee et al. (1998) allow $\lambda$ to vary the estimate of $\eta$ increases from 0.04 to 0.29. As Islam (1998, p. 326) concedes, ‘when, in addition, heterogeneity in growth rates is allowed convergence becomes, in essence, an empty concept’. If Lee et al. (1998) had allowed ln $A$ to vary as well, then the estimated speed of convergence would have tended towards infinity. Convergence, within MRW’s framework of the augmented Solow growth model, is indeed an empty concept.

The fact that there is an inverse correlation between the growth of regional productivity and the initial level of productivity does show regional per capita GDP levels are converging, but there could be many reasons for this. It may be because the regions with lower levels of productivity are technologically more backward and are benefiting from the diffusion of innovations from the more advanced regions. There are strong political pressures to reduce income per capita disparities, especially in the advanced countries. This is accomplished not only through fiscal transfers, but also through the spatial pattern of government expenditure and regional subsidies, either as the result of an explicit or implicit regional policy. In the enlarged EU many of the low productivity regions are those that depended heavily on agriculture, and with the relative decline of this sector and the transfer of labour to the higher productivity industrial sector, regional productivity can be expected to rise in the less developed regions. Thus, the estimation of either absolute or conditional beta convergence is not a test of the (augmented) Solow model.

The endogenous growth models, to the extent that they use the aggregate production function, also suffer from the problem of the accounting identity. Take, for example, the ‘linear-in-capital’ or $Q=AK$ model (expressed here in physical terms), where the exponent of capital is unity and captures the effect of both the direct contribution of capital and the effect of induced technical change. $A$ is now a constant. Let us assume the Kaldorian stylized facts of constant factor shares, no trend in the rate of profit, and no discernible growth in the output-capital ratio (see Simon (1986, Appendix A, pp. 172–183) for an explanation of why value data will give a constant capital output ratio this irrespective of the ratio between physical output and capital). Under these circumstances the accounting identity, Equation (13), may be written as:

$$\dot{V}_t \equiv (1-a)\dot{w}_t + (1-a)\dot{L}_a + a\dot{J}_t \equiv (1-a)(\dot{V}_t - \dot{L}_a) + (1-a)\dot{L}_a + a\dot{J}_t \equiv \dot{J}_t$$  (27)

Consequently, integrating this equation gives $V_t=AJ_t$ where $A$ is the constant of integration and does not vary over time.
A second generation of endogenous growth models allows $A$ to change over, which is interpreted as the ‘stock of knowledge’ or ‘ideas’. Consequently, $A$ is produced by a separate R&D production function of the form:

$$\Delta A_t = \delta L_{R&D}^\sigma A_t^\rho$$

(28)

where $L_{R&D}$ is the number of scientists and $0 < \sigma < 1$ and $\phi < 1$ are parameters of the R&D production function (see Martin & Sunley (1998), Armstrong & Taylor (2000), and Roberts & Setterfield (2010) for discussions of endogenous growth in a regional context). But, as Steedman (2003) points out, this requires $A$ to be a cardinal measure and no explanation has been put forward to show that it is. Consequently, the ‘analysis of the [New Growth Theory] equations having no meaning cannot yield convincing conclusions’ (p. 133).

Izushi’s (2008) attempt to test various specifications of Equation (28) using data for the NUTS1 regions of the European Union clearly demonstrates this problem. $\dot{A}$ was calculated as the growth of total factor productivity, which, as we have seen, is equal to $(1 - a)\dot{w}$, given the assumption that there is no rate of change in the rate of profit. The only statistically significant relationship found by Izushi, and which ‘shows strong support for the … Lucasian framework’ (Izushi, 2008, p. 955), is the one where regressing the growth of total factor productivity on private sector R&D workers gives a slope coefficient of 0.25. This value is most implausible. It implies increasing the growth rate of R&D workers by, say, 10 percentage points per annum will increase the growth of real wages in the EU regions by over 2.5 percentage points per annum. Such an increase in the growth rate of R&D workers should not be difficult, given the supposed economic benefits, as they only consist of about 1% of total EU employment. (The results also suggest that, from the estimated coefficient of the spatial-lag term, a one percentage point increase in the growth of the wage rate of the surrounding regions will implausibly increase the region’s growth rate by 0.6 percentage points. Increasing the share of R&D personnel in total employment also had a further positive impact on the growth of the real wage.) Clearly, to test the endogenous growth model satisfactorily requires an independent measure of the stock of knowledge.

6. A Way Forward?

The above discussion is somewhat nihilistic and is what Lawson (2004) claims is an exercise in under-labouring; or as Locke (1690) put it: ‘removing some of the rubbish that lies in the way of knowledge’ (cited by Lawson, 2004, p. 317). But, as Kuhn (1970) pointed out, one paradigm, no matter how (logically) flawed, is only abandoned if it is replaced by another paradigm. It is beyond the scope of this paper to discuss alternative paradigms and we merely make a few suggestions as to possible ways forward.

At the aggregate level, a more fruitful way of analysing regional growth processes is the post-Keynesian demand-oriented export-led growth approach which emphasizes not only increasing returns and agglomeration economies but the role of the structure of production (captured through differences in income elasticities of demand for a region’s exports). It is flexible enough to capture the importance of monetary factors, interregional capital flows, and sovereign debt problems that are currently threatening the existence of the euro zone and on which the neoclassical approach is noticeably silent (Rowthorn, 2010; McCombie,
One central tenet of this approach is the Verdoorn law (see McCombie et al., 2002), a linear specification of Kaldor’s technical progress function. It has similarities to the production function although Kaldor eschews the marginal productivity theory of factor pricing and the distinction between movements along and shifts of the production function. What then are the implications of the critique for the Verdoorn law given the need to use value data here as well? There are a number of specifications of the Verdoorn law, one of which explicitly includes the growth of capital and takes the form

\[ TFP_t = \lambda + \psi \hat{V}_t \]  

(29)

where \( TFP \) is the growth of total factor productivity, \( \lambda \) the exogenous growth of total factor productivity, \( \hat{V} \) the growth of output and \( \psi \) is the Verdoorn coefficient (\( 1 > \psi > 0 \) has been interpreted as encompassing increasing returns to scale). More elaborate specifications include other variables, such as the density of production, and the model is now routinely estimated using spatial econometric methods. What are the implications of the accounting identity critique for the Verdoorn law? The Verdoorn law does have behavioural implications, notwithstanding the accounting identity critique applies. From the accounting identity in growth rate form (Equation (13)), Equation (29) may be alternatively expressed as:

\[ TFP_t \equiv a \hat{r}_t + (1 - a)\hat{w}_t = \lambda + \psi \hat{V}_t \]  

(30)

or

\[ (a \hat{r}_t + (1 - a)\hat{w}_t) = \frac{\lambda}{\psi} + \frac{\psi}{(1 - \psi)} (a \hat{f}_t + (1 - a)\hat{L}_t) \]  

(31)

where \( \lambda \) is a constant.

Consequently, the Verdoorn law does, in fact, demonstrate an empirical relationship, even allowing for the accounting identity. That is to say, a faster growth of the region’s weighted factor inputs will increase the weighted growth rate of the factor prices.\(^{10}\) It is, however, best regarded as another Kaldorian stylized fact.

The importance of agglomeration economies and clusters is being extensively studied by Michal Porter’s cluster mapping project which uses statistical analysis to determine the extent of clusters. While the concept of the cluster used by Porter is not without its critics (Martin & Sunley, 2003), there is no doubt that it is providing useful insights into regional growth. The approach, with its emphasis on clusters determining the export performance of the region (Porter, 2003, p. 572, fn 13), complements the post-Keynesian approach noted above (McCombie, 2012).\(^{11}\) Porter (2003) presents some statistical estimates of the importance of clusters without using a production function. A more recent study by Delgado et al. (2011) also confirms the importance of clusters. It is found that the growth of employment of four-digit SIC clusters is determined by the degree of regional specialization in the industry, the relative strength of the cluster environment surrounding that industry, and a measure of cluster strength of other industries used in the cluster. Delgado et al. (2010) also demonstrate statistically the importance of clusters in the start-up of new firms which reflects the degree of entrepreneurship.

There is also a considerable literature explaining regional growth disparities through regional innovation systems, industrial districts and regional inter-linkages, which in many cases use case studies and do not rely on regional production
functions. There is not space to discuss these approaches here but surveys can be found in Clark et al. (2000), Part IV, ‘The geography of learning’; see also, inter alia, the articles in Dunning (2000) and Breschi & Malerba (2005).

7. Conclusions

The aggregate production function has had a chequered history ever since it was first introduced by Cobb & Douglas (1928). In particular, it is now well-established theoretically that micro Cobb–Douglas production functions cannot be summed to give an aggregate production function, except under most implausible assumptions. These reservations (together with those of the capital controversies) were discussed in most textbooks on economic growth prior to around 1975 and then were conveniently forgotten—there is no mention of them in growth textbooks, such as Barro & Sala-i-Martin (2004) (see the discussion in McCombie (2011)). The standard instrumentalist defence that aggregate production functions ‘work’, in that they empirically give close statistical fits with plausible estimates, is unsound. These estimations of the aggregate production function have to use constant-price monetary data for output and capital and underlying accounting is responsible for the goodness of fit. The only case where this problem does not occur is when physical, or engineering data, are used and such studies, which almost certainly must be at the plant level, are few and far between.

The aggregate production function has now become widely used in spatial economics, for example, to estimate the degree of agglomeration economies, to calculate the rate of regional technical progress, to model regional economic growth and the rate of convergence or divergence, and to estimate the elasticity of substitution. It also has important policy implications. For example, the elasticity of substitution of the regional production function has been used to estimate the effect of regional capital and/or labour subsidies, but such calculations are fatally flawed, as the aggregate elasticity of substitution does not exist.

This paper serves as a warning to the continued uncritical use of the aggregate production function in economic geography (or, indeed, in macroeconomics, per se).

Notes

1. See Felipe & McCombie (2005a, 2012a). Simon & Levy (1963) first applied the critique to cross-sectional data. Shaikh (1974, 1980) generalized the criticism to time-series estimations of supposed production functions and Simon (1979a) thought the problem so important he mentioned it in his Nobel prize speech (see also Simon, 1979b). However, the problems have generally been ignored by economists who use the production function, with the exception of Temple who misunderstands the issues (see Felipe & McCombie (2012b) and the references cited therein).

2. ‘Finding Waldo’ is a children’s game where the character Waldo has to be found where he is hidden in a coloured picture with many other individuals. Solow (2000) uses this as an analogy for the ad hoc linear restriction necessary in many endogenous growth models to ensure a steady-state solution, but which is often merely implicit rather than explicitly stated.

3. Ideally, the production function should include land as a factor production. However, this is often excluded because of data availability and we shall abstract from this factor of production here.

4. It is sometimes argued that the critique does not hold because neoclassical production theory assumes perfect competition (i.e. there are no economic rents). This is erroneous for the reasons set out in Felipe & McCombie (2007) and, for reasons of space, will not be discussed here.

5. Times-series estimation of production functions sometimes give a poor fit to the data with the estimates of the output elasticities taking often implausible values. This is because proxying the weighted logarithm of the wage rate and the profit rate by a linear time trend (or their weighted growth rate by a constant) is misspecified

as the wage rate and, especially, the rate of profit has a pronounced cyclical fluctuation. It is always possible to find a more flexible proxy that makes the estimating equation equivalent to the accounting identity.

6. Although, as O’Donnell & Swales (1979) and Armstrong & Taylor (2000) point out, other factors such as the supply elasticities of capital and labour are also important.

7. Note that, as we have seen above, the accounting identity critique does not depend upon these stylized facts. If they do not hold, then we know immediately from the accounting identity that MRW’s specification will give a poor statistical fit.

8. These are OECD, East Asia, South Asia, Sub-Saharan Africa, Western Hemisphere, Middle East and North Africa, and Europe.

9. The growth of total factor productivity was adjusted for the rate of change in human capital. Generally, this does not make a great deal of difference as the latter is proxied by a function of the growth of the change in ratio of the wage rates of the skilled and unskilled workers, which is generally relatively small, and could even be negative (Solow, 1988, p. 313).

10. Estimating Equation (31) directly generally finds the implicit estimate of $\psi$ not significantly different from zero, while estimating Equation (29) gives $\nu$ approximately equal to one half. This is due to econometric reasons and will not be discussed here. Suffice it to say, Equation (29) is generally seen as the preferred specification. It should be noted that the use of the factor shares as weights is, in the light of the accounting critique, somewhat of an ad hoc assumption. The neoclassical interpretation of them as reflecting output elasticities is invalid.


References


