Decomposing the Great Stagnation: Baumol's cost disease vs. "ideas are getting harder to find"*

Basil Halperin
Stanford

J. Zachary Mazlish
Oxford and GPI

November 2024
Preliminary

Abstract

We decompose the post-1973 productivity growth slowdown into three causes: structural change (Baumol's cost disease), input misallocation, and pure productivity effects. We do this by constructing sector-level productivity from 1947 to 2016, using the recent BEA-BLS Integrated Level Production Accounts (Eldridge et al. 2020) adjusted with our own industry-specific markup estimates. We find that Baumol's cost disease explains ∼25% of the productivity slowdown. The magnitude of the input misallocation channel is sensitive to methodology, reflecting uncertainty about whether or not aggregate markups have risen: input misallocation can account for between 0-20% of the productivity slowdown. Finally, we also show that linear growth fits the US data better than exponential growth, though under either exponential or linear growth there has been a post-1973 productivity slowdown.

* *Email:* `basilh@stanford.edu, john.mazlish@economics.ox.ac.uk`
Thanks to XXX

1 Introduction

US productivity growth has slowed substantially since the early 1970's: depending on the exact data used, annual total factor productivity growth was around 0.9-1.2pp higher from 1947-1973 then it has been since.

One prominent explanation for why productivity growth might "naturally" slow down as an economy matures is Baumol's cost disease, originated in Baumol (1967). The idea is simple: some sectors (i.e. manufacturing) are intrinsically easier to automate, and thus higher productivity growth sectors. Other sectors (i.e. services) are intrinsically hard to automate, and thus low in productivity growth. Therefore, costs in the service sector rise over time. Assuming people continue to want to consume services as an economy grows, the service sector will become an ever-larger share of economy-wide output, and aggregate productivity growth will be increasingly determined by the slow rate of productivity growth in the services sector.

How much of the slowdown in US productivity growth does Baumol's cost disease explain? In this paper, we show that across a few different ways of calculating TFP, the cost-disease is (annually) responsible for 0.24-0.27pp of the post-1973 slowdown, or 23-27%.

Our results are very similar to the results in Duernecker, Herrendorf, and Valentinyi (2023), which also provides a model that is used to predict how much Baumol's cost-disease will curb *future* US productivity growth. This piece is intended as a short note to complement the more extensive treatment found in Duernecker, Herrendorf, and Valentinyi (2023). Relative to their paper, our contribution is fourfold: 1) we adjust industry-level productivity data for the presence of profits. 2) we use data that is consistent with the US national accounts. 3) we provide a novel decomposition of TFP growth into "pure productivity", "Baumol", "capital allocation", and "labor allocation" terms. 4) we show that using our profit-adjusted data, post-war US productivity growth is clearly not exponential and may even be sub-linear.

We now briefly expand on each of those points before presenting the results.

Profits. Construction of the national accounts data relies on the assumption of zero profits, in order to impute capital compensation as the residual of value-added minus labor compensation. Recently, increasing attention has been given to the possibility that average profits are quite high in US industries and have been rising over time (e.g. De Loecker, Eeckhout, and Unger 2020). Concomitantly, Baqaee and Farhi (2020) have provided important theoretical results which show how to correctly measure productivity in inefficient economies.

Both Vollrath (2024) and Comin et al. (2020) make use of the results in Baqaee and Farhi (2020) to adjust US industry-level accounts data for the presence of profits. We conduct a very similar exercise here, using a slightly different source of US accounts data.

Data. The data we use are from the BEA-BLS Integrated Level Production Account (ILPA), which was recently extended to cover the period 1947-2016. The advantage of the ILPA is that it is internally consistent with the official BEA industry-level national accounts, which are typically seen as the gold standard for measuring economy-wide US output growth. Relative to the other papers just cited, Duernecker, Herrendorf, and Valentinyi (2023) use World KLEMS data, while both Vollrath (2024) and Comin et al. (2020) use BLS private business sector productivity data. There is nothing wrong with these alternative data sources, but we believe it is of interest to conduct this exercise on economy-wide data (including the government and housing sectors) that is consistent with the US national accounts. More details on the BEA-BLS ILPA are provided in Eldridge et al. (2020).

Decomposition. The conceptual basis for measuring the contribution of Baumol's cost-disease comes from Nordhaus (2002). Nordhaus' method is to construct the rate of aggregate productivity growth that would have obtained if industry's shares of aggregate GDP were held fixed at their year X level. The difference between the realized rate of productivity growth since year X and this counterfactual productivity growth rate represents the contribution of Baumol's cost disease. In Nordhaus (2002), when decomposing *labor* productivity growth, he separates the Baumol effect, the "pure productivity" effect, and the "Dension" effect, named after Edward Denison. The Denison effect measures the productivity gains that come from reallocating labor towards higher productivity *level* (rather than growth) industries.

Relative to Nordhaus (2002), our contribution is to show how aggregate TFP can be decomposed into "pure productivity", "Baumol", "labor reallocation", and "capital reallocation" terms. Our "labor reallocation" term is analogous to the Denison effect, except that it takes into account industry-level profits and is part of TFP, rather than labor productivity. Our "capital reallocation" term is conceptually similar, showing how the reallocation of capital across industries with different *levels* of capital elasticities impacts aggregate TFP growth.

We find that the contribution of labor and capital reallocation to aggregate TFP growth are highly sensitive to what assumptions are made about the level of industry profits/markups. If the economy is assumed to be efficient or near-efficient, the capital and labor reallocation terms are negligible. Under our set of assumptions which leads to the largest markups, capital and labor reallocation went from being a 0.26pp annual boost to TFP growth pre-1973 to an only 0.04pp boost afterwards, helping to explain around 20% of the productivity growth decline.

Additive Growth? In a recent paper, Philippon (2022) argues that US productivity growth is better modeled as linear, rather than exponential as is typically assumed in economic models. Furthermore, under the null of linear TFP growth, there has been no productivity growth slowdown in the post-war US.

As discussed more in the body, our measures of US TFP growth use different data and adjust for profits, leading to different results from Philippon (2022). Like Philippon (2022), we find that linear TFP growth fits the US data better than exponential. However, we find that even under a null linear model, it appears that US TFP growth has declined.

Outline. The paper proceeds as follows. In section 2, we explain how we measure aggregate TFP in inefficient economies. In section 3, we show how to decompose productivity growth to separate out the Baumol effect. Section 4 presents the results on the historical strength of the Baumol effect. Section 5 shows that aggregate TFP growth in the post-war US has been sub-linear.

This paper is intended as a short note to present our results, which build heavily on the existing literature. As such, we keep our presentation concise and refer readers to the other cited papers for greater detail.

2 Measuring TFP

Our method for measuring TFP in inefficient economies comes from Baqaee and Farhi (2020), and is presented carefully and clearly in Vollrath (2024). Relative to the standard national accounts method for measuring TFP, there are two important differences. First, we must adjust industry-level TFP series for the presence of profits. Second, we adjust how we aggregate industry-level TFP series to get aggregate TFP.

2.1 Adjusting industry productivity for profits

The typical measure of industry-level TFP is as follows:

$$\Delta\left(A_{it}\right) = \Delta\left(Y_{it}\right) - \alpha_{it}\Delta\left(K_{it}\right) - (1-\alpha_{it})\Delta\left(L_{it}\right) \tag{1}$$

Here, A_{it} is TFP in industry i at time t; Y is value-added; K is capital; L is labor; and α and $(1-\alpha)$ are capital and labor output elasticities. The use of $(1-\alpha)$ embeds the assumption that industry-level productivity is constant returns to scale — an assumption we will maintain. In national accounts, $1-\alpha$ is measured as the labor share of revenue in that industry:

$$(1-\alpha_{it}) = \frac{W_{it}L_{it}}{P_{it}Y_{it}} \tag{2}$$

Where W and P are wages and prices, meaning that the elasticity of output with respect to labor is the share of the industry wage bill in nominal value-added. Measuring the labor elasticity this way immediately gives the capital elasticity α_{it} as well.

However, in an economy with profits, the output elasticity with respect to labor is *not* the revenue-share of labor; it is the cost-share of labor.[1] The challenge in measuring

[1]Under the assumption that industry-representative firms are cost-minimizers *and* are price-takers in

the output-elasticity of labor is then that either output elasticities need to be directly estimated or that cost-shares need to be directly estimated, which requires that capital costs are directly measured in some way, rather than imputed as a residual from nominal value added minus labor costs.

2.1.1 Measuring the output elasticities

We take two different approaches to measuring the true output elasticity, both of which rely on Compustat firm-level data.

Production-Function Approach. Our first approach directly uses the industry-level output elasticity estimates from the De Loecker, Eeckhout, and Unger (2020) replication package. The underlying estimates are derived using the Olley and Pakes (1996) control-function method on Compustat firm-level data.

Accounting Approach. First, in Compustat, we construct π_{ij}, the profit of firm j in sector i, as the ratio of firm revenue minus cost of goods sold (COGS) minus sales, general, and administrative (SGA) expenses, minus capital expenditures (KEXP) to firm revenues. Firm profit rates are then aggregated to industry-level profit rates π_i using the harmonic mean of firm profits times firms share of industry sales, denoted s_{ij}. Using the harmonic mean following Baqaee and Farhi (2020):

$$(1 + \pi_i) = \left(\sum_i (1 + \pi_{ij})^{-1} s_{ij} \right)^{-1} \tag{3}$$

Then, capital costs are constructed as industry gross output divided by one plus the profit rate, minus the cost of labor and intermediate inputs:

$$r_i K_i = \frac{P_i Q_i}{1 + \pi_i} - W_i L_i - X_i \tag{4}$$

Where Q_i is gross output and X_i is nominal intermediate input expenditure. We use gross output and intermediate expenditure in this setting since intermediate costs are part of COGS and SGA.

The output elasticity with respect to capital is then given by $\alpha_i = \frac{r_i K_i}{r_i K_i + W_i L_i}$, and the labor elasticity is $1 - \alpha_i$.

Making Compustat consistent with ILPA. De Loecker, Eeckhout, and Unger (2020) and our accounting output elasticity estimates are for 22 different industries, while our BEA-BLS ILPA data covers 44 industries. Most industries in the BEA-BLS ILPA data are subsets of the Compustat industry classification, and if so, we assign each sub-industry to

input markets. The latter assumption is not innocuous.

have the same output elasticity as its parent. However, in the ILPA, ”Retail Trade” and ”Transportation and warehousing” are each actually super-sets of two different two-digit industries in the Compustat data.[2] For each of those, we weight the sub-industry output elasticities by their share of combined industry sales to get an aggregate industry output elasticity. Finally, for the industries ”Federal,” ”State and local,” and “Management of companies and enterprises”, we have no corresponding Compustat data, and we revert to the typical zero-profits assumption.

While Compustat data is 1955-2016, ILPA data extends back to 1947. We take the average of each industry’s output elasticity over the period 1955-59 and extend that back as a constant industry output elasticity for the years 1947-54.

2.2 Aggregating industry productivity

As shown by Baqaee and Farhi (2020), in inefficient economies, the correct way of constructing aggregate productivity growth is as follows:

$$\Delta\left(A_t\right) = \Delta\left(Y_t\right) - \omega_t \Delta\left(K_t\right) - (1-\omega_t)\Delta\left(L_t\right) \tag{5}$$

Aggregate growth in value-added $\Delta\left(Y_t\right)$ is, as usual, given by the sum of industry value-added growth weighted by industry’s shares of nominal value-added. What is different here relative to the typical Solow residual is that ω (and $1-\omega$) are cost-based factor elasticities, which depend on the rate of profits in each industry *and* the input-output structure of the economy. The traditional Solow residual would set ω to be the aggregate factor revenue share. Similarly to how we weight aggregate factor growth by cost-shares, the growth in aggregate factor inputs $\Delta\left(K_t\right)$ comes from weighting the growth of industry factor inputs $\Delta\left(K_{it}\right)$ by their shares of industry factor costs:

$$\Delta\left(K_t\right) = \sum_i \frac{r_{it}K_{it}}{r_t K_t} \Delta\left(K_{it}\right) \tag{6}$$

Where $rK = \sum_i r_i K_i$ and $r_i K_i$ is constructed based off either the production function or accounting approach.

The aggregate factor elasticities (or factor cost shares) ω_t and $1-\omega_t$ are the last two entries of the following 46x1 vector Λ_t:

$$\Lambda_t = b_t'(I - \tilde{\Omega}_t)^{-1} \tag{7}$$

Where b' is a 46x1 vector representing for each industry i the share that their produced good has in final expenditure, and $\tilde{\Omega}$ is the 46x46 cost-based input-output matrix. Both the final goods share vector and the cost-based input-output matrix have 46 industries because they include the 44 industries from ILPA and are augmented to treat factors cap-

[2]”Retail trade” is composed of NAICS codes 44 and 45, while ”Transportation and warehousing” is NAICS 48 and 49.

ital and labor as ”industries” which produce no final good and purchase no intermediate goods from other industries but are purchased by the other industries.

The matrix and vector are constructed using the BEA input-output tables, with our output elasticities at the industry level allowing us to construct the cost-based IO matrix.[3]

For a fuller treatment the above, we refer the reader to Baqaee and Farhi (2020) and Vollrath (2024).

2.3 Timing of cost-share weights

Many of these constructions rely on multiplying some input or output's growth rate with the share of that input or output in the industry or aggregate. For example, the α_{it} term in (1) represents the share of capital in industry i costs at time t. However, across all usages of these sorts of shares, we do not use the time t share, but instead use the average of the time t and $t-1$ shares. This is the correct way of weighting growth rates from $t-1$ to t to best approximate the ”true” continuous time growth rate in discrete time. Duernecker, Herrendorf, and Valentinyi (2023) have more discussion. We will continue to label these shares as the time t shares for notational convenience.

3 Decomposing productivity growth

In order to decompose aggregate TFP growth so that the Baumol effect can be isolated, we plug industry productivity growth (1) solved for ΔY_i into the sum of industry value-added growth terms which defines Δy in (5) and expand out the growth in aggregate factor-input terms:

$$\Delta\left(A_t\right) = \sum_i \beta_{i,0}\Delta\left(A_{it}\right) + \sum_i \Delta\left(A_{it}\right)\left(\beta_{it}-\beta_{i,0}\right) + \sum_i \Delta\left(K_{it}\right)\left(\beta_{it}\alpha_{it} - \omega_t\frac{r_{it}K_{it}}{r_tK_t}\right) \tag{8}$$

$$+\sum_i \Delta\left(L_{it}\right)\left(\beta_{it}(1-\alpha_{it}) - (1-\omega_t)\frac{W_{it}L_{it}}{W_tL_t}\right) \tag{9}$$

The new notation here is that β_{it} represents the nominal value-added share of industry i in aggregate nominal value-added: $\beta_{it} = \frac{P_{it}Y_{it}}{P_tY_t}$. The term $\beta_{i,0}$ is industry i's share of nominal value-added at some starting time 0.

Each sum in the decomposition is a different term. Those terms are (in-order) the ”pure productivity” effect, the Baumol effect, the capital reallocation effect, and the labor reallocation effect. The pure productivity effect weights industry TFP growth by that industry's share of value-added at the *beginning* of the sample. The idea is that it shows how much of aggregate productivity growth stems from industry productivity growth,

[3] For each of the three different sets of IO tables used (1947-62, 1963-96, and 1997-2018), industries in the IO table which are subsets of the ILPA industries are combined by summing their input expenditures and the consumption of their output by other industries.

holding fixed the importance of each industry. Obviously, this effect is dependent on when one chooses to start the sample.

The Baumol effect weights industry productivity growth by the *change* in their value-added share between time t and time 0. Therefore, it captures how the changing importance of different industries contributes to aggregate productivity growth. The Baumol term will be negative in a given year if relatively faster productivity growth industries tend to be diminishing in size ($\beta_{it} - \beta_{i,0} < 0$).

The capital reallocation term weights changes in industry-level capital inputs by the difference between the industry factor output elasticity α and the aggregate factor output elasticity ω, weighting the industry factor output elasticity by the industry's share of value-added, and the aggregate factor output elasticity by the the industry's share of nominal factor expenditure. The basic intuition is that it is good for aggregate productivity if capital moves to industries where it is more productive than the economy-wide average unit of capital ($\alpha_i > \omega$), and this effect is stronger the more important is that industry to output relative to costs ($\beta_i > \frac{r_i K_i}{rK}$). Another way of thinking about it is that industry's which are weighted more in output than costs are high markup industries, making it more valuable to reallocate factors toward them and expand their input usage.

The labor reallocation term has the same exact interpretation, just with respect to labor. It is similar in spirit to the Denison effect identified in Nordhaus (2002), which also shows how moving factors to higher productivity industries contributes to aggregate productivity growth.

4 Results

Figure 1 shows productivity growth when using the production-function (PF) approach (thick red line) and the accounting profits approach (thick blue line), as well as the counterfactual productivity growth that would have occurred in the absence of the Baumol effect (dashed lines).

Under the PF approach, TFP would have grown 30pp more over the period 1947-2016 in the absence of the Baumol effect. Under the accounting approach, TFP would have grown 35pp more in the absence of the Baumol effect. These effects are comparable, though somewhat larger, than the 27pp difference Duernecker, Herrendorf, and Valentinyi (2023) found for the 1949-2019 period (using different data and no profit adjustment).

As the graph makes clear, differing profit adjustments make a substantial difference to the estimate of the overall *amount* of productivity growth. The PF approach implies there has been more productivity growth because the PF approach finds larger average markups.[4] Larger average markups imply that, relative to a zero profits baseline, the output elasticity with respect to capital is lower and the output elasticity with respect to

[4]Using the Vollrath (2024) measure of aggregate nominal value-added over nominal capital costs plus nominal labor costs as a measure of the "aggregate value-added markup," the PF approach produces an average markup of 17.4%, while the accounting approach produces an average markup of 7.0%.

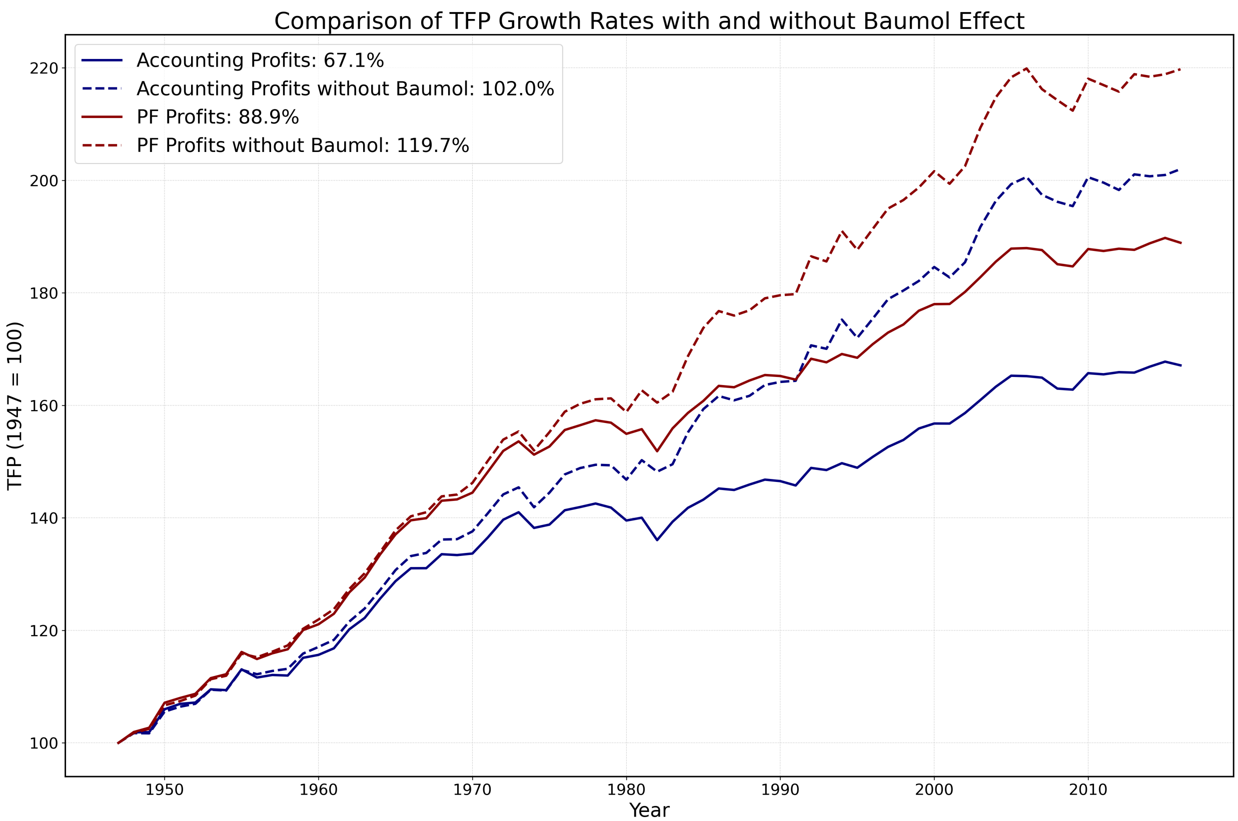

Figure 1: The Baumol Effect

labor is higher. In the US, capital inputs have grown faster than labor inputs, so pushing down the capital elasticity (and up the labor elasticity) leaves more of a residual for TFP to explain.

Tables (1), (2), and (3) present the full results of our decomposition. Table (1) uses the production function approach to measuring output elasticities; table (2) uses the accounting profits approach; and (3) uses the traditional zero-profits assumption.

Table 1: Production Function Approach

	Full Sample	Pre-1973	Post-1973
Aggregate Productivity	0.93	1.69	0.50
PP	1.04	1.49	0.78
Baumol	-0.22	-0.05	-0.32
Capital Allocation	0.07	0.12	0.04
Labor Allocation	0.05	0.14	-0.00

Table 2: Accounting Approach

	Full Sample	Pre-1973	Post-1973
Aggregate Productivity	0.75	1.35	0.41
PP	1.04	1.52	0.77
Baumol	-0.28	-0.13	-0.37
Capital Allocation	-0.01	-0.04	0.01
Labor Allocation	-0.00	-0.00	-0.01

Table 3: Zero Profits

	Full Sample	Pre-1973	Post-1973
Aggregate Productivity	0.65	1.25	0.31
PP	0.93	1.37	0.69
Baumol	-0.28	-0.12	-0.37
Capital Allocation	0.00	0.00	0.00
Labor Allocation	0.00	0.00	0.00

On an annual basis, the Baumol effect has been somewhere between a -0.22pp and -0.28pp drag on productivity growth. Comparing pre and post-1973 periods, the Baumol effect *reduced* post-1973 productivity growth relative to its pre-1973 annual rate by somewhere between 0.24-0.27pp. Given that the fall in aggregate productivity growth was between 0.94-1.19pp, the Baumol effect is responsible for between 23-27% of that decline.

The tables also make clear how important it is to correctly measure profits in order to firmly assess historical productivity growth. It could be that either one is closer to the truth, but the 0.28pp gap in annual post-war productivity growth between the PF approach and the baseline zero-profit approach is quite significant. The 0.10pp gap between the zero-profit approach and the accounting approach is similar to the 0.12pp gap between a zero-profit approach and a user-cost based approach to measuring markups found in Comin et al. (2020), though they look at 1987-2020 data. In Vollrath (2024), the gap between zero-profit and profit-adjusted based productivity growth is anywhere between 0.05pp and 0.31pp.

Another point the tables make clear is that the majority of post-1973 TFP slowdown is due to *within* industry slowdowns in productivity growth. Annual pure productivity growth is between 0.68-75pp lower post-1973. This suggests that some other explanation, such as "ideas are getting hard to find" (Bloom et al. 2020) is needed to fully account for the productivity growth slowdown.

Finally, turning to the capital and labor allocation terms, we see that under either the zero-profits or the accounting approach, their contribution is negligible. The exact zeros in the zero profits table is not by construction — it just so happens that the terms are minuscule.

By contrast, under the PF approach, both capital and labor reallocation contributed substantially positively to TFP growth in the pre-1973 period, and then declined to relative insignificance post-1973.

5 Linear Growth?

Before wrapping up, we briefly look at what this national accounts consistent productivity data says about the overall trend in US TFP growth. Philippon (2022) makes the observation that if you look at post-war US TFP up to 1983 and afterwards, a linear trend fits the data, while the traditionally used exponential trend does not. Accordingly, he argues that if the null growth model is additive, then there has been no "slowdown" in US productivity growth.

Philippon (2022) uses two different data sources in his work. One is Fernald (2014), which takes the BLS data which encompasses just the US business sector and adjusts for capacity utilization. The other is from Bergeaud, Cette, and Lecat (2016), which includes the entire economy. Neither makes an adjustment for profits, and Bergeaud, Cette, and Lecat (2016) does not have a human capital adjustment to their labor input series either.

Turning to our national accounts consistent BEA-BLS ILPA data, figure (2) shows that on both the accounting profits and the PF approach to measuring aggregate TFP, *neither* an exponential nor a linear model can fit the post-war US TFP process. It is not shown here, but the same pattern holds for the zero-profits TFP series in our data.

In the figure, the exponential trend is fit on the pre-1973 data, while linear trends are fit on both pre-1973 and pre-1983 data, for comparability with the 1983 cutoff used in Philippon (2022). However, the post-1983 (and even moreso for 1973) realized data clearly lies below the linear fit.

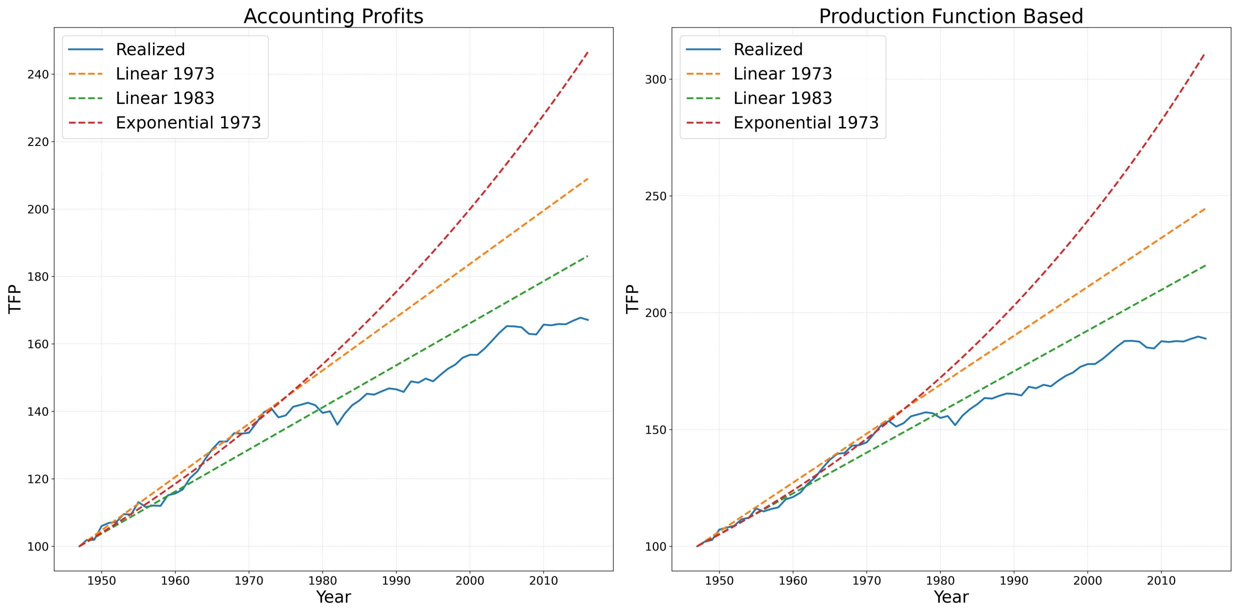

Figure 2: Exponential, Linear, or Neither?

The bottom line is that, in our data, US productivity growth has slowed down, even if the null model for growth is an additive one.

References

Baqaee, David Rezza and Emmanuel Farhi (2020). "Productivity and misallocation in general equilibrium". In: *The Quarterly Journal of Economics* 135.1, pp. 105–163.

Baumol, William J (1967). "Macroeconomics of unbalanced growth: the anatomy of urban crisis". In: *The American economic review* 57.3, pp. 415–426.

Bergeaud, Antonin, Gilbert Cette, and Rémy Lecat (2016). "Productivity trends in advanced countries between 1890 and 2012". In: *Review of Income and Wealth* 62.3, pp. 420–444.

Bloom, Nicholas et al. (2020). "Are ideas getting harder to find?" In: *American Economic Review* 110.4, pp. 1104–1144.

Comin, Diego et al. (2020). "A New Measure of Utilization-Adjusted TFP Growth for Europe and the United States". In: *NBER Working Paper* w28008.

De Loecker, Jan, Jan Eeckhout, and Gabriel Unger (2020). "The rise of market power and the macroeconomic implications". In: *The Quarterly Journal of Economics* 135.2, pp. 561–644.

Duernecker, Georg, Berthold Herrendorf, and Ákos Valentinyi (2023). "Structural Change within the Services Sector and the Future of Cost Disease". In: *Journal of the European Economic Association*, jvad030.

Eldridge, Lucy P et al. (2020). "Toward a bea-bls integrated industry-level production account for 1947–2016". In: *Measuring economic growth and productivity*. Elsevier, pp. 221–249.

Fernald, John (2014). "A quarterly, utilization-adjusted series on total factor productivity". In: Federal Reserve Bank of San Francisco.

Nordhaus, William D (2002). "Productivity growth and the new economy". In: *Brookings Papers on Economic Activity*, pp. 211–267.

Olley, G Steven and Ariel Pakes (1996). "The Dynamics of Productivity in the Telecommunications Equipment Industry". In: *Econometrica* 64.6, pp. 1263–1297.

Philippon, Thomas (2022). *Additive growth*. Tech. rep. National Bureau of Economic Research.

Vollrath, Dietrich (2024). "The elasticity of aggregate output with respect to capital and labor". In: *American Economic Journal: Macroeconomics* 16.4, pp. 470–504.