Returns to Scale and Productivity

Joel Kariel* Anthony Savagar†

February 15, 2022

Abstract

We estimate returns to scale in the U.K. economy from 1998 - 2014 and show that returns to scale have increased over time. We show that there is a positive relationship between returns to scale and productivity across industries. However, within an industry, there is a negative relationship between returns to scale and productivity. We reconcile these results in a firm dynamics model with endogenous returns to scale. In the model industries have different fixed costs. High fixed-cost industries only sustain high-productivity firms (a selection effect). However, within an industry high-productivity firms are large, and they have lower returns to scale because they utilize their fixed cost more.

JEL: E32, E23, D21, D43, L13.
Keywords: Returns to Scale, Productivity, Markups, Market Structures, Firm Dynamics.

*Oxford University and University of Kent, joel.kariel@economics.ox.ac.uk
†a.savagar@kent.ac.uk. This research is funded under ESRC project reference ES/V003364/1.
Disclaimer: *This work was produced using statistical data from ONS. The use of the ONS statistical data in this work does not imply the endorsement of the ONS in relation to the interpretation or analysis of the statistical data. This work uses research datasets which may not exactly reproduce National Statistics aggregates.*

Can changing returns to scale explain recent trends in market power and productivity? We address this question by estimating markups, returns to scale and productivity at the firm-level for a panel of U.K. firms. And, we explain our empirical findings by developing a heterogeneous firm model with endogenous returns to scale. Returns to scale are crucial to understand both the causes and implications of recent trends in market power. A rise in returns to scale will increase markups, but it also means that higher markups will not translate one-to-one into higher economic profits since higher returns to scale increase the marginal product of inputs which raises their price and therefore reduces profits. Consequently, the policy responses, macroeconomic transmission mechanisms, and implications for productivity are different from a scenario where returns to scale are unchanged.

Our findings show that aggregate markups in the U.K. have risen from 1998 to 2014, and the rise in aggregate markups coincide with a rise in returns to scale. We find that on average the U.K. economy has decreasing returns to scale, but these have risen to be closer to constant returns to scale. At a more granular level, we show there is a nuanced relationship between productivity and returns to scale. There is a positive relationship between returns to scale and productivity across industries: higher productivity industries have have higher returns to scale. However, within an industry, there is a negative relationship between returns to scale and productivity: high productivity firms have lower returns to scale. We show that these results are consistent with a heterogeneous firm model with endogenous returns to scale. The core idea is that fixed costs vary across industries but they are constant within an industry. As a result, industries with higher fixed costs (*i.e.* higher returns to scale) have higher productivity firms because low productivity firms cannot pay the fixed cost (a selection effect). However, within industries all firms have the same fixed cost, so the more productive industries are larger and make greater use of their fixed cost, such that they have lower returns to scale.

We estimate production functions for a panel of firms in the U.K. from 1998-2014. Our main dataset is the Annual Business Survey (ABS) which is a representative sam-

ple of roughly 50,000 firms in the U.K. each year – a well-known application of the dataset is Aghion, Bloom, Blundell, Griffith, and Howitt (2005). To estimate production functions we use control function approaches by Levinsohn and Petrin (2003) and Ackerberg, Caves, and Frazer (2015), and we estimate both Cobb-Douglas and Translog functional forms. The control function estimation approaches address endogeneity issues caused by unobserved productivity at the firm level. The estimation results provide coefficients that represent the elasticity of output to factors of production. We use these estimated elasticities to infer returns to scale and markups based on optimization conditions from a firm's cost minimization problem. The residuals of these regressions provide us with productivity measures. Our heterogeneous firm model is based on Hopenhayn (1992) and Restuccia and Rogerson (2008). Our innovations on the baseline frameworks are to introduce endogenous returns to scale through fixed costs and imperfect competition.

Recent research studying changing market structures – rising markups, rising profit shares, declining business dynamism – falls under two broad hypotheses. First, *technological changes* have caused changing market structures (Van Reenen 2018; Autor, Dorn, Katz, Patterson, and Van Reenen 2017; Lashkari, Bauer, and Boussard 2019; Bessen 2020; De Ridder 2019; Aghion, Bergeaud, Boppart, Klenow, and Li 2019). Technological factors are usually represented as primitive parameters in a production function. An implication of this literature is that 'superstar' firms that have harnessed the latest technologies to grow at the expense of others, but this is necessarily not bad for productivity and efficiency. Second, *behavioural changes* have caused changing market structures (Gutierrez, Jones, and Philippon 2019; Barkai 2020; De Loecker, Eeckhout, and Mongey 2021). Behavioural changes refer to outcomes that change due to the optimizing behaviour of firms, such as markups and profits, and are affected by factors such as antitrust regulation and strategic interactions among firms. Primitives that effect market structures, such as barriers to entry or strategic interactions, lead to higher markups. Returns to scale provide a theoretical link between these two concepts. Both of these hypotheses are consistent with rising returns to scale. Re-

turns to scale can be expressed as a function of technical parameters of a production function (fixed costs and returns to variable production) or they can be expressed as a function of profits and markups. In addition to papers that focus on the causes of changing market structures, there is also literature that attempts to understand the implications for other macroeconomic variables and welfare (Edmond, Midrigan, and Xu 2021; De Loecker, Eeckhout, and Mongey 2021; Gutiérrez, Jones, and Philippon 2021; Eggertsson, Robbins, and Wold 2021). These papers develop dynamic general equilibrium models with firm entry and imperfect competition. Baqaee, Farhi, and Sangani (2020) focus directly on returns to scale to show how reallocation of output towards larger firms can be welfare enhancing if they benefit from returns to scale. Rotemberg and Woodford (1999) and Kim (2004) provide explicit analyses of the role of returns to scale in shaping macroeconomic dynamics in DSGE models, and Benhabib and Farmer (1994) and Farmer and Guo (1994) initiate a theoretical literature which explores the implications of increasing returns to scale for multiple equilibria in RBC models.

Economists have long tried to measure returns to scale empirically. Early work by Hall (1988) and Hall (1990) finds increasing returns to scale in the U.S. using instrumental variable estimation with value-added data. Caballero and Lyons (1992) also find increasing returns, though their concept of returns to scale is *external* returns to scale that arises firm output is aggregated. The "Hall regression" technique yields returns to scale and markup measures at an industry level. It has been widely adopted in applied studies, and forms the methodological foundation for recent developments to measure markups at the firm level (De Loecker and Warzynski 2012a). Harrison (1994) and Levinsohn (1993) apply the approach in an international trade context, whilst Basu and Fernald (1997) explain that the measure is misleading when applied to value-added data with non-constant returns and imperfect competition. Instead, they provide production function estimates on gross-output which yields constant or decreasing returns, and is consistent with Burnside (1996) who performs the same approach and finds returns to scale of 0.9. Altug and Filiztekin (2002) and Kee (2004)

provide helpful reviews of this literature, and Feenstra (2003, Ch.10) provides a textbook treatment of the Hall methodology. Using more recent methods, Lashkari, Bauer, and Boussard (2019), Ruzic and Ho (2019), and Gao and Kehrig (2021) provide estimates of returns to scale in the U.S. economy. Gao and Kehrig (2021) estimate returns to scale of 0.96 overall in U.S. manufacturing firms, 0.92 in construction, 0.95 in nondurable manufacturing and 0.97 in durable manufacturing. They also highlight variation in returns to scale across 4-digit industries, ranging from 0.86 to 1.3. Ruzic and Ho (2019) estimate returns to scale at the industry level. They find a decline in returns to scale from 1.2 in 1982 to 0.96 in 2007. Lastly, Lashkari, Bauer, and Boussard (2019) find returns to scale ranging from 0.75 to 1.06, with smaller firms obtaining larger scale economies. The most recent estimates of returns to scale in the U.K. economy are Oulton (1996), Harris and Lau (1998), and Girma and Görg (2002). These studies document constant or slightly decreasing returns to scale. The research uses smaller datasets of manufacturing firms and different estimation strategies based on data availability and econometric methodology at the time of publication.

1 Simple Model

In this section we present the theoretical framework for our our main model. The point of the section is to clarify our interpretation or returns to scale and show how this is related to productivity.

1.1 Returns To Scale

Returns to scale are a feature of a firm's production function. They describe the change in firm's output as inputs are changed, holding other factors constant.[1] Returns to scale are described as increasing, decreasing or constant depending on whether firm output changes more than, less than or proportionally to a change in inputs. Produc-

[1] Economies of scale are a related concept that capture the cost advantages or disadvantages of production at different scales. Returns to scale, on the other hand, are related to a firm's production function.

tivity measures the amount of output a firm produces for a given amount of factor inputs. At a first-pass, the concepts appear tautological: firms with a production technology that yields more output for a given increase in input should be more productivity.

Definition 1. *Returns to scale (*RTS*) are the inverse cost elasticity. The inverse cost elasticity is the ratio of average cost to marginal cost. Thus,*

$$\text{RTS} \equiv \left(\frac{\partial \mathcal{C}}{\partial y} \frac{y}{\mathcal{C}} \right)^{-1} = \frac{\mathcal{AC}}{\mathcal{MC}} \tag{1}$$

where $\text{AC} \equiv \mathcal{C}/y$ *and* $\text{MC} \equiv \partial \mathcal{C}/\partial y$.

We define returns to scale as increasing, decreasing or constant as follows:

$$\text{RTS} \equiv \begin{cases} \text{Increasing returns,} & \text{if RTS} > 1 \\ \text{Constant returns,} & \text{if RTS} = 1 \\ \text{Decreasing returns,} & \text{if RTS} < 1 \end{cases}$$

Figure 1 presents returns to scale for a firm with U-shaped average cost curve due to upward-sloping marginal cost and a fixed cost, which is consistent with our model. At the intersect of average and marginal cost a firm has constant returns. To the left-hand side of the minimum, average cost exceeds marginal cost so there are increasing returns, and to the right-hand side of the minimum, average cost is less than marginal cost so there are decreasing returns. Hence, size and returns to scale are negatively related at the firm level.

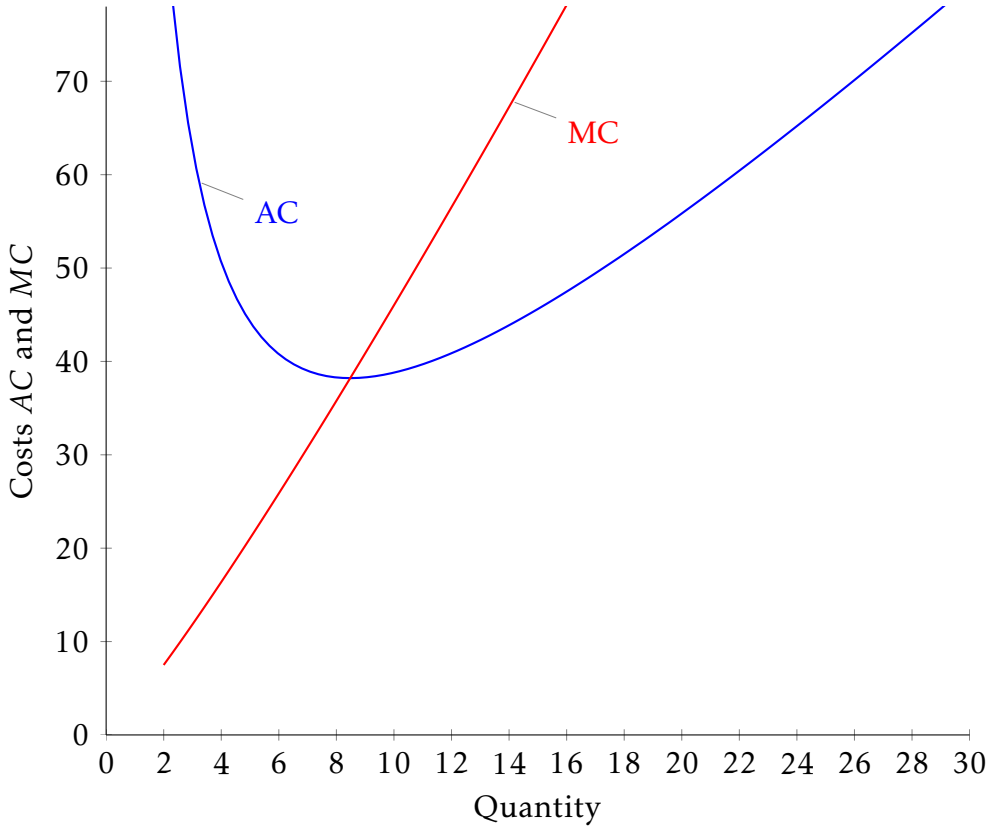


Figure 1: Fixed Cost with Increasing MC, U-Shaped AC Curve

We do not observe total and marginal costs, so directly computing returns to scale using equation (1) is not possible. Thus we use theory to obtain different expressions for returns to scale.

1.2 Net Output

A firm's net output is given by:

$$y = zF(k,\ell) - \phi, \tag{2}$$

where y is net output, z is Hicks-neutral productivity, k is capital, ℓ is labour, and ϕ is an output-denominated fixed cost.[2] We assume that $F(\cdot)$ is twice continuously differentiable, strictly concave, and homogeneous of degree ν.

[2]The specification follows a large literature which cites Rotemberg and Woodford (1993) and Rotemberg and Woodford (1999). The output-denominated fixed cost is output which is taken from total production before sold, such as HR costs, defects, or asset disposals. In other words, it is non-revenue generating output.

1.3 Elasticities

In linearized form net output is given by

$$\hat{y} = \hat{z} + \varepsilon_{yk}\hat{k} + \varepsilon_{y\ell}\hat{\ell}, \tag{3}$$

where hat notation represents deviation from steady-state or trend. The coefficients represent output elasticity to capital and output elasticity to labour respectively, these are given by:

$$\varepsilon_{yk} \equiv \frac{\partial y}{\partial k}\frac{k}{y} = zF_k(k,\ell)\frac{k}{y} \quad \text{and} \quad \varepsilon_{y\ell} \equiv \frac{\partial y}{\partial \ell}\frac{\ell}{y} = zF_\ell(k,\ell)\frac{\ell}{y}. \tag{4}$$

Later, we shall show that when firms cost minimize output elasticities are equal to the markup multiplied by the factor's cost share in sales.[3] The sum of the output elasticities yields

$$\varepsilon_{yk} + \varepsilon_{y\ell} = \nu(1 + s_\phi), \quad \text{where } s_\phi \equiv \frac{\phi}{y}. \tag{5}$$

The variable s_ϕ is the fixed cost share in net output. The result uses Euler's homogeneous function theorem that $F_k k + F_\ell \ell = \nu F$. The sum of the two coefficients represent the change in net output from a change in both inputs which, from intuition, represents returns to scale. For example, if they sum to one, then a change in capital and labour cause a proportional change in net output. In the next section we will show formally that this represents returns to scale by showing that $\nu(1 + s_\phi)$ is the ratio of average cost to marginal cost.

1.4 Firm problem

Firms maximize profits by solving a two-stage problem. First, they choose inputs to minimise variable costs subject to a production function which yields a cost function under optimal input choices, second, they select output to maximise profits subject

[3] A firm's cost minimization conditions yield that nominal wage is given by $w = \mathcal{M}\mathcal{C}zF_\ell$, then dividing by product price gives $w/p = zF_\ell/\mu$ where $\mu \equiv p/\mathcal{M}\mathcal{C}$. Therefore, this factor market equilibrium expression relates to the output elasticity for labour as follows $\varepsilon_{y\ell} = \frac{w\ell}{py} \times \mu$.

to demand and the cost-function from stage one. The factor market is competitive, so factor prices are given.

1.4.1 Cost minimization

A cost-minimising firm will set total variable costs equal to:

$$\begin{aligned} \mathcal{C} :- \min_{\ell,k} \quad & w\ell + rk, \\ \text{s.t.} \quad & y + \phi \leq z k^{\alpha} \ell^{\beta} \end{aligned}$$

The equilibrium conditions yield

$$r = \lambda \alpha \frac{y + \phi}{k} \tag{6}$$

$$w = \lambda \beta \frac{y + \phi}{\ell} \tag{7}$$

$$y = zF(k, \ell) - \phi. \tag{8}$$

The Lagrange multiplier λ is also the marginal cost. Substituting optimal factor prices into the objective function yields the nominal cost function is

$$\mathcal{C} = \nu \lambda (y + \phi). \tag{9}$$

If we divide by y, we get average cost, then divide by λ yields the ratio of average cost to marginal cost, which gives

$$\text{RTS} = \nu(1 + s_{\phi}). \tag{10}$$

This is the same expression as the sum of the net output elasticities and confirms that summing the firm-level production function coefficients yields returns to scale.

We do not need to determine the markup from profit maximization and the demand condition, in order to present an alternative form of returns to scale. It is suffi-

cient to assume that nominal profits are given by

$$\pi = py - \mathcal{C}. \tag{11}$$

From the profit statement, we can see an alternative way of presenting the ratio of average cost to marginal cost which is returns to scale:

$$\text{RTS} = \mu(1 + s_\pi), \text{ where } s_\pi \equiv \pi/py. \tag{12}$$

Full derivation in appendix. Use the profit statement and our cost function to show an alternative expression for returns to scale. We consider firms on a productivity continuum $j \in (0,1)$.[4] Using our expression for nominal costs in the definition of nominal profits yields

$$\begin{aligned}
\pi(j) &= p(j)y(j) - \nu\lambda(j)(y(j) + \phi) && (13)\\
&= \left(1 - \frac{\nu}{\mu}\right)p(j)(y(j) + \phi) - p(j)\phi, \text{ where } \mu \equiv p/\lambda, && (14)\\
&= \left(1 - \frac{\nu(1 + s_\phi(j))}{\mu}\right)p(j)y(j), \text{ where } s_\phi \equiv \phi/y(j), && (15)\\
&= \left(1 - \frac{\text{RTS}(j)}{\mu}\right)p(j)y(j). && (16)
\end{aligned}$$

Hence returns to scale is negatively related to the profit share in revenue. The expression shows the equivalence of our two definitions of returns to scale

$$\mu(1 - s_\pi(j)) = \nu(1 + s_\phi(j)). \tag{17}$$

1.4.2 Profit Maximization

Profit maximization subject to a demand function determines the difference between price and marginal costs, i.e. the markup. Under monopolistic competition the markup

[4]For our model, firms operate in sectors and the productivity process is described in more detail, but the key results of this section still hold.

is

$$\mu \equiv \frac{\theta}{\theta - 1} \tag{18}$$

where θ is substitutability among producers. Whereas under Cournot competition with homogeneous goods within a sector and product differentiation across sectors, the markup is

$$\mu \equiv \frac{\theta N_t}{\theta N_t - 1} \tag{19}$$

where θ_I is substitutability across sectors and N_t is the number of firms in a sector.

1.5 Cut-off Productivity

Firms only live for one period, so their value is either the profits they earn if they decide to operate or zero if they decide not to operate. Therefore firm value is given by the discontinuous function

$$\pi = \max\left\{0, \pi^{\mathrm{P}(j)}\right\}.$$

This recognises that when profits from production are negative a firm will choose to produce 0. A firm with $p(j)y(j) > p(J)y(J)$ makes positive profit. A firm that has paid an entry cost will not produce below $p(J)y(J)$ as they would make negative profits, so would prefer to make zero profits by not producing. The threshold firm is also the smallest firm. The threshold firm obtains productivity $J \in (0,1)$ and makes zero profits:

$$\left(1 - \frac{\nu}{\mu}\right) p(J)(y(J) + \phi) = p(J)\phi \tag{20}$$

Hence the threshold firm size is

$$y(J) = \frac{\nu}{\mu}\left(1 - \frac{\nu}{\mu}\right)^{-1} \phi \tag{21}$$

And, the threshold fixed cost share is

$$\frac{\phi}{y(J)} = \frac{\mu}{\nu} - 1 \tag{22}$$

Furthermore, returns to scale of the threshold firm are

$$\text{RTS}(J) = \nu(1 + s_\phi(J)) = \mu \tag{23}$$

This is an important result. The smallest firm – that exists at the productivity threshold – has increasing returns to scale which are equivalent to the markup. Larger firms will have a positive profit share and strictly lower returns to scale, which could be constant or even decreasing.

Substituting the fixed cost from equation (20) into the profit condition in (13), we obtain the profit of an individual producer:

$$\begin{aligned}
\pi(j) &= \left(1 - \frac{\nu}{\mu}\right) p(j)\left(y(j) + \phi\right) - \frac{p(j)}{p(J)}\left(1 - \frac{\nu}{\mu}\right) p(J)\left(y(J) + \phi\right) \\
&= \left(1 - \frac{\nu}{\mu}\right) p(J)\left(y(J) + \phi\right)\left(\frac{p(j)}{p(J)}\frac{\left(y(j) + \phi\right)}{\left(y(J) + \phi\right)} - \frac{p(j)}{p(J)}\right) \\
&= \left(1 - \frac{\nu}{\mu}\right) p(j)\left(y(J) + \phi\right)\left(\frac{\left(y(j) + \phi\right)}{\left(y(J) + \phi\right)} - 1\right) \\
&= p(j)\phi\left(\frac{y(j) + \phi}{y(J) + \phi} - 1\right) \\
&= \phi\left(\frac{a(j)}{a(J)} - 1\right)
\end{aligned}$$

The final bridge comes from the result that gross revenue $p(j)(y(j) + \phi)$ and factor inputs are employed proportionally to productivity dispersion $a(j)/a(J)$, which is a result of cost minimization and perfect factor markets. The result shows that a producing firm with productivity index $j > J$ will have higher profits the further its productivity draw is from the threshold level.

A firm's expected profits after paying the entry cost but before receiving the produc-

tivity draw is:

$$\mathbb{E}(\pi) = \phi \int_J^1 p(j) \left[\frac{y(j) + \phi}{y(J) + \phi} - 1 \right] dj = \phi \int_J^1 \left[\frac{a(j)}{a(J)} - 1 \right] dj = \phi (1 - J_t) \left[\frac{\bar{a}(\bar{j})}{a(J_t)} - 1 \right] \tag{24}$$

The final bridge uses the mean-value theorem for integrals to represent average productivity $\bar{a}(\bar{j})$ for $\bar{j} \in [J_t, 1]$. The result shows that expected profits are increasing in the fixed cost and average productivity, but they are decreasing in the cut-off.

1.6 Productivity

We have established that returns to scale RTS $= \mu(1 - s_\pi)$ and profits are negatively related and that the smallest, least profitable, firm will have the highest returns to scale exactly equivalent to the markup. It is the firm that pays the same fixed cost as everyone else, but gets the least use out of it given their break-even productivity draw. Given the negative relationship between returns to scale and profit, higher returns to scale reduce profits, and therefore raise the threshold productivity J to survive. Gao and Kehrig (2021) study a similar relationship with perfect competition. Thus, Equation (24) highlights an important result. Increases in returns to scale, and thus cut-off J, affect the productivity distribution of incumbents in two ways. The first is to narrow the continuum over which the integral is computed. The second is is to raise the value of the term on the denominator of the fraction $a(J)$.

We call this the *selection effect*. When returns to scale are high, average productivity is higher, as only more productive firms can exist in an environment with greater average-to-marginal-cost ratios. Notice that this occurs at the industry level (here we consider just one sector). At the firm level, as shown in Equation (5), returns to scale and productivity are negatively related.

Heterogeneous productivity leads to a heterogeneous degrees of returns to scale across all inputs (labour, capital, overhead) used in production. If a firm has low productivity, it will be small and the initial fixed cost will dominate the decreasing returns in variable production leading to increasing returns to scale. If the firm has

high productivity, it will be large and the decreasing returns in variable inputs will offset the increasing returns from the overhead, so overall the firm has decreasing returns. This framework captures our main empirical finding that returns to scale and productivity tend to be positively related across industries, but negatively related within industries.

2 Model

There is a representative household which owns all firms. It has preferences over consumption and labour supply. Final output is produced by perfectly competitive firms using inputs from a continuum of sectors, each of which has a discrete number of imperfectly competitive heterogeneous firms. These firms combine labour and capital in a production function with homogeneity of degree ν, and pay an output-denominated fixed cost ϕ. Firms are heterogeneous in productivity, which combine idiosyncratic and sector-specific components from Pareto distributions.

This model framework is closely related to Edmond, Midrigan, and Xu (2015) and Edmond, Midrigan, and Xu (2021). We introduce non-unity *variable* returns to scale, and endogenous returns to scale with our net output specification. We document how these extensions affect aggregate productivity, and the relationship between returns to scale, productivity, and firm size.

2.1 Household

The representative household maximizes utility subject to a budget constraint:

$$\max_{C_t, L_t, K_{t+1}} \quad \sum_{t=1}^{\infty} \rho^t \left(\ln(C_t) - \psi \frac{L_t^{1+\eta}}{1+\eta} \right) \tag{25}$$

$$\text{s.t.} \quad C_t + I_t = w_t L_t + r_t K_t + \Pi_t \tag{26}$$

$$I_t = K_{t+1} - (1-\delta) K_t \tag{27}$$

C_t is consumption of the final good, I_t is investment, K_t is physical capital, L_t is labour, w_t is wage from labour r_t is the return from renting capital to firms, and Π_t are the aggregate net profits from owning firms. This yields the following optimization conditions for the firm

$$w_t = \psi C_t L_t^{\eta} \tag{28}$$

$$1 = \rho \frac{C_t}{C_{t+1}} (r_{t+1} + 1 - \delta) \tag{29}$$

Equation (28) is the households intratemporal condition which determines labour supply, and (29) is the household intertemporal condition that determines investment choices.

2.2 Firm

2.2.1 Final Good Producer

The final good Y is produced by perfectly competitive firms using inputs $y(s)$ from a continuum of intermediate sectors:

$$Y = \left(\int_0^1 y(s)^{\frac{\theta-1}{\theta}} ds \right)^{\frac{\theta}{\theta-1}}.$$

In each sector s, output is produced by $n(s)$ firms:

$$y(s) = n(s)^{1+\epsilon} \left(\frac{1}{n(s)} \sum_{j=1}^{n(s)} y_j(s)^{\frac{\eta-1}{\eta}} \right)^{\frac{\eta}{\eta-1}},$$

where $\eta > \theta$ is the elasticity of substitution across goods *within sector s*. Love of variety is exhibited by $\epsilon > 0$, but we set $\epsilon = 0$ for our benchmark model.

Final Good Producer Problem

The final good price is P. Final good firms choose to purchase inputs $y_j(s)$ to maximise:

$$\max_{y_j(s)} \quad PY - \int_0^1 \sum_{j=1}^{n(s)} p_j(s) y_j(s) ds,$$

subject to the definition of the final good and the sector input. This yields the demand function:

$$y_j(s) = \left(\frac{p_j(s)}{p(s)}\right)^{-\eta} \left(\frac{p(s)}{P}\right)^{-\theta} Y,$$

with the price indices $P = \left(\int_0^1 p(s)^{1-\theta} ds\right)^{\frac{1}{1-\theta}}$ and $p(s) = n(s)^{\frac{1}{\eta-1}} \left(\sum_{j=1}^{n(s)} p_j(s)^{1-\eta}\right)^{\frac{1}{1-\eta}}$.

2.2.2 Intermediate Goods Producer

Intermediate firm j produces according to:

$$y_j(s) = z_j(s) k_j(s)^{\alpha} \ell_j(s)^{\beta} - \phi(s).$$

The production technology is homogeneous of degree $\nu \equiv \alpha + \beta$, and ϕ is an output-denominated fixed cost.

Intermediate Goods Producer Cost Min Problem

Taking factor prices as given, cost minimising firms solve:

$$\mathcal{C}_j(s) := \min_{k_j(s), \ell_j(s)} \; w\ell_j(s) + rk_j(s) \quad \text{s.t.} \quad y_j(s) \geq z_j(s) F(k_j(s), \ell_j(s)) - \phi(s).$$

This yields the solutions:

$$rk_j(s) = \lambda_j(s) \alpha (y_j(s) + \phi(s)) \tag{30}$$

$$w\ell_j(s) = \lambda_j(s) \beta (y_j(s) + \phi(s)) \tag{31}$$

$$y_j(s) = z_j(s) F(k_j(s), \ell_j(s)) - \phi(s) \tag{32}$$

By Euler's Homogeneous Function Theorem, total costs are:

$$\mathcal{C}_j(s) := rk_j(s) + w\ell_j(s) = \nu\lambda_j(s)(y_j(s) + \phi(s)).$$

The input demand conditions imply that the marginal cost can be written:

$$\lambda_j(s) = \frac{1}{z_j(s)}\left(\frac{r}{\alpha}\right)^{\alpha}\left(\frac{w}{\beta}\right)^{\beta}.$$

Returns to Scale and Profits

With the cost function we can present returns to scale and also the reduced-form profit function. Returns to scale is the ratio of average to marginal costs, as it is the elasticity of output with respect to costs:

$$\mathrm{RTS}_j(s) = \frac{\partial y_j(s)}{\partial \mathcal{C}_j(s)}\frac{\mathcal{C}_j(s)}{y_j(s)} = \frac{\mathcal{AC}_j(s)}{\mathcal{MC}_j(s)}.$$

Average costs are total costs divided by output: $\nu\lambda_j(s)\left(\frac{y_j(s)+\phi(s)}{y_j(s)}\right)$. Dividing this by marginal costs $\lambda_j(s)$ yields:

$$\mathrm{RTS}_j(s) = \nu\left(\frac{y_j(s) + \phi(s)}{y_j(s)}\right).$$

This is equal to returns to scale in the variable production function multiplied by one plus the fixed cost share in output. The returns to variable production are equivalent to the slope of the marginal cost curve. Profits are equal to revenue minus costs. Using the expression for total costs from cost minimisation and rearranging, we get:

$$\pi_j(s) = \left(1 - \frac{\nu}{\mu_j(s)}\right)p_j(s)\Big(y_j(s) + \phi(s)\Big) - p_j(s)\phi(s),$$

where profits are equal to gross revenue $p(y+\phi)$ multiplied by one minus the revenue elasticity $\frac{\nu}{\mu}$, minus the value of the output-denominated fixed cost.

Intermediate Goods Producer Profit Max Problem

Intermediate producers engage in Cournot competition by solving the following problem whilst recognising that their output choices will affect sectoral outcomes:

$$\max_{y_j(s)} p_j(s) y_j(s) - \mathcal{C}. \tag{33}$$

subject to the inverse demand function from the final goods producer and the cost-function from the first stage of the intermediate goods problem. The first-order conditions imply that firms price at a markup that depends on their market share and how differentiated they are from other firms and other sectors:

$$\mu_j(s) = \left[\left(1 - \frac{1}{\eta}\right) + \left(\frac{1}{\eta} - \frac{1}{\theta}\right) \omega_j(s) \right]^{-1},$$

The variable $\omega_j(s)$ is the market share of firm j in sector s.

2.3 Market Clearing

The aggregate market clearing condition for the final good is

$$Y_t = C_t + I_t. \tag{34}$$

2.4 Aggregation

At each level of aggregation, we define net output as value-added multiplied by productivity, minus the total output-denominated fixed costs. The productivity terms will be calculated to ensure consistency across the different levels of aggregation.
The equation for sectoral net output is:

$$y(s) = z(s) F(k(s), \ell(s)) - n(s)\phi(s). \tag{35}$$

where $k(s) = \sum_{j=1}^{n(s)} k_j(s)$ and $\ell(s) = \sum_{j=1}^{n(s)} \ell_j(s)$ are the sums of firm input choices within

sectors.

And aggregate net output is denoted analogously:

$$Y = ZF(K,L) - \Phi. \tag{36}$$

where $K = \int_0^1 k(s)ds$, $L = \int_0^1 \ell(s)ds$ and $\Phi = \int_0^1 n(s)\phi(s)ds$.

With these definitions in hand, computing productivity aggregates is straightforward, following Edmond, Midrigan, and Xu (2015). Aggregate productivity can be written:

$$Z = \frac{Y+\Phi}{F(K,L)} = \frac{Y+\Phi}{\int_0^1 \frac{y(s)+n(s)\phi(s)}{z(s)}ds} = \left(\int_0^1 \frac{1}{z(s)}\frac{y(s)+n(s)\phi(s)}{Y+\Phi}\right)^{-1} \tag{37}$$

Sectoral productivity can be written:

$$z(s) = \frac{y(s)+n(s)\phi(s)}{F(k(s),\ell(s))} = \frac{y(s)+n(s)\phi(s)}{\sum_{j=1}^{n(s)}\frac{y_j(s)+\phi(s)}{z_j(s)}} = \left(\sum_{j=1}^{n(s)}\frac{1}{z_j(s)}\frac{y_j(s)+\phi(s)}{y(s)+n(s)\phi(s)}\right)^{-1} \tag{38}$$

Notice that setting $\phi(s) = 0$, as in Edmond, Midrigan, and Xu (2021), we get the result that aggregate productivity is simply the sector-share-weighted harmonic mean of sectoral productivity.

Misallocation

Let $\widetilde{y}$ denote gross output, the sum of net output and fixed costs. The productivity aggregates are harmonic means of productivity at one level below, weighted by the gross-output-shares, for example $\frac{\widetilde{y(s)}}{\widetilde{Y}}$ in Equation 37. We can rewrite $\widetilde{y(s)}/\widetilde{Y}$ as below, using the definition of the demand function:

$$\begin{aligned}\frac{\widetilde{y(s)}}{\widetilde{Y}} &= \frac{\widetilde{y(s)}}{y(s)}\frac{Y}{\widetilde{Y}}\frac{y(s)}{Y} \\ &= \underbrace{\frac{y(s)+n(s)\phi(s)}{y(s)}}_{1+s_\phi(s)}\underbrace{\frac{Y}{Y+N\Phi}}_{(1+s_\phi)^{-1}}\left(\frac{p(s)}{P}\right)^{-\theta}\end{aligned}$$

where s_ϕ is the fixed cost share in net output.

Similarly we can rewrite $\widetilde{y_j(s)}/\widetilde{y(s)}$ as:

$$\frac{\widetilde{y_j(s)}}{\widetilde{y(s)}} = \frac{1+s_{\phi,j}(s)}{1+s_\phi(s)}\left(\frac{p_j(s)}{p(s)}\right)^{-\eta}$$

Prices take the following form:

$$p_j(s) = \mu_j(s)\lambda_j(s), \quad p(s) = \mu(s)\lambda(s)$$

where:

$$\lambda_j(s) = \frac{\Omega}{z_j(s)}, \quad \lambda(s) = \frac{\Omega}{z(s)}$$

and $\Omega = \left(\frac{r}{\alpha}\right)^\alpha \left(\frac{w}{\beta}\right)^\beta$ doesn't vary with the level of aggregation.

The gross output ratios can now be written:

$$\frac{\widetilde{y(s)}}{\widetilde{Y}} = \frac{1+s_\phi(s)}{1+s_\phi}\left(\frac{\mu(s)}{\mu}\right)^{-\theta}\left(\frac{Z}{z(s)}\right)^{-\theta}$$

$$\frac{\widetilde{y_j(s)}}{\widetilde{y(s)}} = \frac{1+s_{\phi,j}(s)}{1+s_\phi(s)}\left(\frac{\mu_j(s)}{\mu(s)}\right)^{-\eta}\left(\frac{z(s)}{z_j(s)}\right)^{-\eta}$$

Plugging these terms into our expressions, and rearranging for aggregate and sectoral productivity respectively yields the following results:

$$Z = \left(\int_0^1 z(s)^{\theta-1}\left(\frac{\mu(s)}{\mu}\right)^{-\theta}\left(\frac{1+s_\phi(s)}{1+s_\phi}\right)ds\right)^{\frac{1}{\theta-1}}$$

$$z(s) = \left(\sum_{j=1}^{n(s)} z_j(s)^{\eta-1} \left(\frac{\mu_j(s)}{\mu(s)} \right)^{-\eta} \left(\frac{1 + s_{\phi,j}(s)}{1 + s_\phi(s)} \right) \right)^{\frac{1}{\eta-1}}$$

Given returns to scale is equal to $\nu(1 + s_\phi)$, it is clear that productivity aggregates depend on the dispersion of returns to scale across firms and sectors. These productivity aggregates highlight *two channels* that can reduce productivity from first-best, as opposed to simply markup dispersion in Edmond, Midrigan, and Xu (2015) and Edmond, Midrigan, and Xu (2021). We call the extra channel **returns to scale dispersion**, which depends on $\phi(s) \neq 0 \forall s$. With positive fixed cost shares, we have an extra wedge which *reduces* aggregate productivity. This wedge represents unequal distribution of *gross* output between firms and industries.

Taking Stock

We introduce *endogenous* returns to scale in a framework which exhibits imperfect competition among heterogeneous firms, endogenous markups, and non-standard aggregation. The distinguishing features from Edmond, Midrigan, and Xu (2015) are (i) a production function with a degree of homogeneity that may differ from one, and (ii) an output-denominated fixed cost which represents lost output before firms go to the market.

3 Data & Estimation

We use data from the ARDx, which is a U.K. firm-level research dataset constructed from two ONS surveys, the Annual Business Inquiry (ABI; 1998 - 2008) and the Annual Business Survey (ABS; 2009 onwards), combined with the Business Register and Employment Survey (BRES; 2009 onwards). In sum, it contains financial and employment data for around 50,000 firms each year, with stratification by industry, size, and region to ensure a representative sample.

The ARDx is essentially a census and a survey: the former for large firms, which are

repeatedly included, and the latter for small firms, which are sampled with specific rules on inclusion to reduce the administrative burden. There are approximately 11 million workers covered by the businesses in the ARDx.

For the purpose of our production function estimation, we exclude certain non-market sectors: Agriculture, Public Sector, Finance & Insurance, Education, and Health.[5] We set out rules for SIC re-coding to ensure compatibility pre- and post-2007, when the classification is changed. For SIC codes post-2007, we simply divide the number by 1000 to match with pre-2007 codes.

We convert firm gross output into real values using the ONS experimental industry deflators.[6] Material inputs are deflated with the ONS producer price inflation data.[7]. Finally, the constructed capital stock is deflated with the ONS gross fixed capital formation deflator.[8] The summary statistics for output and inputs, overall and by sector, are contained in Table 8 of the Appendix.

To reduce the influence of outliers, which may represent measurement or recording errors in the surveys, we remove the firms with the top and bottom 1% of material inputs in each industry × year.

3.1 Production Function Estimation

Firm-level productivity estimation is computed using control function methods. These require firm-level capital stock. This section explains the Perpetual Inventory Method (PIM) used to construct a measure of capital stock, followed by the productivity estimation approaches.

The Perpetual Inventory Method (PIM) allows construction of firm-level capital stocks when such data is unavailable, but investment data is present. The method here follows Martin (2002). The PIM is constructed using the following equation:

[5] 2-digit Standard Industrial Classification (SIC) 2007 codes: A, K, O, P, Q.

[6] https://www.ons.gov.uk/economy/inflationandpriceindices/datasets/experimentalindustrydeflatorsuknonseasonallyadjusted

[7] https://www.ons.gov.uk/economy/inflationandpriceindices/datasets/producerpriceindex

[8] https://www.ons.gov.uk/economy/grossdomesticproductgdp/timeseries/ybfu/ukea

$$K_t = (1-\delta)K_{t-1} + i_t.$$

where K_t is the capital stock in period t, and i_t is investment in period t. However, to use this method, we need K_0 - the initial capital stock of a firm - which is not in this survey. To construct this series, each firm's K_0 is a revenue-weighted share of the industry-level capital stock in the first year that firm appears in the panel, as in (Hwang and Savagar 2020). Capital stock is then constructed for all future years with the above equation, with missing investment data interpolated. The depreciation rate is taken to be 18.195%, which is a weighted average of ONS depreciation rates for the three different capital categories: Building, Vehicles, Other.

Consider the production function:

$$Y_{it} = Z_{it} K_{it}^{\varepsilon_k} L_{it}^{\varepsilon_l} M_{it}^{\varepsilon_m}.$$

where Y_{it}, K_{it}, L_{it} represent revenue, capital stock, and employment respectively, while the ε's are the production elasticities to be estimated. Taking logarithms, we get:

$$y_{it} = \varepsilon_0 + \varepsilon_k k_{it} + \varepsilon_l l_{it} + \varepsilon_m m_{it} + \epsilon_{it}.$$

where $\ln z_{it} = \varepsilon_0 + \epsilon_{it}$. Firms draw productivity Z_{it} which is unobserved by the econometrician, leading to potential omitted variable bias, as clearly the optimal firm input choices will be correlated with this variable.

We briefly comment on how this links to the theoretical section. Our theoretical concept of net output $y = zF(k,\ell) - \phi$ can be mapped onto this production function quite simply: the function $F(\cdot)$ is Cobb-Douglas (with materials as an extra input, but the results still hold); z is unobserved idiosyncratic productivity for which we must control; and $y + \phi$ is the gross output measure we observe in the data.

We proceed with the control function methods of Levinsohn and Petrin (2003) and Ackerberg, Caves, and Frazer (2015), which make assumptions on the timing of in-

put choices to achieve identification, and uses investment as a proxy for unobserved productivity shocks. The full details of the estimation procedure, including how we estimate markups and time-varying elasticities, is contained in Appendix A.

4 Mapping the model to data

The estimated elasticities obtained using control function methods are *revenue elasticities*, because we only observe revenue, and not output. As discussed in the theory section, if firms have price-setting power (such that $\mu \neq 1$), then the revenue and output elasticities are not equal. In this case, simply summing the revenue elasticities gives biased estimates of returns to scale. We multiply the sum of revenue elasticities by the markup in order to obtain the correct estimate of returns to scale as in Ruzic and Ho (2019).

Consider logged Cobb-Douglas production function estimation with deflated revenue:

$$\frac{p_{it}y_{it}}{p_t} = \varepsilon_0 + \varepsilon_k k_{it} + \varepsilon_l l_{it} + \epsilon_{it}.$$

Clearly if $\frac{p_{it}}{p_t}$ doesn't vary across firms, then an appropriate instrument will identify the output elasticities. However, when firms have price-setting power, this is not the case. The estimators will be biased:

$$\varepsilon_X = \frac{\varepsilon_{YX}}{\mu}.$$

This is because the estimated coefficients are the revenue elasticities, which will be equal to the output elasticities divided by the markup.

5 Empirical Results

5.1 Returns to Scale

In this section, we present results on returns to scale in the U.K. economy between 1998 and 2014. Our preferred estimation procedure follows Ackerberg, Caves, and Frazer (2015), with value-added as the dependent variable, and a translog production function. This approach uses materials inputs as an instrument, for which most firms have available data, unlike investment (see Figure ?? in the Appendix). In addition, the translog assumption is both more flexible, and permits the estimation of time-varying elasticities. Variations from our preferred methodology produce similar estimates, which are reported.

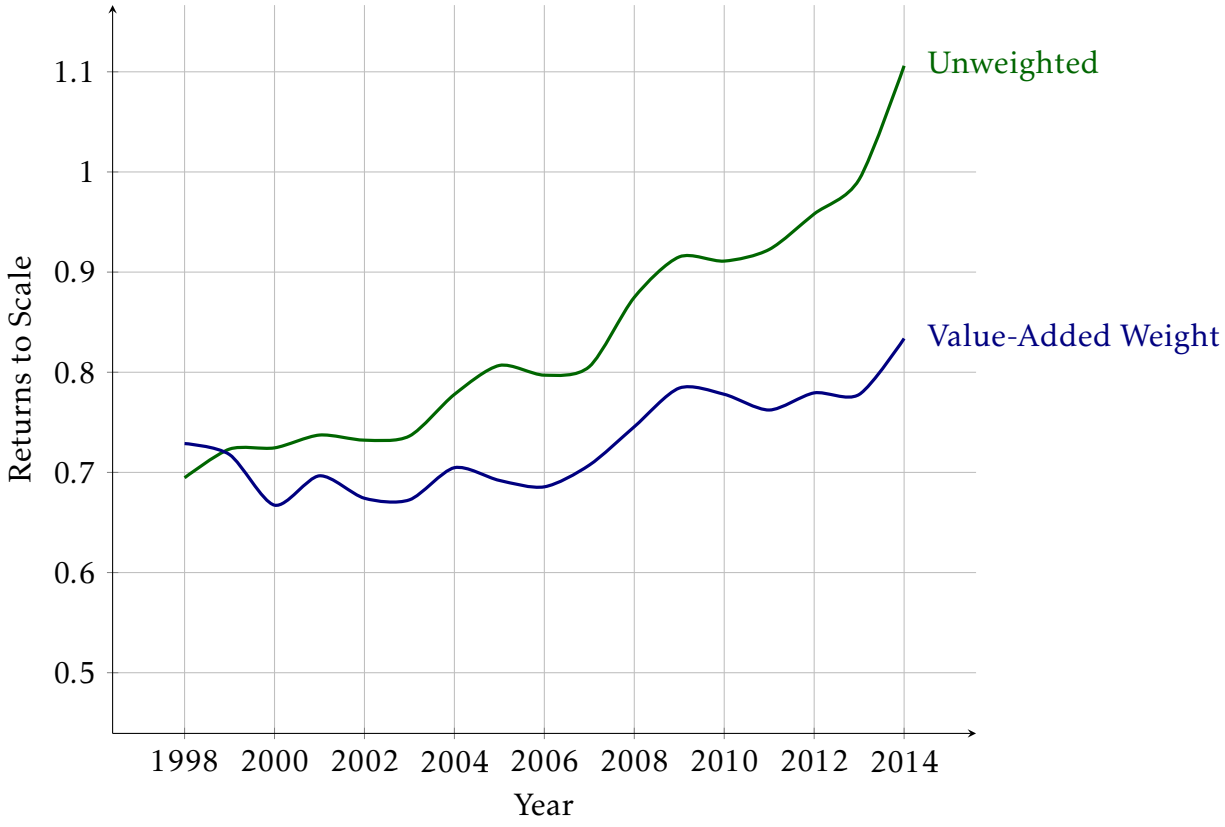


Figure 2: Aggregate Returns to Scale.

Recall that returns to scale can be described as the markup multiplied by one minus the profit share: RTS = $\mu(1 - s_\pi)$. There is significant evidence of rising markups in the U.K. over this time period (Hwang and Savagar 2020). Thus, we expect a rise in

returns to scale, assuming that profit shares have not risen dramatically (as seems to be the case, shown in Figure 22 in the Appendix).

Figure 2 presents our estimates of returns to scale across the whole economy. There are two key findings. Firstly, our results for the U.K. are that returns to scale are below unity, across a range of estimation methods. Secondly, the levels have risen over time, towards constant returns to scale. Both of these results generally hold across sectors, with further details in Appendix B.

Across the economy, we find returns to scale rising from values around 0.7 in 1998, to 0.8 - 1.1 by 2014, depending on the weighting scheme, production function used, and aggregation method.

Returns to scale is informative about the ability of firms to scale up output as they adjust inputs. To the extent that more productive firms are larger, we would expect returns to scale and firm size to be related similarly, as discussed in Section **??**.

5.2 Productivity Estimation

We showed in Section 3.1 how TFP can be estimated using the control function methods. Figure 3 shows the steady rise in productivity up to 2008, followed by the plateauing from the late 2000s. This result is robust across estimation methods (Olley and Pakes 1996; Levinsohn and Petrin 2003; Ackerberg, Caves, and Frazer 2015, e.g.). The general trend also holds across macro sectors, although the flattening of productivity growth is least noticeable in Services (see Figure 14 in the Appendix).

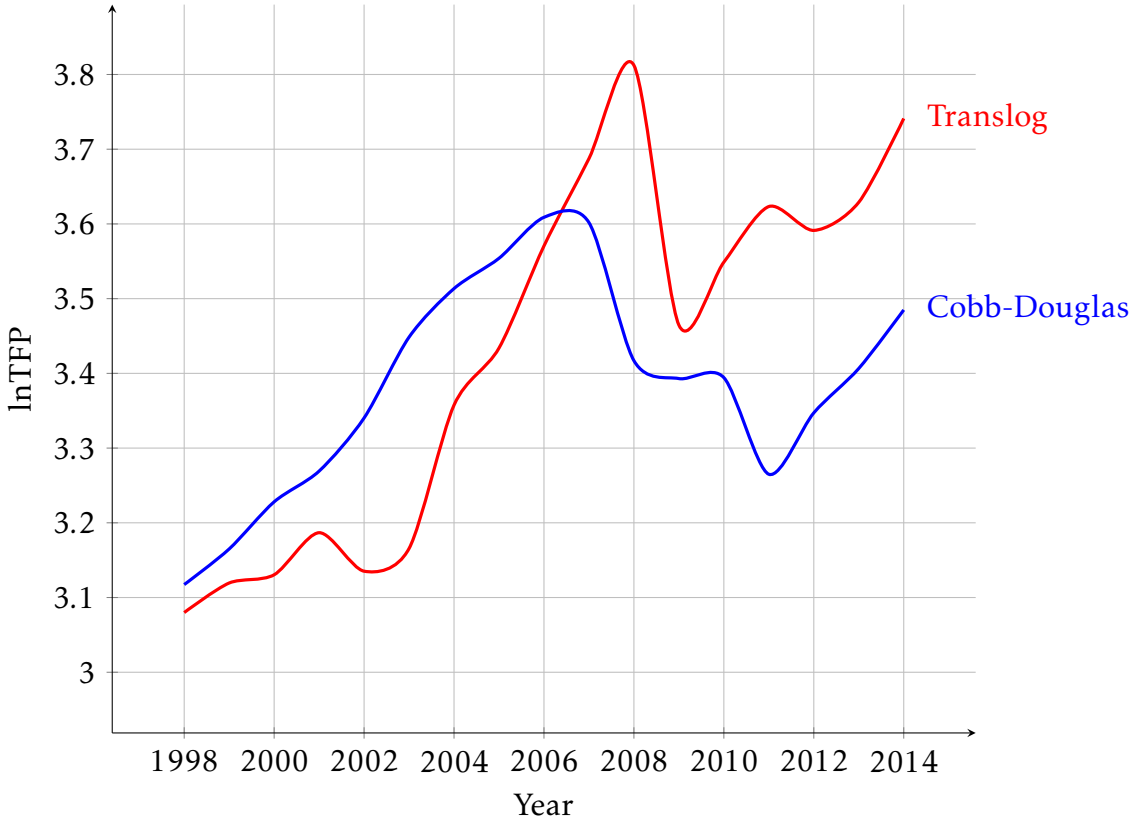


Figure 3: Aggregate TFP.

At the aggregate level, returns to scale have risen, as has productivity. However, the trends are different: productivity experienced gradual growth until around the Great Financial Crisis, followed by a slowdown. On the other hand, returns to scale followed a different trajectory, with quite limited change up to the mid-2000s, followed by a noticeable rise in the last decade. Clearly it is too simplistic to consider these aggregate trends alone. As discussed in Section **??**, the relationship between productivity and returns to scale is not straightforward.

To investigate this relationship further, we present results from a firm-level fixed-effects panel regression of returns to scale on log TFP in Table 1.

Simply looking at the correlation between returns to scale and productivity produces an estimate barely different from zero. Both of these variables exhibit trends at the aggregate level, but a year fixed effect doesn't move the needle much on the estimated relationship at the firm level. However, once we include firm fixed effects, we get a robust negative relationship between returns to scale and productivity between

Table 1: Regression: Returns to Scale on Productivity at Firm Level

	Dependent variable: Returns to Scale				
Log TFP	−0.003	0.007	−0.915***	−0.944***	−0.635***
	(0.026)	(0.026)	(0.094)	(0.093)	(0.053)
N	423,337	423,337	260,951	260,951	256,966
Year FE:		✓		✓	✓
Firm FE:			✓	✓	✓
Remove outliers:					✓

*Note: Estimates statistically significant at levels of 0.1%: ***, 1%: **, 5%: *. Robust standard errors clustered at the level of the fixed effects included. Weighted by value-added at the firm level. Outliers are the top and bottom 1% of firms by returns to scale.*

firms. Removing outlier firms (in terms of returns to scale) does not materially affect the result. This confirms the prediction from our endogenous returns to scale theory, where RTS $= \nu(1 + s_\phi)$: returns to scale is falling in productivity, if fixed costs are held constant. We find that *within firms*, those that become more productive experience falling returns to scale over time.

There is, however, evidence that returns to scale and productivity are positively correlated *across industries*. This is in line with findings by Gao and Kehrig (2021), and standard static models of heterogeneous firms with a productivity cut-off and exogenous returns to scale. We also find this result, as documented in Table 2. Industries with higher average productivity also exhibit higher returns to scale.

Table 2: Regression: Returns to Scale on Productivity at Industry Level

	Dependent variable: Returns to Scale			
Mean log TFP	0.191***	0.211***	0.125***	0.228***
	(0.014)	(0.013)	(0.018)	(0.021)
N	1,009	1,009	1,009	1,009
2-digit SIC FE:			✓	✓
Year FE:		✓		✓

*Estimates statistically significant at levels of 1%: ***, 5%: **, 10%: *. Robust standard errors clustered at the level of the 2-digit SIC.*

Therefore, we have a paradox! Within firms, becoming more productive leads to

a fall in scale economies. Across industries, this relationship is flipped. And at the aggregate level, a slowdown in the productivity *growth* has coincided with a gradual rise in returns to scale. This suggests that reallocation across firms has had significant aggregate implications. We use our model to try to reconcile these facts.

6 Model Calibration

Productivity $z_j(s)$ is made up of a firm-specific and sector-specific component, as in Edmond, Midrigan, and Xu (2015):

$$z_j(s) = a(s)x_j(s),$$

where $a(s) \geq 1$ is draw from an i.i.d. Pareto distribution with shape parameter $\xi_a > 0$ across sectors. Within each sector, $x_j(s) \geq 1$ are i.i.d. Pareto draws across firms with shape parameter $\xi_x > 0$.

The output-denominated fixed cost $\phi(s)$ varies across sectors. The distribution across sectors is exponential, with mean parameter ξ_ϕ, which is calibrated to match returns to scale.

Table 3: Calibration Parameters

	Parameter	**Value**	**Target**
Assigned			
ρ	Discount rate	0.96	Match annual IR
α	Capital elasticity	0.25	Production function estimation
β	Labour elasticity	0.45	Production function estimation
δ	Depreciation rate	0.08	ONS
θ	Across-sector elasticity of substitution	1.2223	Regress markups on market shares
Calibrated			
ξ_ϕ	Exponential mean, fixed cost	54.0	Match returns to scale
ξ_a	Pareto shape, sector productivity	0.45	Distribution of sectoral shares
ξ_x	Pareto shape, idiosyncratic productivity	7.19	Within-sector concentration
η	Within-sector elasticity of substitution	11.50	Match markups

The model replicates the data well. Table 4 summarises a host of moments, and indicates which were targeted in the calibration procedure and which were not.

Results

Table 4 shows that the average markup is increasing as we move from the firm-level to the sectoral level to the aggregate. This is intuitive, as the markup is always rising in firm size, and higher levels of aggregation are computed using a revenue-weighted harmonic mean.

Table 4: Moments in Data and Model

	Model	**Data**
Targeted		
Aggregate Returns to Scale	0.766	0.820
Aggregate Markup	1.263	1.441
Concentration Ratios:		
CR5	0.081	0.080
CR10	0.120	0.120
CR20	0.167	0.170
CR50	0.244	0.252
HHI	940	940
Mean market share	0.05	0.01
Median market share	0.01	0.006
Sectoral Returns to Scale Percentiles:		
10^{th}	0.70	0.49
25^{th}	0.71	0.57
50^{th}	0.73	0.70
75^{th}	0.82	0.80
90^{th}	0.94	0.94
Non-targeted		
Mean firm-level Returns to Scale	0.702	0.734
Mean sectoral Returns to Scale	0.700	0.725
Mean firm-level Markup	1.132	1.189
Mean sectoral Markup	1.155	1.220

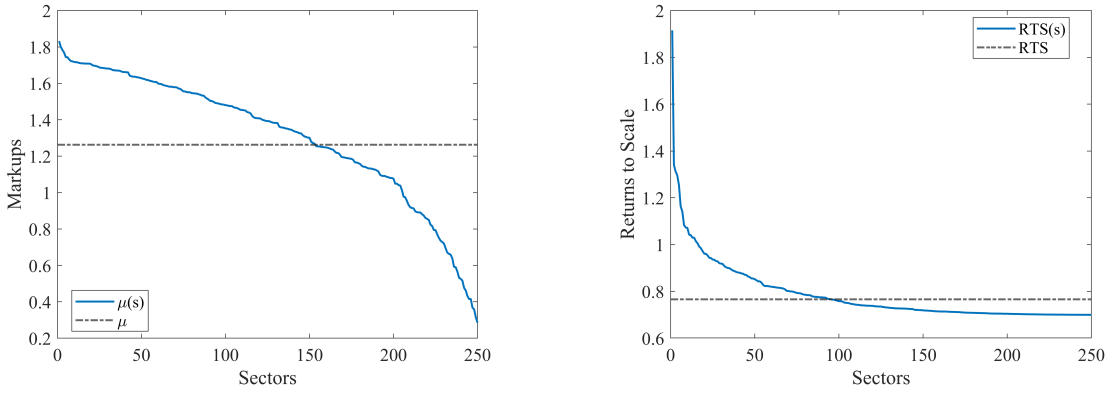
On the other hand, this relationship is not replicated for returns to scale. At the firm-level, returns to scale is decreasing in firm size. However, at the sector level, returns to scale rises with sectoral output, due to the *selection effect* described in the model. Therefore, we should expect the average firm-level returns to scale to exceed the average sector-level returns to scale, but both to be lower than aggregate returns to scale.

Figures 4 show the distribution of markups and returns to scale over sectors, and the aggregate levels.

Productivity and Misallocation

Productivity is quite fat-tailed, at the firm and sector levels, due to the Pareto calibration. Figure 5 shows the distribution of log productivity over sectors, with the aggregate plotted to show the extent of the variation.

Misallocation occurs through two channels. The total loss - from the efficient



(a) Markups (b) Returns to Scale

Figure 4: Returns to Scale and Markup Distributions.

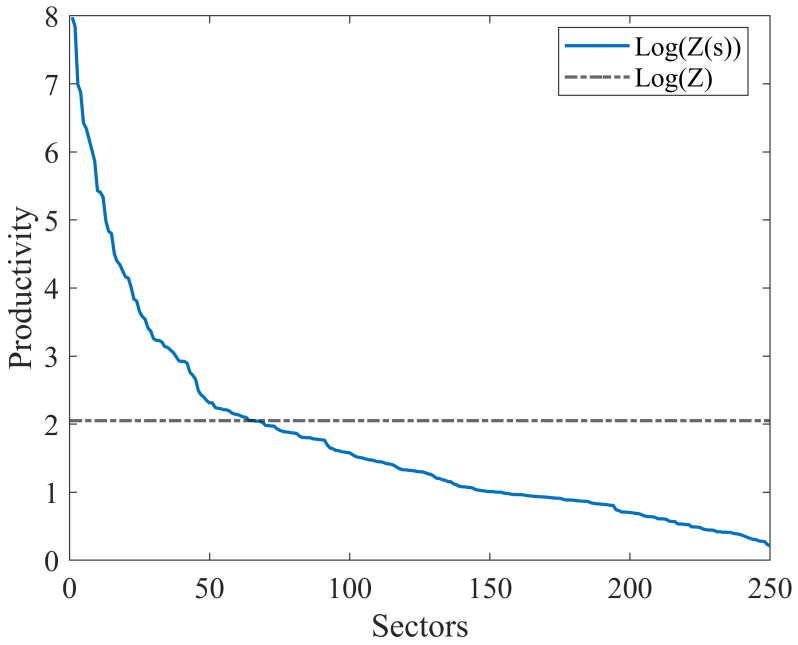


Figure 5: Productivity distribution.

benchmark - is around 26.6%. This can be decomposed into the two components:

1. Markup dispersion: 16.1%

2. Returns to scale dispersion: 10.7%

The model indicates that misallocation across heterogeneous firms reduces total productivity due to *both* markup and returns to scale dispersion. Both are quantitatively important channels. As markups and returns to scale have risen from 1994 - 2014, these are two potential explanations for the fall in productivity growth in the

UK over this time period. Further quantitative exercises with this model will focus on this issue.

7 Conclusion

We describe the theory to help understand and measure returns to scale. We highlight two ways to understand this concept: as a function of technical parameters of the production function, and as a function of behaviour outcomes from strategic interactions.

Our contribution to understanding returns to scale in the U.K. is threefold: (1) we present updated estimates of scale economies across a services-dominated economy from 1998 - 2014, and report results at various levels of aggregation, (2) we estimate time-varying returns, and present evidence of a rise over time, (3) we simultaneously estimate time-varying markups, and adjust our revenue elasticity estimates to correctly describe returns to scale.

We find that most industries have decreasing returns to scale, with significant heterogeneity across both services and manufacturing. We report a rise in returns to scale over time, especially since the mid-2000s. We show a strong negative relationship between scale economies and productivity within firms over time. However, across industries, the relationship is positive. These results are consistent with an endogenous returns to scale theory, hinging on an output-denominated fixed cost, alongside a standard productivity cut-off with fixed costs varying across industries. We combine these elements in a heterogeneous firm model with non-unity production function homogeneity and the presence of a fixed cost to drive endogenous returns to scale. We match the model to the data and show that the model can reproduce the relationships between markups, returns to scale, and productivity across different levels of aggregation.

References

Ackerberg, Daniel A., Kevin Caves, and Garth Frazer (2015). "Identification Properties of Recent Production Function Estimators". In: *Econometrica* 83.6, pp. 2411–2451.

Aghion, Philippe, Antonin Bergeaud, Timo Boppart, Peter J. Klenow, and Huiyu Li (Nov. 2019). *A Theory of Falling Growth and Rising Rents*. NBER Working Papers 26448. National Bureau of Economic Research, Inc.

Aghion, Philippe, Nick Bloom, Richard Blundell, Rachel Griffith, and Peter Howitt (2005). "Competition and innovation: An inverted-U relationship". In: *The Quarterly Journal of Economics* 120.2, pp. 701–728.

Altug, Sumru and Alpay Filiztekin (2002). "Scale effects, time-varying markups, and the cyclical behaviour of primal and dual productivity". In: *Applied Economics* 34.13, pp. 1687–1702.

Autor, David, David Dorn, Lawrence F Katz, Christina Patterson, and John Van Reenen (2017). "Concentrating on the Fall of the Labor Share". In: *American Economic Review* 107.5, pp. 180–85.

Baqaee, David, Emmanuel Farhi, and Kunal Sangani (May 2020). *The Darwinian Returns to Scale*. Working Paper 27139. National Bureau of Economic Research.

Barkai, Simcha (2020). "Declining labor and capital shares". In: *The Journal of Finance* 75.5, pp. 2421–2463.

Basu, Susanto and John Fernald (1997). "Returns to scale in US production: Estimates and implications". In: *Journal of political economy* 105.2, pp. 249–283.

Benhabib, Jess and Roger Farmer (1994). "Indeterminacy and Increasing Returns". In: *Journal of Economic Theory* 63.1, pp. 19–41.

Bessen, James (2020). "Industry Concentration and Information Technology". In: *The Journal of Law and Economics* 63.3, pp. 531–555.

Burnside, Craig (Apr. 1996). "Production function regressions, returns to scale, and externalities". In: *Journal of Monetary Economics* 37.2-3, pp. 177–201.

Caballero, Ricardo J and Richard K Lyons (1992). "External effects in US procyclical productivity". In: *Journal of Monetary Economics* 29.2, pp. 209–225.

De Loecker, Jan, Jan Eeckhout, and Simon Mongey (2021). *Quantifying market power and business dynamism in the macroeconomy*. Tech. rep. Working paper (Sept 2021).

De Loecker, Jan and Frederic Warzynski (2012a). "Markups and firm-level export status". In: *The American Economic Review* 102.6, pp. 2437–2471.

— (2012b). "Markups and firm-level export status". In: *American Economic Review* 102.6, pp. 2437–71.

De Ridder, Maarten (Mar. 2019). *Market Power and Innovation in the Intangible Economy*. Discussion Papers 1907. Centre for Macroeconomics (CFM).

Edmond, Chris, Virgiliu Midrigan, and Daniel Yi Xu (2015). "Competition, markups, and the gains from international trade". In: *American Economic Review* 105.10, pp. 3183–3221.

— (May 2021). *How Costly Are Markups?* NBER Working Papers 24800. National Bureau of Economic Research, Inc.

Eggertsson, Gauti B., Jacob A. Robbins, and Ella Getz Wold (2021). "Kaldor and Piketty's facts: The rise of monopoly power in the United States". In: *Journal of Monetary Economics* 124. The Real Interest Rate and the MarginalProduct of Capital in the XXIst CenturyOctober 15-16, 2020, S19–S38.

Farmer, Roger and Jang-Ting Guo (June 1994). "Real Business Cycles and the Animal Spirits Hypothesis". In: *Journal of Economic Theory* 63.1, pp. 42–72.

Feenstra, Robert C (2003). *Advanced International Trade: Theory and Evidence*. Princeton University Press.

Gao, Wei and Matthias Kehrig (July 2021). "Returns to Scale, Productivity and Competition: Empirical Evidence from US Manufacturing and Construction Establishments". In: *Working Paper (Jul 2021)*.

Girma, S. and H. Görg (2002). "Foreign Ownership, Returns to Scale and Productivity: Evidence from UK Manufacturing Establishments". In: *CEPR Discussion Paper Series*.

Gutierrez, German, Callum Jones, and Thomas Philippon (Feb. 2019). *Entry Costs and the Macroeconomy*. Working Paper 25609. National Bureau of Economic Research.

Gutiérrez, Germán, Callum Jones, and Thomas Philippon (2021). "Entry costs and aggregate dynamics". In: *Journal of Monetary Economics* 124. The Real Interest Rate and the MarginalProduct of Capital in the XXIst CenturyOctober 15-16, 2020, S77–S91.

Hall, Robert (1988). "The relation between price and marginal cost in US industry". In: *Journal of political Economy* 96.5, pp. 921–947.

— (1990). "Invariance Properties of Solow's Productivity Residual". In: *Growth, Productivity, Unemployment: Essays to Celebrate Bob Solow's Birthday*. MIT Press, pp. 71–112.

Harris, Richard and Eunice Lau (Apr. 1998). "Verdoorn's law and increasing returns to scale in the UK regions, 1968–91: some new estimates based on the cointegration approach". In: *Oxford Economic Papers* 50.2, pp. 201–219.

Harrison, Ann E (1994). "Productivity, imperfect competition and trade reform: Theory and evidence". In: *Journal of international Economics* 36.1-2, pp. 53–73.

Hopenhayn, Hugo (1992). "Entry, Exit, and Firm Dynamics in Long Run Equilibrium". In: *Econometrica* 60.5, pp. 1127–1150.

Hwang, Kyung-In and Anthony Savagar (2020). "Product Market Power and TFP". In: *Working Paper*.

Kee, Hiau Looi (Oct. 2004). "Estimating Productivity When Primal and Dual TFP Accounting Fail: An Illustration Using Singapore's Industries". In: *The B.E. Journal of Economic Analysis & Policy* 4.1, pp. 1–40.

Kim, Jinill (2004). "What determines aggregate returns to scale?" In: *Journal of Economic Dynamics and Control* 28.8, pp. 1577–1594.

Lashkari, Danial, Arthur Bauer, and Jocelyn Boussard (2019). *Information Technology and Returns to Scale*. 2019 Meeting Papers 1380. Society for Economic Dynamics.

Levinsohn, James (1993). "Testing the imports-as-market-discipline hypothesis". In: *Journal of International Economics* 35.1-2, pp. 1–22.

Levinsohn, James and Amil Petrin (2003). "Estimating production functions using inputs to control for unobservables". In: *The Review of Economic Studies* 70.2, pp. 317–341.

Martin, Ralf (2002). *Building the capital stock*. CeRiBA Working Paper. The Centre for Research into Business Activity.

Melitz, Marc J (Nov. 2003). "The impact of trade on intra-industry reallocations and aggregate industry productivity". In: *Econometrica* 71.6, pp. 1695–1725.

Olley, G. Steven and Ariel Pakes (1996). "The dynamics of productivity in the telecommunications equipment industry". In: *Econometrica* 64.6, pp. 1263–1297.

Oulton, Nicholas (1996). "Increasing Returns and Externalities in UK Manufacturing: Myth or Reality?" In: *The Journal of Industrial Economics* 44.1, pp. 99–113.

Restuccia, Diego and Richard Rogerson (2008). "Policy distortions and aggregate productivity with heterogeneous establishments". In: *Review of Economic dynamics* 11.4, pp. 707–720.

Rotemberg, Julio J. and Michael Woodford (Oct. 1993). *Dynamic General Equilibrium Models with Imperfectly Competitive Product Markets*. Working Paper 4502. National Bureau of Economic Research.

— (1999). "The cyclical behavior of prices and costs". In: *Handbook of Macroeconomics*. Ed. by J. B. Taylor and M. Woodford. Vol. 1. Handbook of Macroeconomics. Elsevier. Chap. 16, pp. 1051–1135.

Ruzic, Dimitrije and Sui-Jade Ho (Aug. 2019). "Returns to Scale, Productivity Measurement, and Trends in U.S. Manufacturing Misallocation". In: *INSEAD working paper*.

Van Reenen, John (2018). "Increasing differences between firms: market power and the macro-economy". In: *CEP Discussion Papers*.

A Production Function Estimation

Consider the production function:

$$y_{it} = \varepsilon_0 + \varepsilon_k k_{it} + \varepsilon_l l_{it} + \varepsilon_m m_{it} + \epsilon_{it}.$$

where $\ln z_{it} = \varepsilon_0 + \epsilon_{it}$. We split up the unobserved residual $\epsilon_{it} = \omega_{it} + \eta_{it}$, where ω_{it} is anticipated and η_{it} is an ex-post shock. Thus, inputs are correlated with ω_{it} only. The following assumptions are required:

1. **Information Sets:** firms' information sets I_{it} include current and past productivity shocks $\{\omega_{i\tau}\}_{\tau=0}^{t}$, but firms know nothing about future shocks. The ex-post shocks η_{it} are expected to be zero on average: $\mathbb{E}\{\eta_{it}|I_{it}\} = 0$.

2. **First-Order Markov Shocks:** productivity shocks follow a First-Order Markov Process, so $\omega_{it} = \mathbb{E}(\omega_{it}|\omega_{i,t-1}) + \nu_{it}$, and $\mathbb{E}\{\nu_{it}|I_{it-1}\} = 0$.

3. **Timing of Input Choices:** in the previous period $i_{i,t-1}$ determines capital in the current period k_{it}, whereas labour is chosen in the current period.

4. **Scalar Unobservable:** investment decisions $i_{it} = f_t(k_{it}, \omega_{it})$ have just one scalar unobservable ω_{it}, so there is no other across firm unobserved heterogeneity (e.g. adjustment costs, investment efficiency, input prices).

5. **Strict Monotonicity:** investment decisions are strictly monotonic in the scalar unobservable ω_{it}, so $i_{it} = f_t(k_{it}, \omega_{it})$.

Given that investment is strictly monotonic in the unobserved anticipated shock, this function can be inverted, and then substituted into the production function:

$$y_{it} = \varepsilon_0 + \varepsilon_k k_{it} + \varepsilon_l l_{it} + \varepsilon_m m_{it} + f_t^{-1}(k_{it}, i_{it}) + \eta_{it}.$$

This inverted function is unknown, so is approximated by a polynomial in capital and investment:

$$y_{it} = \varepsilon_0 + \varepsilon_k k_{it} + \varepsilon_l l_{it} + \varepsilon_m m_{it} + \vartheta_0 + \vartheta_1 k_{it} + \vartheta_2 i_{it} + \vartheta_3 k_{it}^2 + \vartheta_4 i_{it}^2 + \vartheta_5 i_{it} k_{it} + \eta_{it}.$$

and estimation takes the standard two-step process. The first-step OLS regression over the above equation yields an estimate $\widehat{\varepsilon_l}$ and an estimate of the "composite" term $\widehat{\Phi}_{it} = \varepsilon_0 + \widehat{\varepsilon_k} k_{it} + \omega_{it}$. To estimate ε_k, we calculate 'implied' ω_{it}'s:

$$\widehat{\omega}_{it}(\varepsilon_k) = \widehat{\Phi}_{it} - \widehat{\varepsilon_k} k_{it}.$$

Then, by the First-Order Markov Process of the productivity shocks, we can non-parametrically regress the implied $\widehat{\omega}_{it}(\varepsilon_k)$'s on their lag $\widehat{\omega}_{it-1}(\varepsilon_k)$, and the residuals $\widehat{\nu}_{it}(\varepsilon_k)$ are the implied innovations in productivity. Finally, the sample analogue of the moment condition $\mathbb{E}\{\nu_{it} k_{it}\} = 0$ is:

$$\frac{1}{N}\frac{1}{T}\sum_i \sum_t \widehat{\nu}_{it}(\widehat{\varepsilon_k}) k_{it} = 0.$$

and we find $\widehat{\varepsilon_k}$ to solve this problem.

Levinsohn and Petrin (2003) use material inputs as a proxy instead of investment. However, there is a potential collinearity problem, as highlighted by Ackerberg, Caves, and Frazer (2015), such that identification of ε_l is not possible. Thus, they use a value-added production function and adjust the timing assumptions, so labour is chosen before material inputs, but after investment.

We estimate returns to scale by summing the estimated output elasticities $\widehat{\varepsilon_k}$ and $\widehat{\varepsilon_l}$. This is computed across the whole sample, but also on four 'macro sectors', and disaggregated by 2-digit industries. Other research looking at U.S. manufacturing has found the returns to scale to be in the range of 0.95 (e.g. Ruzic and Ho 2019; Gao and Kehrig 2021), implying diminishing returns, although industrial heterogeneity gives results both above and below this value.

This estimation procedure also allows us to extract firm-level productivity estimates, and markups. Productivity is computed from the estimated ω_{it}, and we aggre-

gate over industries to investigate productivity trends. Markups are calculated using the method of De Loecker and Warzynski (2012b):

$$\mu_{it} = \frac{\widehat{\varepsilon_X}}{\alpha_{it}^X}. \tag{39}$$

The markup is the ratio of the output elasticity of flexible input X to the expenditure share of that input in total sales. The numerator is estimated using the aforementioned control function approach, and the denominator is an observed ratio in the data.

In order to obtain returns to scale at the level of the firm and year, a slightly different estimation procedure is required. We need to estimate time- and firm-specific output elasticities $\theta_{it}^k, \theta_{it}^l$. This can be achieved by generalising the production function to translog, as in De Loecker and Warzynski (2012b):

$$y_{it} = \varepsilon_0 + \varepsilon_k k_{it} + \varepsilon_l l_{it} + \varepsilon_{ll} l_{it}^2 + \varepsilon_{kk} k_{it}^2 + \varepsilon_{lk} l_{it} k_{it} + \epsilon_{it}.$$

Then follow the same procedure as used with a Cobb-Douglas production function to estimate the coefficients and productivity process. The time- and firm-specific output elasticities are easily computed:

$$\widehat{\theta_{it}^k} = \widehat{\varepsilon_k} + 2\widehat{\varepsilon_{kk}} k_{it} + \widehat{\varepsilon_{lk}} l_{it}.$$

$$\widehat{\theta_{it}^l} = \widehat{\varepsilon_l} + 2\widehat{\varepsilon_{ll}} l_{it} + \widehat{\beta_{lk}} k_{it}.$$

Allowing heterogeneity across firms and over time in the output elasticities permits computation of distributions of returns to scale, and tracking the changes over time.

B Returns to Scale Estimates

Economy-wide returns to scale over almost two decades are below unity. Results are included in Table 5 across different sectors of the economy. These are estimates for

constant markups and constant revenue elasticities over the time period.

Table 5: Returns to Scale Estimates using Levinsohn and Petrin (2003)

	Manufacturing	*Construction*	*Trade, Wholesale + Transport*	*Services*
RTS	0.682	0.651	0.742	0.785
N	120,7142	51,786	181,985	138,011

Estimated RTS using Cobb-Douglas production function, Levinsohn and Petrin (2003) control function method.

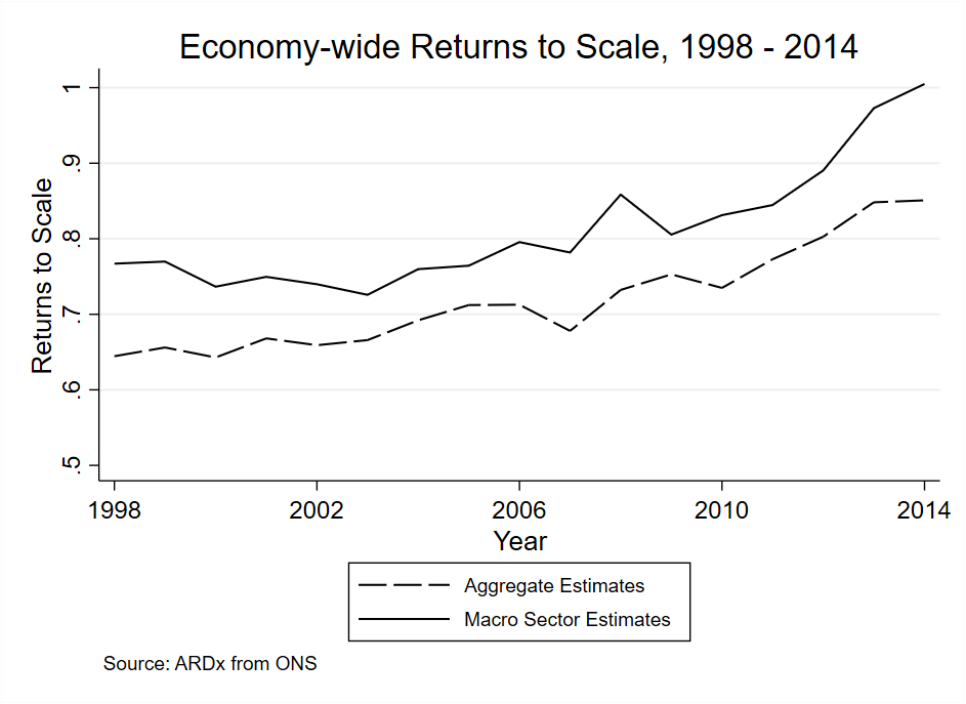


Figure 6: Economy-wide returns to scale estimates from a Cobb-Douglas production function using the Levinsohn and Petrin (2003) approach, with two different levels of aggregation (i.e. the level at which production function was estimated, before compiling a weighted average).

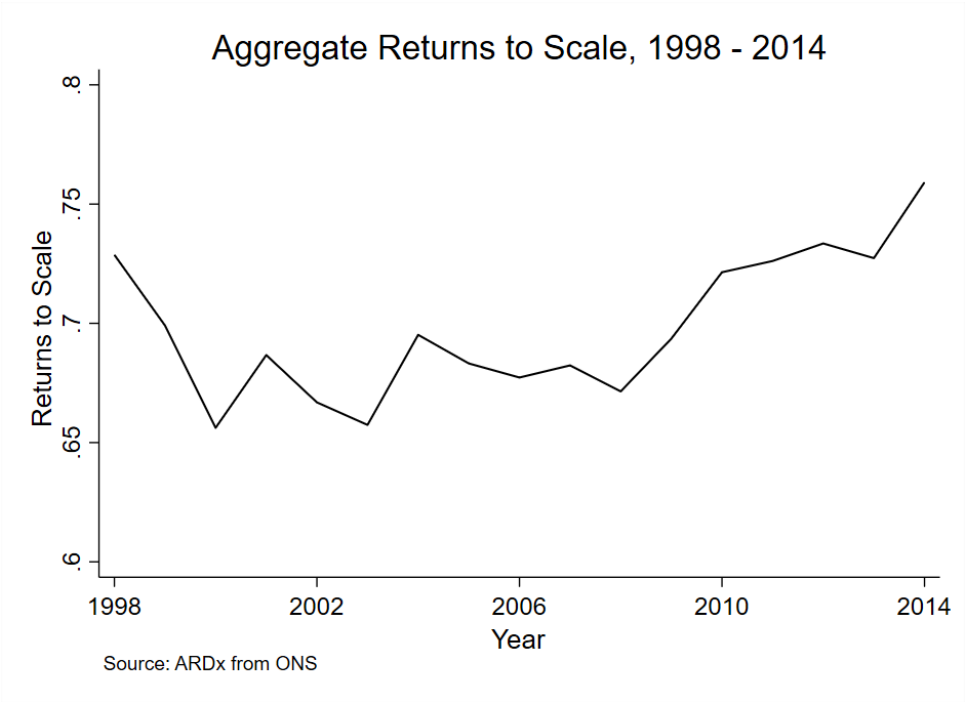


Figure 7: Economy-wide returns to scale estimates from a translog production function using the Levinsohn and Petrin (2003) approach.

The sectoral differences in returns to scale are informative about the systematic long-run differences in industry characteristics. We are interested in both the levels and the changes over time. We will focus on the time-varying estimates here in Figure 8. The patterns broadly match our constant estimates from Table 5: Manufacturing and Construction have lower scale economies than Services and Trade, Wholesale + Transport. However, the time-varying estimates highlight some interesting trends. For Manufacturing, returns to scale fell quite substantially in the early years from 0.67 in 1998 to 0.54 by 2002, before rebounding somewhat, and eventually plateauing at around 0.55. This story is not replicated in other sectors, where there was an overall rise in scale economies in the latter half of the sample. In Construction firms, returns to scale shot up from around 0.5 to 0.7 in the early 2000s, and has stayed there. Both Services and Trade, Wholesale + Transport fluctuated in the region 0.7 - 0.8, before experiencing a gradual rise in the last ten years of the data.

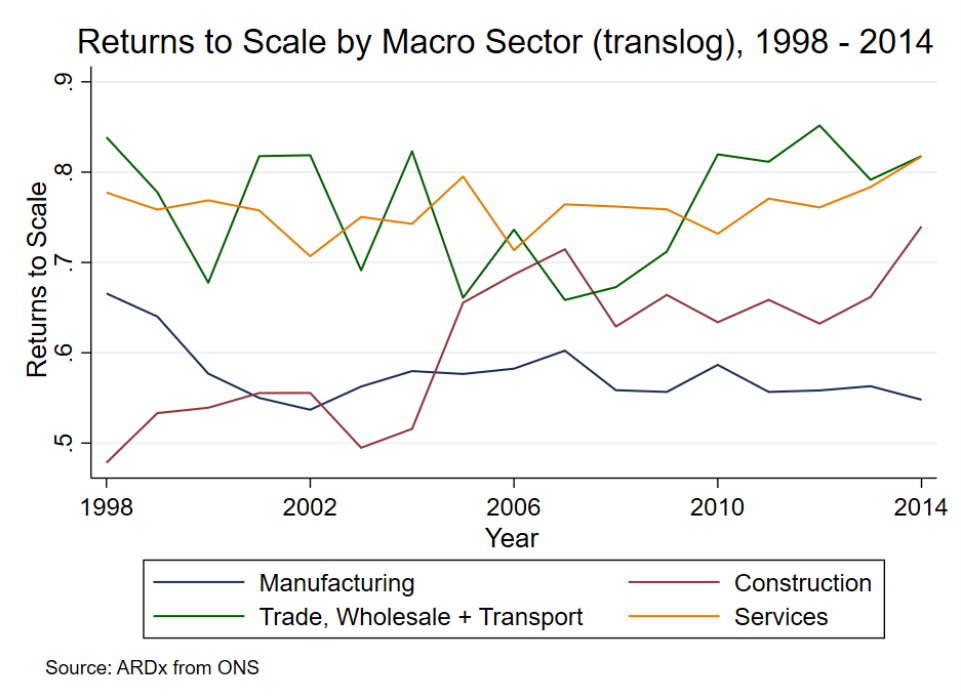


Figure 8: Macro sector returns to scale estimates from a translog production function using the Levinsohn and Petrin (2003) approach.

We explore returns to scale at one lower level of aggregation: 2-digit SICs. Figure 9 presents the time-varying estimates across industries from 1998 - 2014. The considerable across-industry heterogeneity is clear, but the within-industry fluctuations obscure the general upwards trend in scale economies. Therefore, we also highlight the (firm-count) weighted-average in the bold dotted line, which pretty much matches up with the aggregate translog estimate in Figure **??**: economy-wide returns to scale sit between 0.66 and 0.73, and we see a gradual rise towards the end of the sample period. The estimates using Cobb-Douglas production functions are in Figure 12, which shows even more substantial across-industry heterogeneity, due to the upwards-biased markup estimates arising from fixed materials-output elasticities.

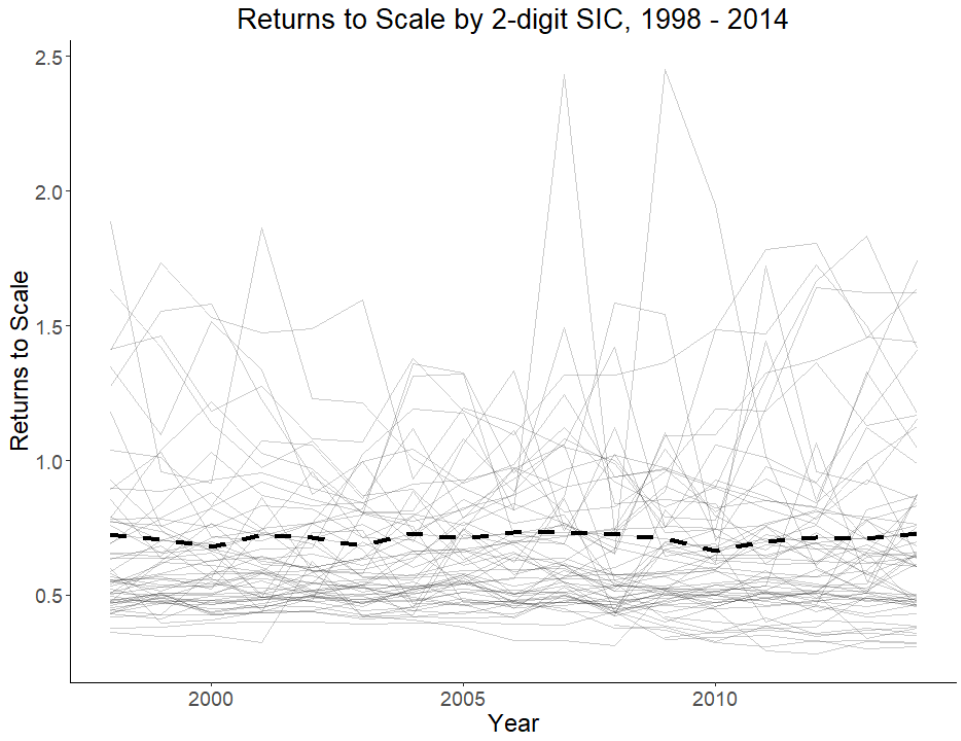


Figure 9: 2-digit SIC returns to scale estimates from a translog production function using the Levinsohn and Petrin (2003) approach.

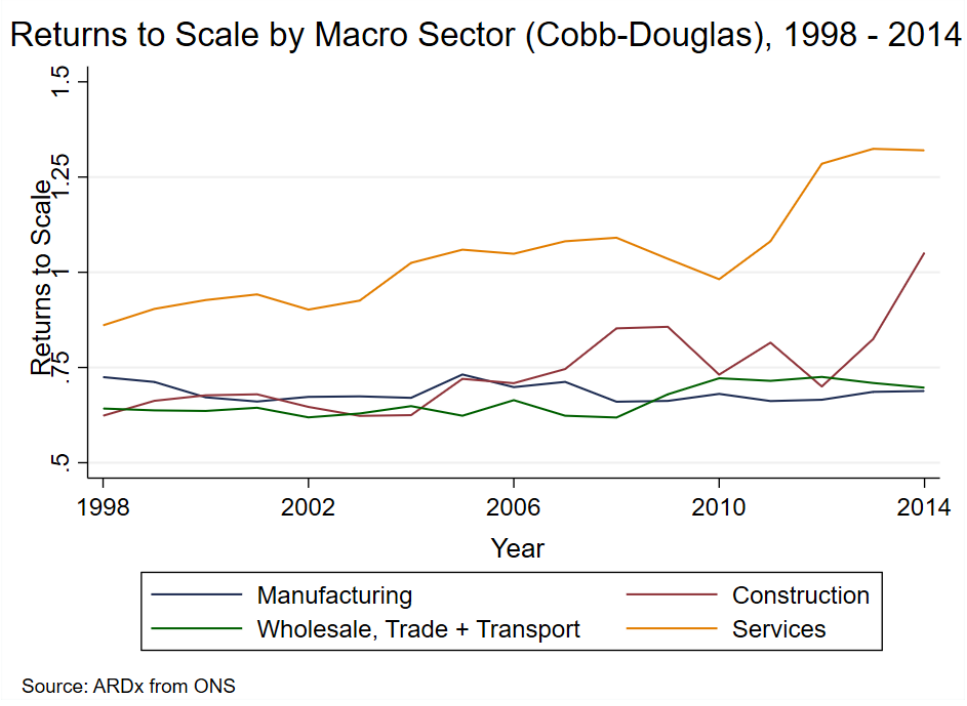


Figure 10: Macro sector returns to scale estimates from a Cobb-Douglas production function using the Levinsohn and Petrin (2003) approach.

Estimates of scale economies by 2-digit SIC can be found in Table 6 below. We

plot these results in Figure 11, which highlight the sectoral heterogeneity of scale economies. Across Manufacturing sectors, there are four in the range of constant returns to scale: Computer, Electronic, & Optical; Basic Metals; Coke and Refined Petroleum; Repair and Installation of Machinery.

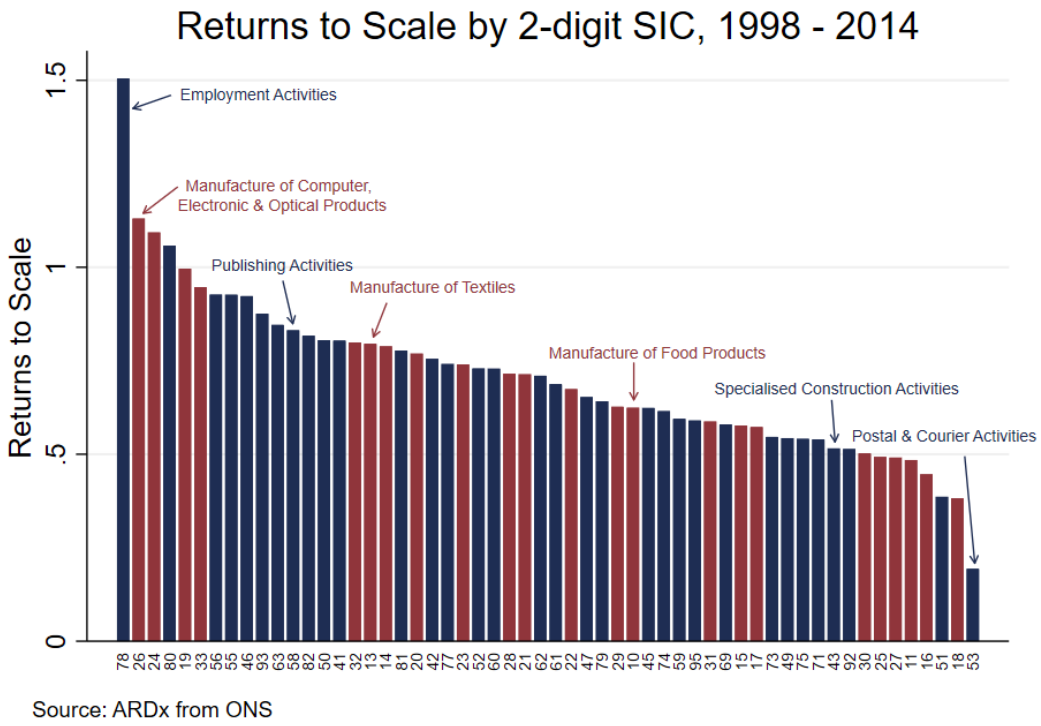


Figure 11: Returns to scale estimates from estimating a constant revenue elasticity and constant markup using Levinsohn and Petrin (2003) for each 2-digit SIC. Manufacturing sectors are coloured maroon.

This section contains returns to scale estimates by industry, at the 2-digit SIC level. The number of firms on which estimation was computed is included. The standard error on the RTS estimate is computed using the delta method, and these are in brackets below the RTS estimate. If the factor elasticity on labour or capital was outside the range of $[0,1]$, then the RTS was not computed.

SIC	N	RTS
10	12,495	0.624
11	1,724	0.485
13	4,981	0.795
14	3,355	0.789
15	841	0.577
16	3,478	0.446
17	4,184	0.573
18	7,521	0.382
19	506	0.996
20	5,733	0.769
21	986	0.714
22	7,776	0.674
23	5,616	0.74
24	4,776	1.093
25	15,597	0.493
26	7,648	1.13
27	4,913	0.491
28	10,899	0.715
29	1,633	0.627
30	1,973	0.502
31	4,060	0.587
32	5,020	0.799
33	4,997	0.947
41	12,218	0.804
42	12,554	0.755
43	27,014	0.516
45	24,639	0.624
46	68,969	0.923

SIC	N	RTS
47	66,171	0.653
49	11,501	0.543
50	1,306	0.805
51	807	0.386
52	8,103	0.729
53	489	0.194
55	8,549	0.927
56	25,219	0.927
58	6,802	0.832
59	2,548	0.595
60	693	0.729
61	1,062	0.688
62	9,061	0.71
63	1,224	0.846
69	10,296	0.58
71	11,953	0.54
73	5,168	0.546
74	4,769	0.615
75	1,482	0.541
77	6,195	0.741
78	9,842	1.505
79	4,136	0.642
80	1,926	1.058
81	6,472	0.777
82	9,624	0.817
92	1,248	0.514
93	7,853	0.876
95	1,889	0.59

Table 6: Estimates of returns to scale across 2-digit SICs, following the Levinsohn and Petrin (2003) approach with a Cobb-Douglas production function.

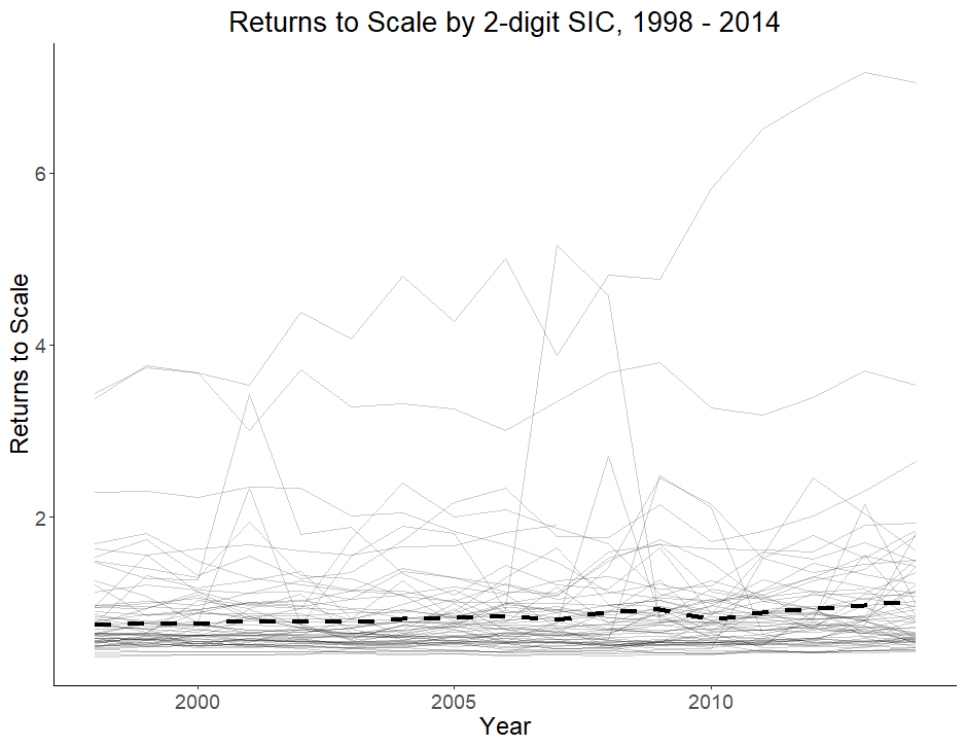


Figure 12: 2-digit SIC returns to scale estimates from a Cobb-Douglas production function using the Levinsohn and Petrin (2003) approach.

The time- and firm-specific returns to scale permits analysis of the distribution across firms, over time. Figure 13 shows the shift in the distribution of scale economies from the start to the end of the period of analysis. Both the mean and median returns to scale are somewhat below unity. From 1998 to 2014, we can see an increase in the mass of firms in the tails of the distribution, and it is especially noticeable that there are more firms with increasing returns to scale between around 1 and 2.

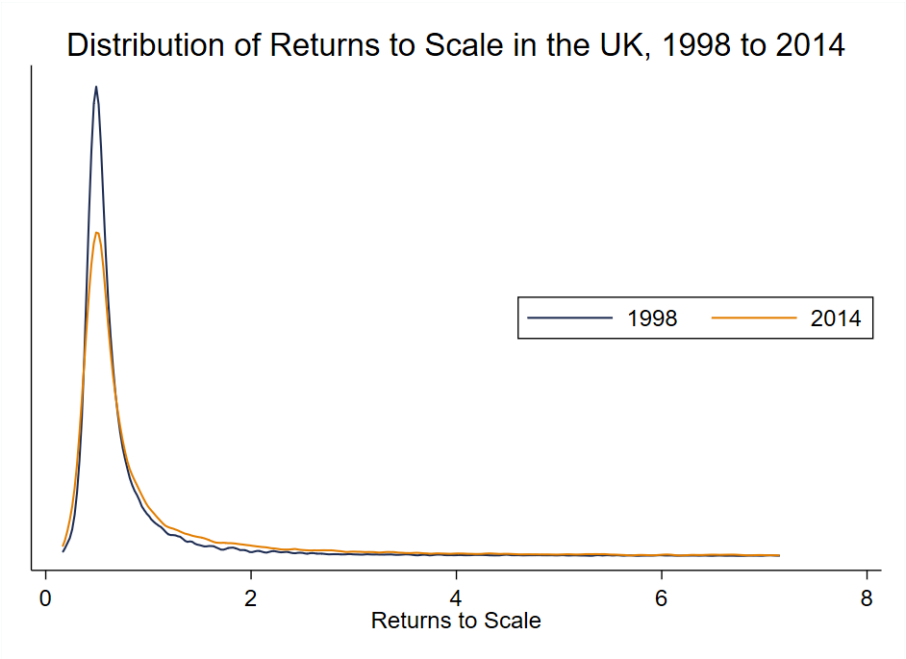


Figure 13: Distribution of firm-level returns to scale estimates, using Levinsohn and Petrin (2003) with a translog production function.

We know from theory that input factor elasticities differ from input shares when markups are not equal to one. Despite this, it may be the case that the sum of output elasticities (i.e. returns to scale) is higher for industries with higher labour shares, and lower materials shares. If firms optimally respond to high scale economies - driven by high labour elasticity - by hiring more labour, then the relationship will be positive.

The results from a regression of returns to scale on the labour and materials shares, with year and industry fixed effects, are presented in Table 7. When scale economies are greater, we see two robust relationships: the labour share is higher, and the materials share is lower.

Table 7: Regression: Returns to Scale and Input Shares

	Dependent variable: Returns to Scale	
Labour Share	3.341^{***}	0.845^{**}
	(0.272)	(0.327)
Materials Share	-2.843^{***}	-2.056^{***}
	(0.243)	(0.306)
N	1,003	1,003
Year FE:		✓
2-digit SIC FE:		✓

Estimates statistically significant at levels of 0.1%: ***, 1%: **, 5%: *. Robust standard errors clustered at the level of the fixed effects included. Weighted by the number of firms in each industry.

C Productivity Estimates

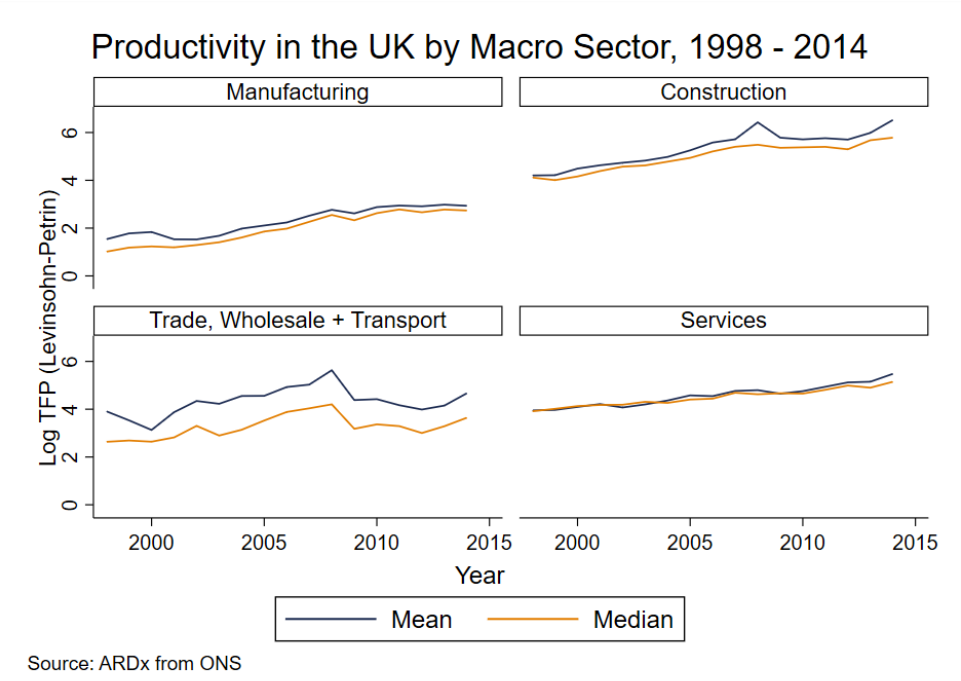


Figure 14: Log productivity estimates from Levinsohn and Petrin (2003) production function estimation, estimated at the macro sector level.

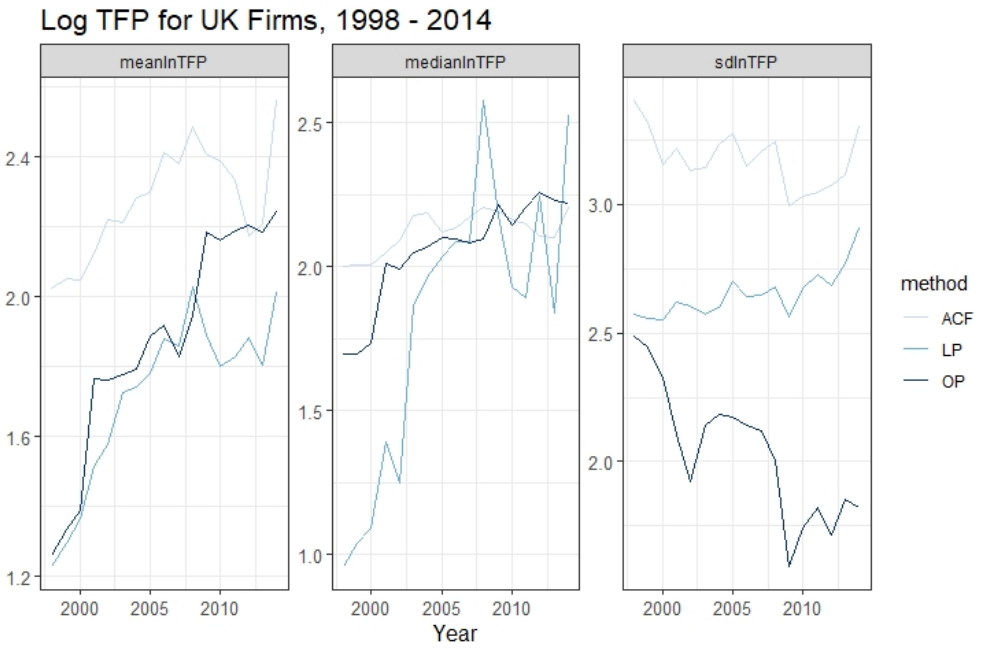


Figure 15: Productivity estimates from the three different control function approaches. Regressions were run at the 2-digit industry level.

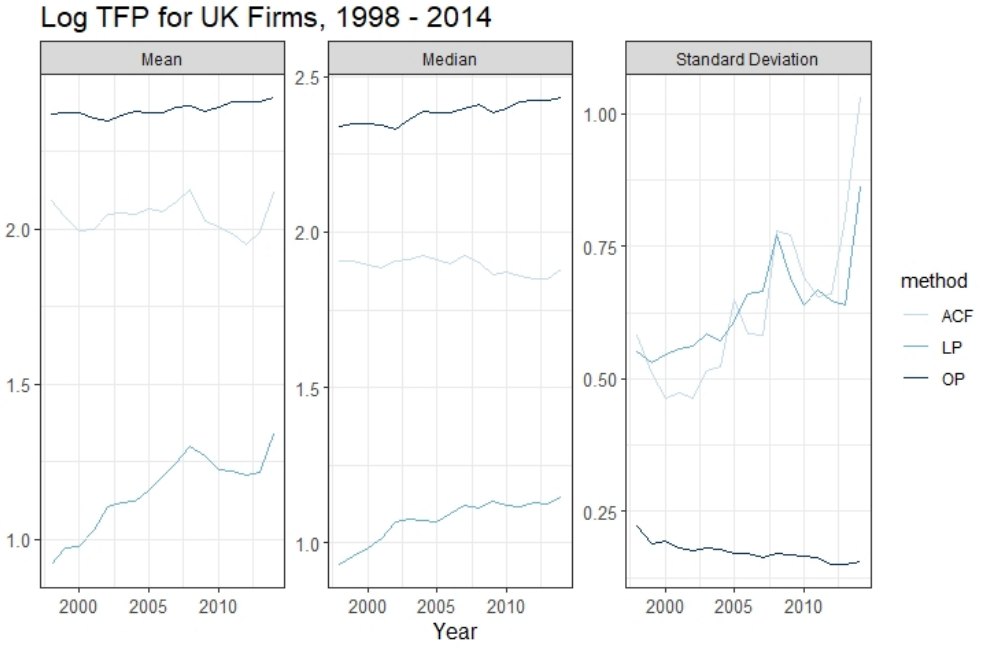


Figure 16: Productivity estimates from the three different control function approaches. Regressions were run at the economy-wide level.

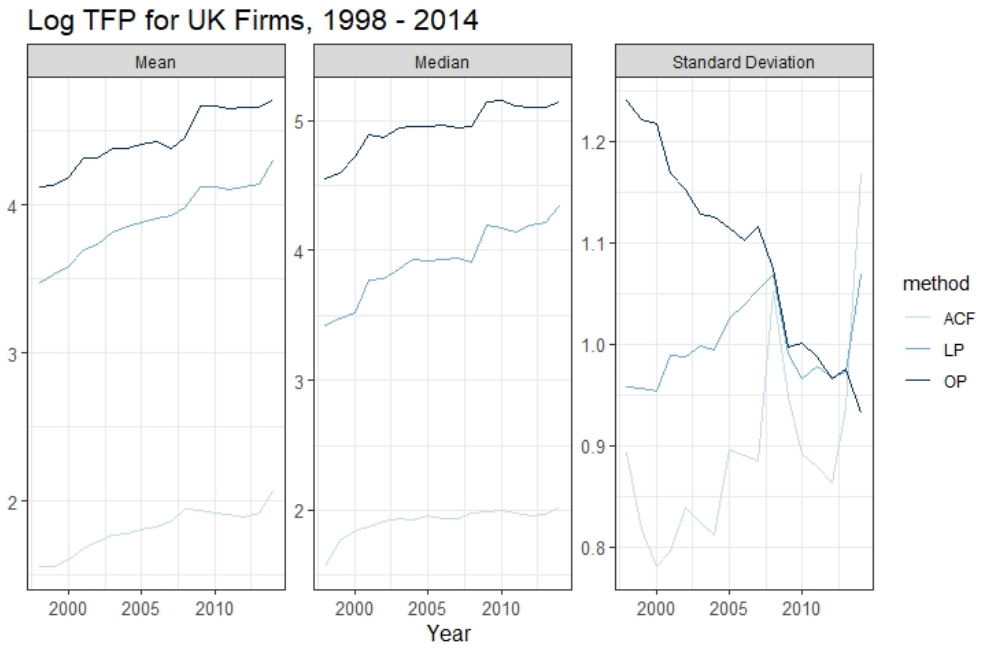


Figure 17: Productivity estimates from the three different control function approaches. Regressions were run at the macro-sector level.

Markup Estimation

Here we present results on markup estimation, which is the ratio of prices to marginal cost and is inferred from the elasticity of the output to the materials input, divided from materials' share of expenditure. The evidence suggests that markups have been increasing over time, replicating results from Hwang and Savagar (2020). The aggregate markup rises from 1.20 in 1998 to 1.96 in 2014.

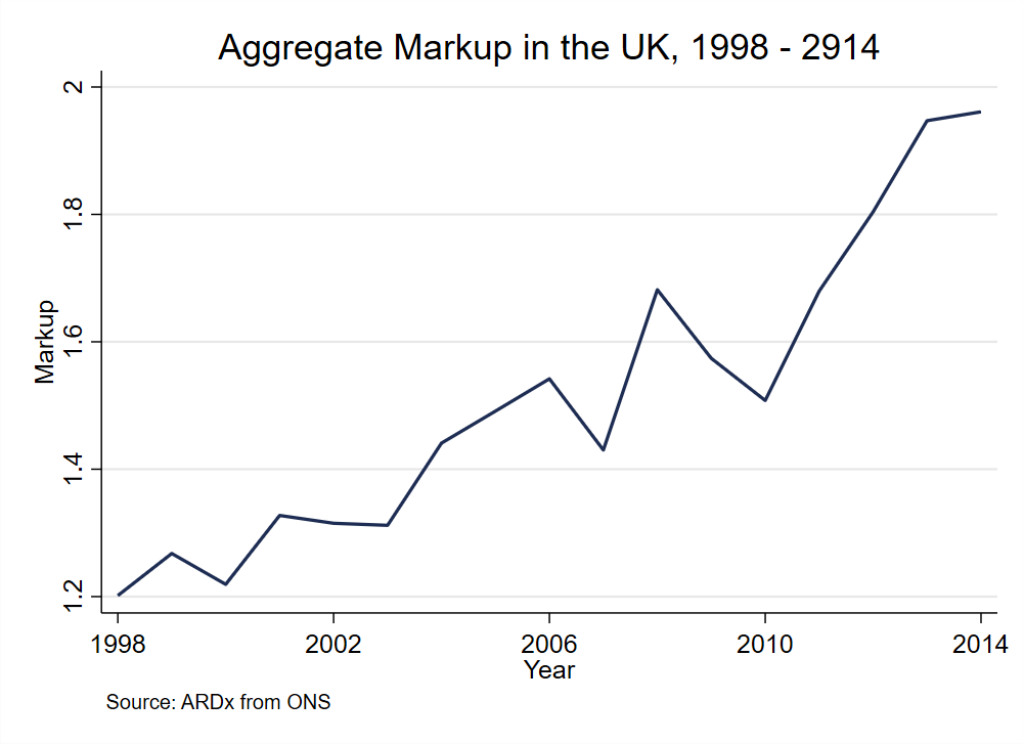


Figure 18: Economy-wide markup, computed as materials output elasticity over its expenditure share, estimated using Levinsohn and Petrin (2003).

Figure 19 presents sector-level markups using fixed estimated materials elasticities. The sector with the largest markup is Services, rising from around 1.6 in 1998 to 2.6 in 2014. The markup in Manufacturing is the lowest, fluctuating slightly from just below one at the start of the period of analysis, and reaching around one in 2014. Otherwise, we observe a significant rise in markups across sectors.

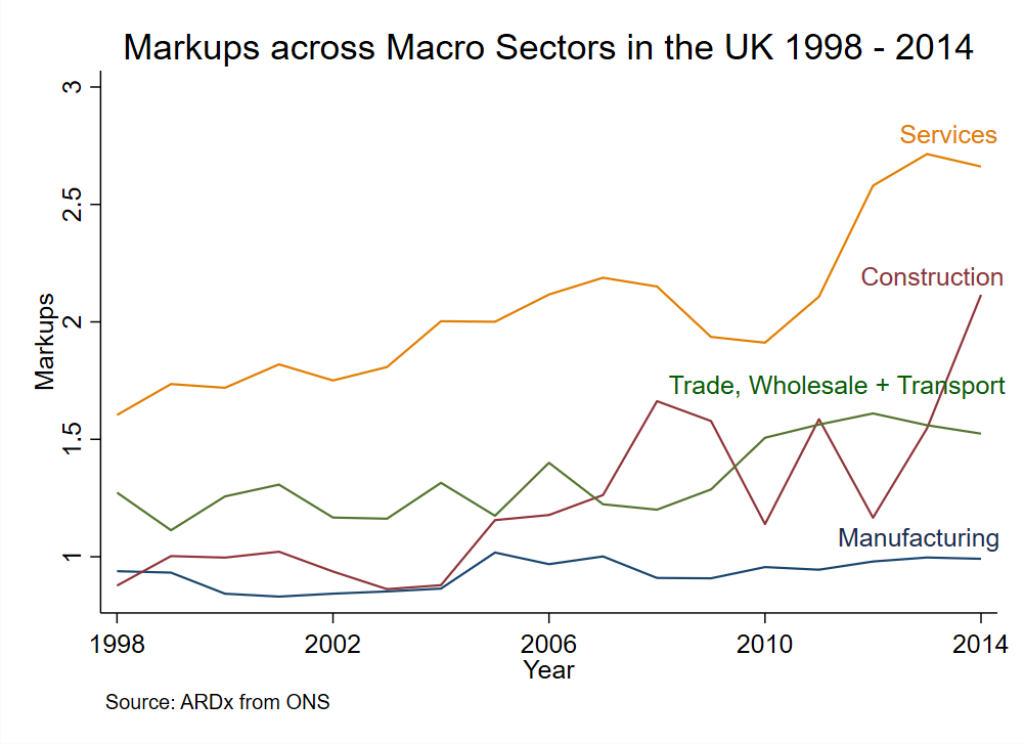


Figure 19: Sector-level markups, computed as materials output elasticity over its expenditure share, estimated using Levinsohn and Petrin (2003).

In Figure 20 we show the distribution of markups across firms in the first and final year of analysis.[9] It clearly highlights the long right-tailed distribution of markups in the U.K., with the vast majority of firms bunching below a value of 2. This graph also shows the shift in markups, both on average and in skewness: by 2014, there is more mass of markups above one, and the tail stretches much further to the right.

[9]Outliers have been removed in each year; either due to typographical errors, or almost-zero materials inputs for firms in certain industries, we occasionally get negative or massive markups.

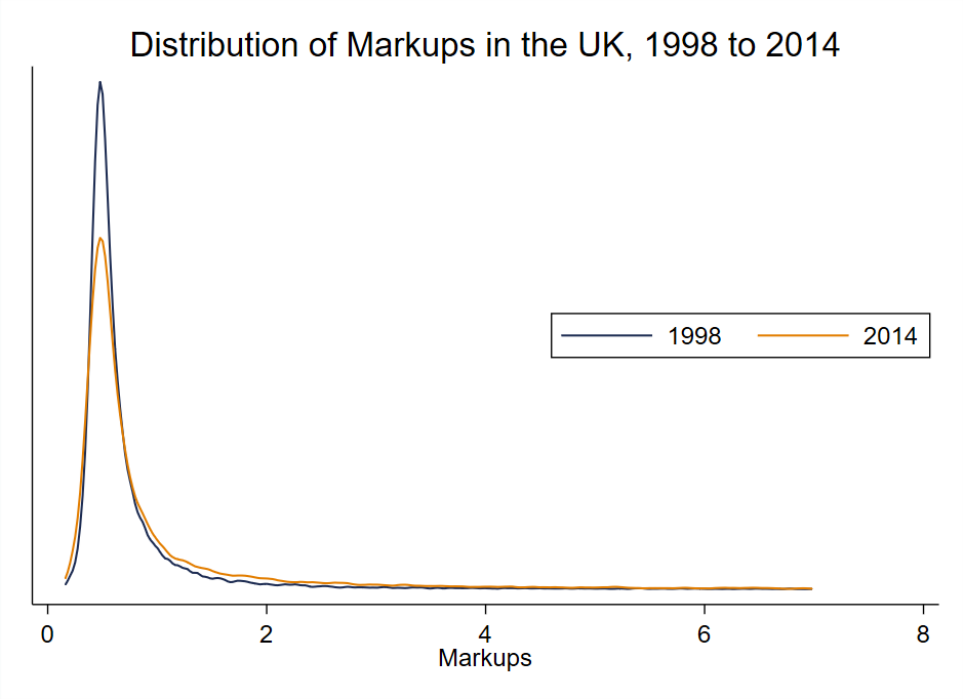


Figure 20: Distribution of firm-level markup estimates, using Levinsohn and Petrin (2003) to obtain output materials elasticity.

D Further Figures & Tables

Table 8: Summary Statistics from the ARDx, 1998 - 2014

Economy-Wide				
	y	k	l	m
Mean	7.6	6	3.2	7
Stdev	2.4	2.2	1.9	2.6
N			527,800	
Manufacturing				
Mean	8.5	7.4	4	8
Stdev	2	1.9	1.6	2.3
N			120,700	
Construction				
Mean	7	4.9	2.6	6.3
Stdev	2.3	2.1	1.8	2.6
N			51,800	
Trade, Wholesale + Transport				
Mean	7.7	5.8	2.9	7.3
Stdev	2.4	2.2	1.8	2.5
N			182,000	
Services				
Mean	7	5.7	3.1	6
Stdev	2.3	2.2	2	2.5
N			173,300	

Summary statistics on panel of firms used for production function estimation. Output, labour, capital, and materials are measured in logs.

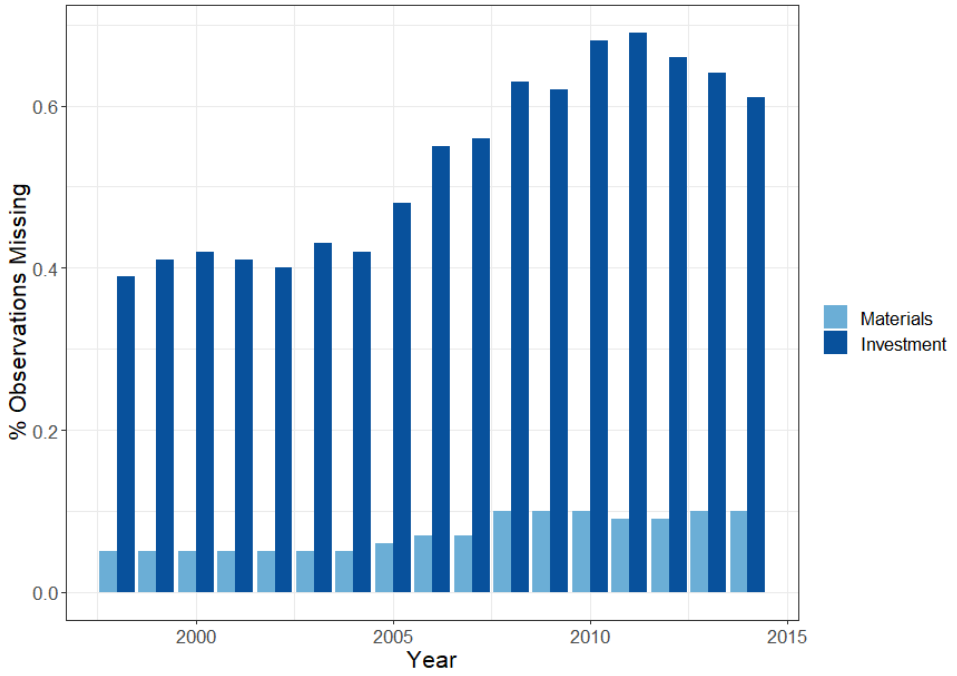


Figure 21: Share of Missing Firm Observations for Materials and Investment, by Year.

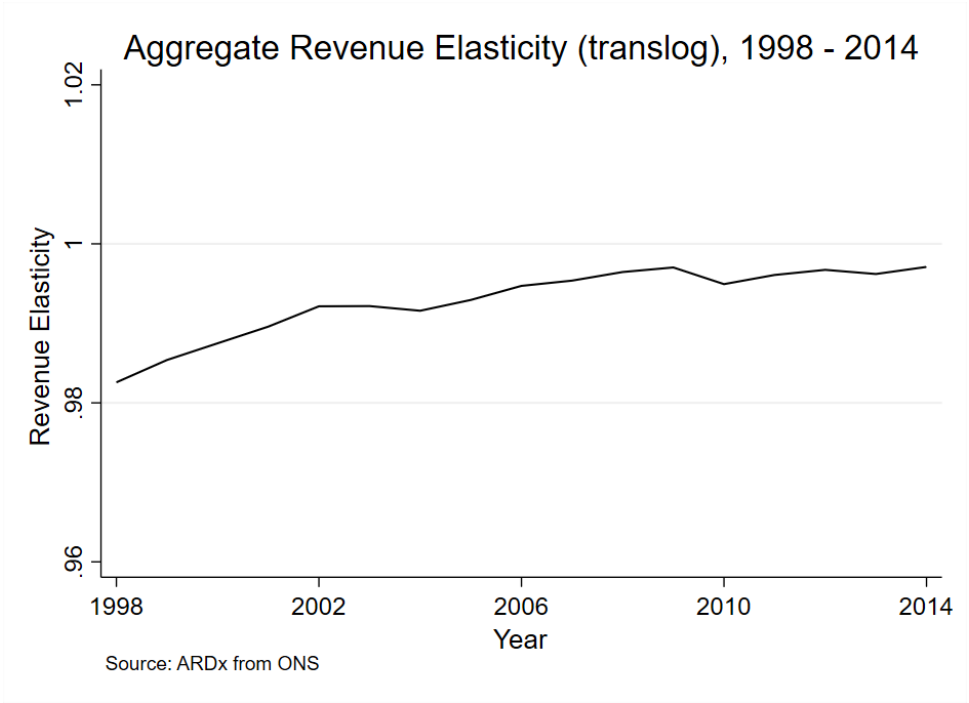


Figure 22: Estimated revenue elasticity using a translog production function, following Levinsohn and Petrin (2003). Can be interpreted as one minus the profit share.

Table 9: Regression: Returns to Scale and Median Log Productivity

	Dependent variable: Returns to Scale			
Median log TFP	0.134***	0.131***	0.131***	0.126***
	(0.009)	(0.009)	(0.025)	(0.031)
N	1,006	1,006	1,006	1,006
Year FE:		✓		✓
2-digit SIC FE:			✓	✓

Estimates statistically significant at levels of 0.1%: ***, 1%: **, 5%: *. Robust standard errors clustered at the level of the fixed effects included. Weighted by the number of firms in each industry.

Table 10: Regression: Returns to Scale and Mean Log Productivity

	Dependent variable: Returns to Scale			
Mean log TFP	0.172***	0.189***	0.265***	0.358***
	(0.015)	(0.014)	(0.013)	(0.015)
N	494,217	494,217	291,933	291,933
Year FE:		✓		✓
2-digit SIC FE:			✓	✓

Estimates statistically significant at levels of 0.1%: ***, 1%: **, 5%: *. Robust standard errors clustered at the level of the fixed effects included. Weighted by revenue at the firm level.

E Model Derivations

E.1 Revenue Elasticity

Price elasticity of demand is given by:

$$-\frac{\partial \mathcal{D}}{\partial p}\frac{p}{y} = -\left(\frac{\partial \mathcal{P}}{\partial y}\frac{y}{p}\right)^{-1}.$$

Firms' revenue elasticity will depend both on the direct effect of changing an input X, as well as the indirect effect of changing input X on demand. Revenue elasticity is defined:

$$\frac{\partial R}{\partial X}\frac{X}{py} = \left[\frac{\partial \mathcal{P}}{\partial y}\frac{\partial y}{\partial X}y + p\frac{\partial y}{\partial X}\right]\frac{X}{py} = \left[\left(\frac{\partial \mathcal{D}}{\partial p}\frac{p}{y}\right)^{-1} + 1\right]\frac{\partial y}{\partial X}\frac{X}{y}.$$

Table 11: Regression: Returns to Scale and Mean Log Productivity

	Dependent variable: Returns to Scale			
Mean log TFP	0.184***	0.203***	0.241***	0.307***
	(0.003)	(0.003)	(0.005)	(0.006)
N	494,217	494,217	291,933	291,933
Year FE:		✓		✓
2-digit SIC FE:			✓	✓

Estimates statistically significant at levels of 0.1%: ***, 1%: **, 5%: *. Robust standard errors clustered at the level of the fixed effects included. Weighted by employment at the firm level.

Given that the markup is the inverse of the final term in square brackets, it follows that revenue elasticity is equal to output elasticity divided by the markup.

E.2 Cost Relationships

From cost minimisation, variable costs are $\mathcal{C} = wL + rK = z\lambda y\left(\frac{\partial y}{\partial L}\frac{L}{y} + \frac{\partial y}{\partial K}\frac{K}{y}\right)$. Euler's homogeneous function theorem states that a function multiplied by its degree of homogeneity will be equal to the sum of partial derivatives multiplied by the arguments:

$$\nu F(X_1, ..., X_N) = \sum_{i=1}^{N} \frac{\partial F(X_1, ..., X_N)}{\partial X_i} X_i$$

Applying this theorem to our minimised variable costs yields:

$$\begin{aligned}
\mathcal{C} &= z\lambda y\left(\frac{\partial y}{\partial L}\frac{L}{y} + \frac{\partial y}{\partial K}\frac{K}{y}\right) \\
&= z\lambda\left(\frac{\partial y}{\partial L}L + \frac{\partial y}{\partial K}K\right) \\
&= \lambda\nu z F(K, L) \\
&= \lambda\nu(y + \phi)
\end{aligned}$$

E.3 Returns to Scale: Markups and Profit Share

Consider the profit function $\pi = \mathcal{P}(y)y - \mathcal{C}(y;w,r)$ and note that the markup $\mu = \frac{p}{MC}$. Some simple algebra yields another equation for returns to scale:

$$\begin{aligned}
\pi &= \mathcal{P}(y)y - \mathcal{C}(y;w,r) \\
\frac{\pi}{\mathcal{P}(y)y} &= 1 - \frac{\mathcal{C}(y;w,r)}{\mathcal{P}(y)y} \\
s_\pi &= 1 - \frac{\lambda}{\lambda}\frac{\text{AC}}{\mathcal{P}(y)} \\
s_\pi &= 1 - \mu^{-1}\frac{\text{AC}}{\text{MC}} \\
\frac{\mathcal{AC}}{\mathcal{MC}} &= \mu(1 - s_\pi)
\end{aligned}$$

E.4 Demand Side

Following Melitz (2003), Cobb-Douglas preferences are defined over J sectors:

$$U = \sum_j \beta_j \ln Y_j.$$

with $\sum_j \beta_j = 1$.

In each sector j, there's a continuum of horizontally-differentiated varieties, and preferences for these are CES:

$$Y_j = \left(\int_{\omega \in \Omega_j} y_j(\omega)^{(\sigma_j - 1)/\sigma_j} d\omega \right)^{\sigma_j/(\sigma_j - 1)}.$$

where $\sigma_j > 1$.

If M is aggregate income, then Cobb-Douglas preferences over sectors yields expenditure $E_j = \beta_j M$. The demand for each variety in each sector is thus:

$$y_j(\omega) = A_j p_j(\omega)^{-\sigma_j}.$$

where $A_j = E_j P_j^{\sigma_j - 1}$, and the price index:

$$P_j = \left(\int_{\omega \in \Omega_j} p_j(\omega)^{(1-\sigma_j)} d\omega \right)^{1/(1-\sigma_j)}.$$

A_j is an index of market demand that scales each firm's residual demand, and is determined by sectoral expenditure and the CES price index. Given a continuum of firms, each firm takes this as given. Dropping the sector and variety notation for simplicity, we can analyse firm behaviour within each sector in this environment.

The demand function for each firm $y(z)$ can be rewritten by combining the equation A_j above, and noting that expenditure equals revenue (so $E = R = py$) in each sector: $y(z) = \left(\frac{p(z)}{P}\right)^{-\sigma} \frac{E}{P} = \left(\frac{p(z)}{P}\right)^{-\sigma} Y$.

F Graphical Illustration of Returns to Scale

Figure 23 illustrates the cost curves of a firm with a fixed cost and increasing marginal cost curve. The firm's marginal cost intersects the average total cost at its minimum. This minimum point is the firm's *minimum efficient scale* (MES) which would arise under perfect competition and at this minimum the firm has constant returns to scale. To the left-hand side of the MES the firm has increasing returns to scale and to the right-hand side the firm has decreasing returns to scale.

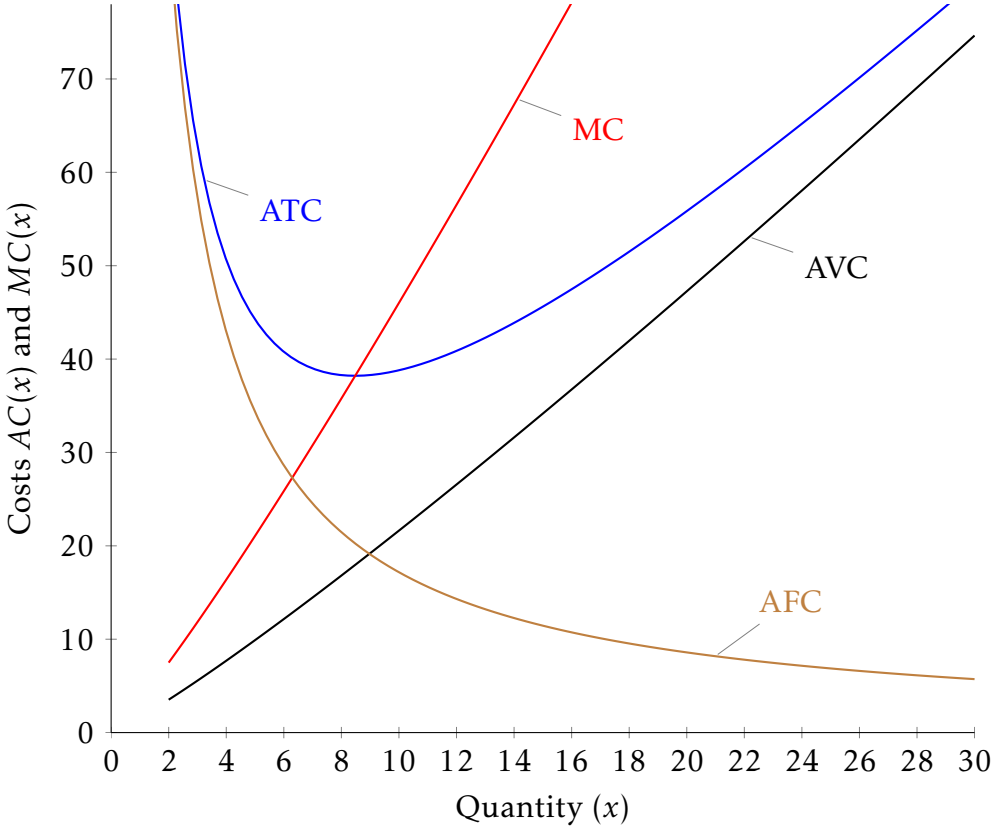


Figure 23: Fixed Cost with Increasing MC, U-Shaped AC Curve

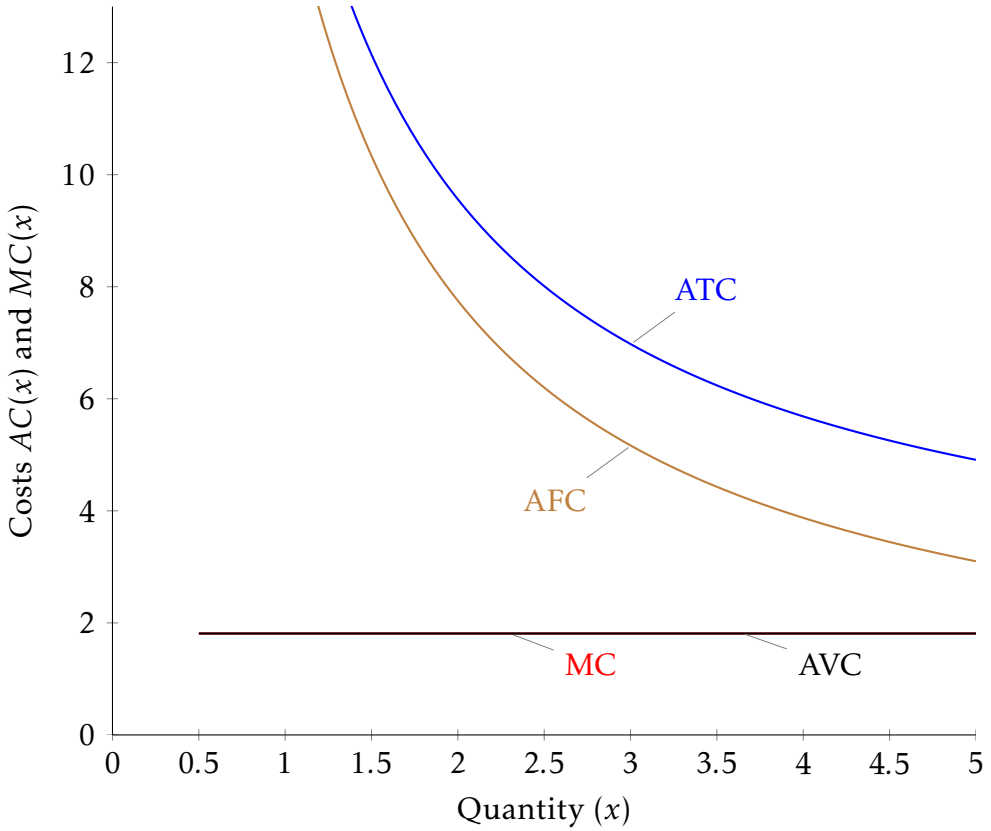


Figure 24: Fixed Cost with Constant MC, Globally Decreasing Returns

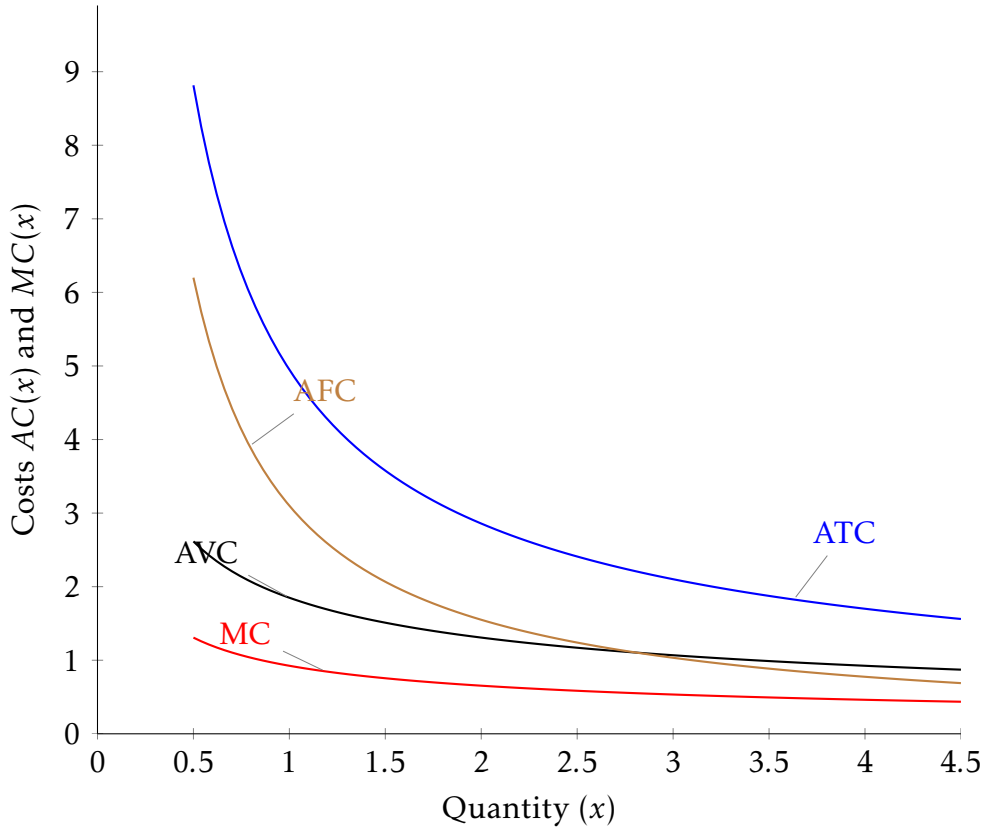


Figure 25: Fixed Cost with Decreasing MC, Globally Decreasing Returns

We denote output by x. Labour and capital inputs are denoted L and K. The price of labour is given by the wage w and the price of capital is given by the interest rate r.

$$\text{Production function} : x = L^{\beta} K^{\alpha}$$
$$\text{Total cost function} : TC(K, L) = wL + rK$$

The fixed parameters β and α are production function elasticities that determine the slope of the marginal cost curve. Specifically, $\partial \ln x / \partial \ln L = \beta$ is the output elasticity to labour and $\partial \ln x / \partial \ln K = \alpha$ is the output elasticity to capital.

F.0.1 Cost Minimization

A cost-minimizing firm chooses inputs such that the relative unit costs of the inputs equals to the marginal rate of technical substitution substitution.

$$|\text{MRTS}| \equiv \frac{\text{MPL}}{\text{MPK}} = \frac{w}{r}.$$

Therefore, a cost-minimizing firm employs capital and labour in order to satisfy the following condition

$$\begin{aligned} \frac{\beta}{\alpha} \frac{L^{\beta-1} K^{\alpha}}{L^{\beta} K^{\alpha-1}} &= \frac{w}{r} \\ \frac{\beta}{\alpha} \frac{K}{L} &= \frac{w}{r} \\ K &= \frac{\alpha}{\beta} \frac{w}{r} L. \end{aligned}$$

Substituting K into the production function gives the demand for L conditional on output x, relative factor prices w/r and production function elasticities β, α:

$$
\begin{aligned}
x &= L^{\beta}\left(\frac{\alpha}{\beta}\frac{w}{r}L\right)^{\alpha} \\
x &= L^{\beta+\alpha}\left(\frac{\alpha}{\beta}\frac{w}{r}\right)^{\alpha} \\
L &= \left(x\left(\frac{\alpha w}{\beta r}\right)^{-\alpha}\right)^{\frac{1}{\beta+\alpha}}.
\end{aligned}
$$

Using this factor demand curve, conditional on output (x), we can derive the cost function. We know $C(x) = wL + rK$ and we'll assume that capital is quasi-fixed, so we replace K with a constant fixed cost FC:

$$
TC(x) = w\left(\frac{\alpha w}{\beta r}\right)^{-\frac{\alpha}{\beta+\alpha}} x^{\frac{1}{\beta+\alpha}} + rFC.
$$

The derivative of the total cost function with respect to output is the marginal cost function:

$$
\begin{aligned}
MC(x) &= \frac{\mathrm{d}C(x)}{\mathrm{d}x} \\
MC(x) &= \frac{w}{\beta+\alpha}\left(\frac{\alpha w}{\beta r}\right)^{-\frac{\alpha}{\beta+\alpha}} x^{\frac{1}{\beta+\alpha}-1}.
\end{aligned}
$$

The average cost function is the total cost function divided by output, x:

$$
\begin{aligned}
AC(x) &= \frac{C(x)}{x} \\
AC(x) &= \frac{w\left(\frac{\alpha w}{\beta r}\right)^{-\frac{\alpha}{\beta+\alpha}} x^{\frac{1}{\beta+\alpha}} + rFC}{x} \\
AC(x) &= w\left(\frac{\alpha w}{\beta r}\right)^{-\frac{\alpha}{\beta+\alpha}} x^{\frac{1}{\beta+\alpha}-1} + \frac{rFC}{x}.
\end{aligned}
$$

Average total cost (ATC) consists of average variable cost (AVC) and average fixed cost (AFC):

$$AVC(x) = w\left(\frac{\alpha w}{\beta r}\right)^{-\frac{\alpha}{\beta+\alpha}} x^{\frac{1}{\beta+\alpha}-1}$$
$$ATC(x) = \frac{rFC}{x}$$

Average variable cost and marginal cost are related as follows

$$\beta + \alpha = \frac{\text{AVC}}{\text{MC}} = \left(\frac{\partial \text{VC}/\partial x}{VC/x}\right)^{-1} = \epsilon_{\text{VC}x}^{-1}.$$

The ratio of average variable cost to marginal cost is the sum of input elasticities. Average variable cost and marginal cost are equivalent with a constant marginal cost. The ratio of average variable cost to marginal cost is the inverse variable cost elasticity.

The elasticity of variable cost is $\nu = \beta + \alpha$ whereas the elasticity of total cost, with output denominated fixed cost, is $\nu(1 + s_\phi)$. And with labour denominated fixed cost it is $\nu + ws_\phi$.