Scale Economies and Aggregate Productivity

Joel Kariel* Anthony Savagar†

July 12, 2024

Abstract

We develop a theoretical framework to investigate the link between rising scale economies and stagnating productivity. Our model features heterogeneous firms, imperfect competition, and firm selection. We demonstrate that scale economies generated by fixed costs have distinct impacts on aggregate productivity compared to those driven by returns to scale. Using UK data, we estimate long-run increases in both fixed costs and returns to scale. Our model implies that this should increase aggregate productivity through improved firm selection and resource allocation. However, increasing markups can o set the productivity gain. Higher markups cushion low-productivity firms' revenues, allowing them to survive, and constrain firm output, which limits exploitation of scale economies.

JEL: E32, E23, D21, D43, L13.

Keywords: Returns to Scale, Scale Economies, Productivity, Market Structures, Firm Dynamics, Fixed Costs, Marginal Costs.

*Competition and Markets Authority and University of Kent, joel.kariel@cma.gov.uk

†University of Kent, 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.*

We thank seminar participants at Cardi , CompNet, Nottingham, York, Durham Macro workshop, EUI, Bank of Italy, MWM Clemson, University of Washington, RES 2023, EMF Bern 2023, Birmingham, St Louis Fed, Lancaster, King's, Bath, IFN Stockholm, EEA-ESEM 2022, IAAE 2022, CEF 2022, RES 2022, SNDE 2022, AMEF 2022, SES 2022, MMF 2022, Kent Firm Dynamics Workshop 2021, Exeter Macro Workshop 2022 and Bristol for their helpful comments. We thank the following people for feedback: Matthias Kehrig, Jan de Loecker, Jan Eeckhout, Mark Bils, Alex Monge, Max Gillman, Mark Wright, Omar Licandro, Julian Neira, Tom Schmitz, Petr Sedláček, Danial Lashkari, John Morrow, Anthony Priolo, Riccardo Silvestrini and Kunal Sangani.

Recent technological advances, such as cloud computing, can raise scale economies allowing firms to expand at lower cost. But, as these technologies have emerged in economies such as the US and UK, productivity has stagnated. In this paper, we develop a theory to relate firm-level scale economies to aggregate productivity. We show that increases in scale economies should have increased aggregate productivity significantly. However, rising markups can offset the productivity gains.

We make three contributions: first, we document rising scale economies from two determinants: higher returns to scale and higher fixed costs. Second, we develop a tractable model to link these determinants of scale economies to aggregate productivity. Third, we conduct a quantitative exercise to replicate growing scale economies but stagnating productivity in the UK economy.

We develop a heterogeneous firm model with monopolistic competition, fixed costs, returns to scale and endogenous entry. We derive firm-level scale economies, which is the inverse cost elasticity or, equivalently, the ratio of average cost to marginal cost. Firm-level scale economies are a function of fixed costs and returns to scale, and they vary endogenously with firm size.[1] The fixed cost and returns to scale determinants of scale economies have different aggregate productivity outcomes. Both tend to increase aggregate productivity, by reducing profits and in turn the number of active firms. Fewer active firms enhances productivity through selection of high technical efficiency firms and exploitation of increasing returns (if present). However, the effect of fixed costs on aggregate productivity is independent of markups, whereas the effect of returns to scale is mitigated by the presence of markups. Quantitatively, estimated increases in fixed costs cannot buoy aggregate productivity sufficiently to offset the negative effect on aggregate productivity from estimated increases in markups. On the other hand, estimated increases in returns to scale can buoy aggregate productivity sufficiently to counteract rising markups.

We decompose aggregate productivity into allocative efficiency and technical efficiency components. *Allocative efficiency* depends on the division of aggregate resources

[1] Returns to scale are returns to scale in variable inputs. This measures the slope of the marginal cost curve and is the sum of output elasticities to variable inputs.

across firms and how this interacts with returns to scale, as well as the fixed cost that each additional firm must pay. Increasing returns favour concentrating resources on fewer, larger producers, while decreasing returns favour the opposite. *Technical efficiency* measures the average technology of *active* firms. Technology is an exogenous productivity characteristic that is revealed to firms upon entry. Given a technology draw, a firm decides to be active or inactive based on a period-by-period fixed cost. Therefore, technical efficiency is determined by firm selection. That is, where the exogenous productivity distribution is truncated.

Our theoretical results show that rising scale economies, either through fixed costs or returns to scale, strengthen selection, thus improving average technical efficiency. However, in high-markup environments, the selection channel is weaker. With high markups, selection weakens because small (low technology draw) firms get more revenue for each unit sold, so it is easier to cover fixed costs and survive. Allocative efficiency declines because markups increase the number of firms which limits the exploitation of scale economies. Therefore, ceteris paribus, increases in scale economies should increase productivity. However, high mark-ups weaken the passthrough of scale economies to productivity.

Our theory emphasizes the importance of the returns to scale *levels* (decreasing, constant, or increasing) in understanding how rising returns to scale or fixed costs impact aggregate productivity. Concentrating capital and labour among fewer firms increases aggregate output if there are increasing returns, but decreases aggregate output if there are decreasing returns. We estimate increasing returns to scale levels in our data and these have increased over time. Our model shows that rising returns to scale or fixed costs reduce the number of active firms through lower profits. This strengthens the selection of high-technology firms, boosting productivity. Additionally, concentrating resources within these fewer firms further enhances productivity from increasing returns. Hence, there are two channels leading greater returns to scale and fixed costs to enhance aggregate productivity. Higher markups counteract this e ect because they increase the number of active firms through higher profits.

Ultimately, the theory stresses the importance of the number of active firms for aggregate output. Changes in underlying parameters affect the number of active firms, and they are a crucial determinant of aggregate productivity as they characterise selection and the division of aggregate resources among production units, which matters in the absence of constant returns.

Our quantitative exercise applies the theoretical insights to UK aggregate productivity. We show that estimated increases in returns to scale accompanied by estimated increases in markups replicate UK aggregate productivity dynamics well. Rising fixed costs cannot explain the data as well. If markups had not increased, UK aggregate productivity would have been 20% higher through efficiency gains from scale economies.

Our paper abstracts from the specific technologies that may have changed scale economies, other than to characterise them by increasing fixed costs or raising returns to scale (reducing MC). Industry studies provide some insight. Ganapati (2021) shows that information technology reduced marginal costs and increased markups in the wholesale sector. For the manufacturing sector, Bloom, Garicano, Sadun, and Van Reenen (2014) study specific information technologies, such as enterprise resource planning, that increase managers' span of control and, therefore, lower marginal costs. Syverson (2019) hypothesises a shift towards products with lower marginal costs, such as software and pharmaceuticals. Lashkari, Bauer, and Boussard (2024) link IT price changes to changing scale economies using French data. They find lower IT prices reallocates business to larger firms, with low scale economies, and this can cause a lower labour share.

Therefore, our conclusion is that emerging technologies have increased returns to scale, which has decreased marginal costs and enhanced scale economies. These scale economies should translate into productivity gains. However, increasing market power limits the exploitation of scale economies and, in turn, productivity gains.

Related Literature

Our paper connects theory on the aggregate impacts of microeconomic production primitives, with the measurement of these features at the firm level. Recent work by Bilbiie and Melitz (2020), Edmond, Midrigan, and Xu (2021), and Baqaee, Farhi, and Sangani (2023) demonstrates the importance of returns to scale for aggregate welfare. This work focuses primarily on external returns to scale (love of variety) that arise from aggregation. However, Baqaee, Farhi, and Sangani (2023) also note that returns to scale at the firm level magnify aggregate returns to scale. Similarly to our analysis, the effects of scale economies are smaller in efficient (low markup) economies. Baqaee and Farhi (2020) provide non-parametric aggregation results for models with scale economies. Both our parametric approach and their non-parametric approaches show that the role of allocative efficiency grows as distortions increase. And, we combine this theory with measurement to show that firm-level scale economies are quantitatively-relevant to replicate UK productivity dynamics.

In order to understand the consequences of rising market power, De Loecker, Eeckhout, and Mongey (2021) present a quantitative model with oligopolistic competition and fixed costs. This allows them to compare the role of technology on the supply-side versus competitive factors on the demand-side. We differ by focusing on analytical results to understand the supply-side mechanisms through which different technologies affect scale economies, and in turn aggregate productivity. Our demand-side is restricted to monopolistic competition for tractability. Collectively, our papers advance the idea that to reconcile changing technologies on the supply side, market power must increase on the demand side.

Recent research in endogenous growth theory shows that changing technologies affect firm cost structures, which in turn explains stagnating growth. De Ridder (2024) models intangible inputs as reducing marginal costs and raising fixed costs. Unlike us, the focus is the level of constant marginal costs, not the slope of marginal costs. Aghion, Bergeaud, Boppart, Klenow, and Li (2023) model a fixed cost that increases with the number of product lines, but as technology improves, the fixed cost becomes less sensitive to the number of products. Our paper differs from this research, which focuses on quantitative endogenous growth models with an important role for R&D, and a main aim of replicating US stagnation facts. We present a parsimonious and tractable analysis based on firm entry to directly link the firm-level determinants of scale economies to aggregate productivity. Conceptually, this body of work, including our paper, contributes to the hypothesis that recent changes in technology have affected firm cost structures, and lead to important aggregate effects. To our knowledge, our work is the first to directly compare the effect of returns to scale and fixed costs, and formalise these effects of new technologies through the economies of scale channel. Informally, it is understood that these are two sources of scalable technologies and are hallmarks of intangible capital (Haskel and Westlake 2017). We formalise that they both affect scale economies in the same way, but can lead to distinct aggregate productivity effects.

Our model is a neoclassical growth model with heterogeneous firms based on Hopenhayn and Rogerson (1993), Restuccia and Rogerson (2008), and Barseghyan and DiCecio (2016). The model is similar to two-factor closed-economy versions of Melitz (2003) and Ghironi and Melitz (2005). We include firm production with fixed costs and returns to scale similar to the models of J. Kim (2004), Atkeson and P. J. Kehoe (2005), Bartelsman, Haltiwanger, and Scarpetta (2013), and D. Kim (2021). Gao and Kehrig (2021) present a partial equilibrium industry model under perfect competition and focus on cross-industry variation in returns to scale and productivity dispersion. Similarly to our theory, they show a positive relationship between productivity and returns to scale across industries, whereby a rise in returns to scale leads to selection of more-productive firms.

Several recent articles provide estimates of returns to scale in the US economy. Gao and Kehrig (2021) estimate slightly decreasing returns to scale in US manufacturing firms. Using similar US data, Ruzic and Ho (2019) find a decline in returns to scale from 1982 to 2007. Using Compustat data, Chiavari (2022) documents rising returns to scale through production function estimation, and De Loecker, Eeckhout, and Unger (2020, Figure 7) documents increasing overhead cost shares as evidence of rising scale economies. Baqaee, Farhi, and Sangani (2023) also document economies of scale in US firms. Lashkari, Bauer, and Boussard (2024) find cost elasticity below one for French corporations, which implies economies of scale. For the UK economy, Oulton (1996), Harris and Lau (1998), and Girma and Görg (2002) document constant or slightly decreasing returns to scale for manufacturing firms.

The remainder of our paper is as follows. In Section 1, we present some foundations on scale economies and returns to scale. In Section 2, we present an empirical motivation which shows rising fixed costs and returns to scale, concurrently with stagnating productivity in the UK. In Section 3, we present our model, equilibrium conditions and characterise some properties of the model. In Section 4, we present comparative statics on the e ects of fixed costs and returns to scale on aggregate productivity. Informed by these theoretical insights, in Section 5 we perform a quantitative analysis which simulates our model for an estimated timeseries of returns to scale, fixed costs and markups. We present counterfactual experiments when each of these components changes independently.

1 Scale Economies Background

In this section, we define some concepts which are occasionally subject to ambiguity.

Internal vs. External Returns to Scale: Our interest is internal returns to scale, not external returns to scale that arise from aggregation. Internal returns to scale and scale economies arise within the firm from the production technology or fixed costs. External returns to scale are gains in aggregate output from changing aggregate inputs. They arise from grouping firms together.[2]

Scale Economies: Scale economies describe the response of firm costs to output changes. They are measured by the inverse cost elasticity, which is the average cost

[2]On the demand-side, with a consumption aggregator, the analogous concept is love-of-variety. Other terms used are 'thick markets' (Caballero and Lyons 1992), Ethier e ects (Ethier 1982), and agglomeration e ects (Krugman 1991).

to marginal cost ratio.[3]

Returns to scale: Returns to scale are a property of the production technology. To be precise, they are captured by the degree of homogeneity of the production function. On the cost side, this parameter represents the slope of a firm's marginal cost curve.[4] For homothetic production functions, the scale elasticity of the cost function equals the returns to the scale of the production function.[5] Fixed costs lead to non-homothetic production functions which break this relationship.

Imprecision over the terms scale economies and returns to scale extends beyond semantics. Erroneous conclusions and calibrations occur when the AC/MC ratio is estimated but is interpreted as the production function returns to scale.[6]

1.1 Graphical Intuition of Scale Economies

To aid understanding throughout the paper, it is helpful to present the cost curve scenarios of the production functions we consider. We define scale economies as the inverse cost elasticity, which is the ratio of average cost to marginal cost. With firm output y, we have:

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

where AC $\equiv \mathcal{C}/y$ and MC $\equiv \partial\mathcal{C}/\partial y$. There are economies of scale if $S(y) > 1$; constant scale economies if $S(y) = 1$; and diseconomies of scale if $S(y) < 1$. Figure 1 presents a firm with a U-shaped average cost curve due to increasing marginal costs and fixed cost.[7] At the intersection of average and marginal cost, a firm has constant

[3]This definition of scale economies is common in industrial organization textbooks (Panzar 1989; Church and Ware 2000; Davis and Garcés 2009), recent examples are Syverson (2019) and Conlon, Miller, Otgon, and Yao (2023). It is sometimes recognised in macroeconomics, for example Rotemberg and Woodford (1993), Basu (2008), Baqaee, Farhi, and Sangani (2023), and Lashkari, Bauer, and Boussard (2024).

[4]Occasionally, researchers recognise this parameter as 'span of control' since it is mathematically analogous to the span of control parameter in Lucas (1978). In that context, it captures diminishing returns in managerial span of control. Hopenhayn (2014) analyses the equivalence with returns to scale.

[5]Silberberg and Suen (2000, Ch. 8) present traditional proofs.

[6]Basu (2008) discusses this in detail. Since homothetic production functions are common in macroeconomics, the term returns to scale is often used universally even in the presence of fixed costs.

[7]In the appendix we present plots considering the three main cases that arise in our theory: a fixed cost with increasing, constant or decreasing marginal cost.

scale economies. To the left there are economies of scale. To the right there are diseconomies of scale. Therefore, the $S(y)$ curve shows that size and scale economies are negatively related at the firm level.[8]

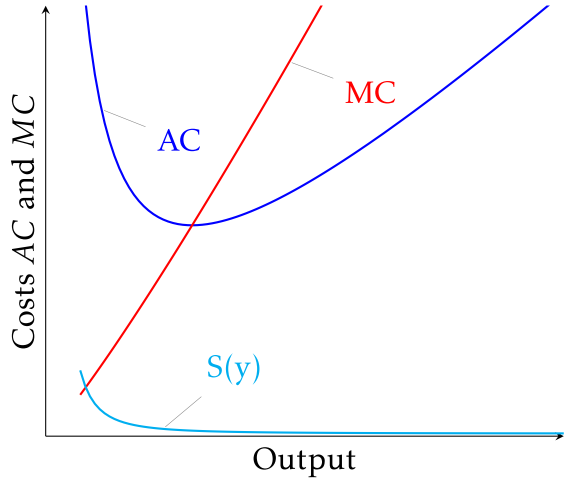


Figure 1: Fixed Cost with Increasing MC, U-Shaped AC Curve

Profits, Markups and Scale Economies: Scale economies can be represented directly from the profit definition. This yields an expression based on market structure, namely markups and profits. Scale economies can also be written in terms of technical properties of the production function, namely fixed costs and the homogeneity parameter. This will depend on the production function and can be derived from the cost function or the production function.[9] Consider the definition of profits as revenue minus costs

$$\text{Profit} = \text{Price} \times \text{Output} - \text{Cost} = \text{Revenue} - \text{Cost}.$$

Divide by revenue, define AC=Cost/Output, and multiply by MC/MC, yields:

$$\frac{\text{AC}}{\text{MC}} = \frac{\text{Price}}{\text{Marginal Cost}} \left(1 - \frac{\text{Profit}}{\text{Revenue}}\right)$$

This shows that a firm's scale economies are its markup multiplied by its profit share remainder (*i.e.* total cost share).[10] A firm that makes zero-profits has scale economies

[8] In the appendix we present a graphical explanation of scale economies from the production side.

[9] In this paper we will show this for labour denominated fixed costs beginning with the production function. Savagar (2021) shows it for output-denominated fixed costs beginning with the cost function.

[10] The total cost share is the sum of the variable cost share and the fixed cost share.

equal to its markup.[11] And, a firm with positive profits will have lower scale economies than the zero-profit firm. Higher scale economies imply higher markups or lower profit shares. Since we develop a framework with constant markups, di erences in scale economies are analagous to di erences in profits shares. Large, high-productivity, firms have large profit shares and low scale economies, whilst small, low-productivity, firms have low profit shares and high scale economies.

Figure 2 illustrates scale economies from the production side. It conveys the idea that small firms have high scale economies, whilst large firms have low scale economies. The figure represents an economy where firm output is produced directly by production labour. In order to produce there is some overhead labour that is the same for both firms. Total labour is the sum of production labour and overhead labour. The figure shows that a 10% rise in total labour at a firm raises production labour by 100% for the small firm, but only 13% for the large firm. Therefore, a proportional change in inputs has a proportionally larger e ect on output for the small firm.

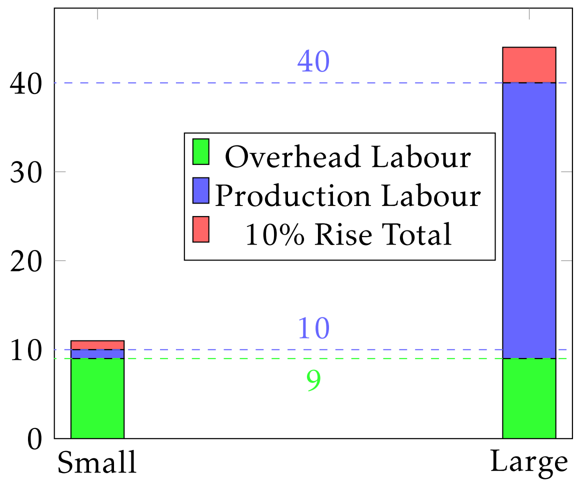


Figure 2: Scale Economies for Large and Small Firm

2 Empirical Motivation

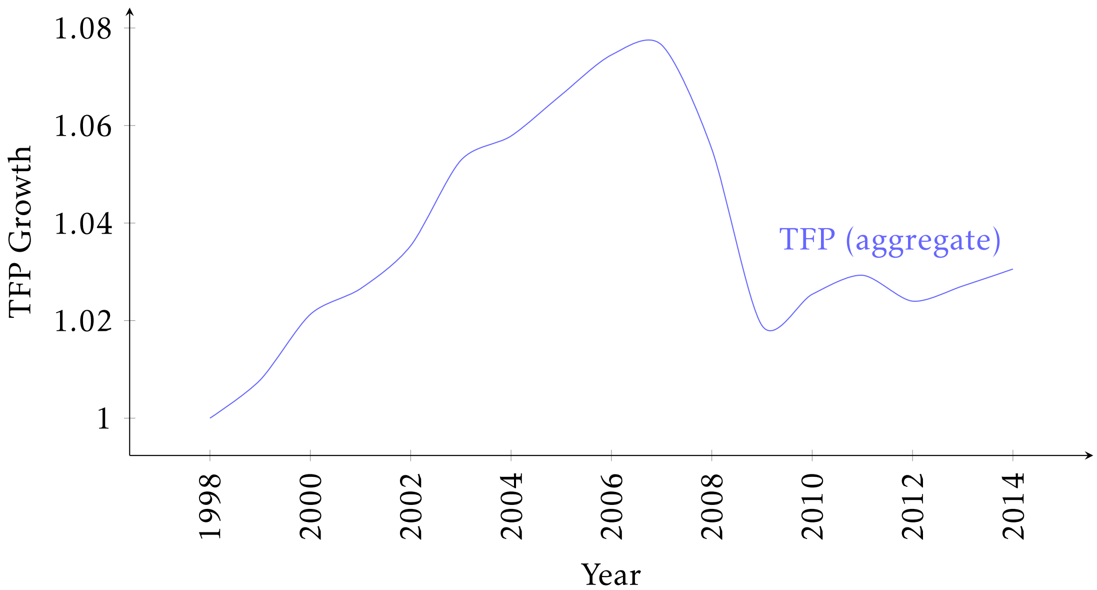
We are motivated by the presence of rising scale economies at the firm level, while aggregate measures of productivity are stagnating.

[11]This result was used in earlier empirical work on returns to scale, when profits in the US economy were close to zero (Basu and Fernald 1997).

2.1 Productivity

Figure 3 shows UK aggregate TFP growth over time. Aggregate productivity growth increases until 2007 but then declines and stagnates. This captures the UK 'productivity puzzle' (Barnett, Batten, Chiu, Franklin, and Sebastia-Barriel 2014; Goodridge, Haskel, and Wallis 2016).

Figure 3: UK TFP Growth, 1998 - 2014



TFP growth (aggregate) is from the Penn World Table 10.01 (Feenstra, Inklaar, and Timmer 2015), accessed from FRED: Total Factor Productivity at Constant National Prices for United Kingdom (RTFPNAGBA632NRUG).

2.2 Returns to Scale in Variable Inputs

To measure returns to scale, we estimate firm-level production functions on UK data from the Annual Respondents Database (ARDx). The data contains approximately 50,000 firms each year, 11 million workers, and two-thirds of gross value added. Firms report a range of production data, including gross output, value added, labour, materials, and investment.[12] We assume that we observe variable inputs, net of fixed costs.

[12]In the appendix, we provide details about the data, data cleaning, deflation, capital construction, SIC code matching, and summary statistics.

We assume that each firm j has the following Cobb-Douglas production function

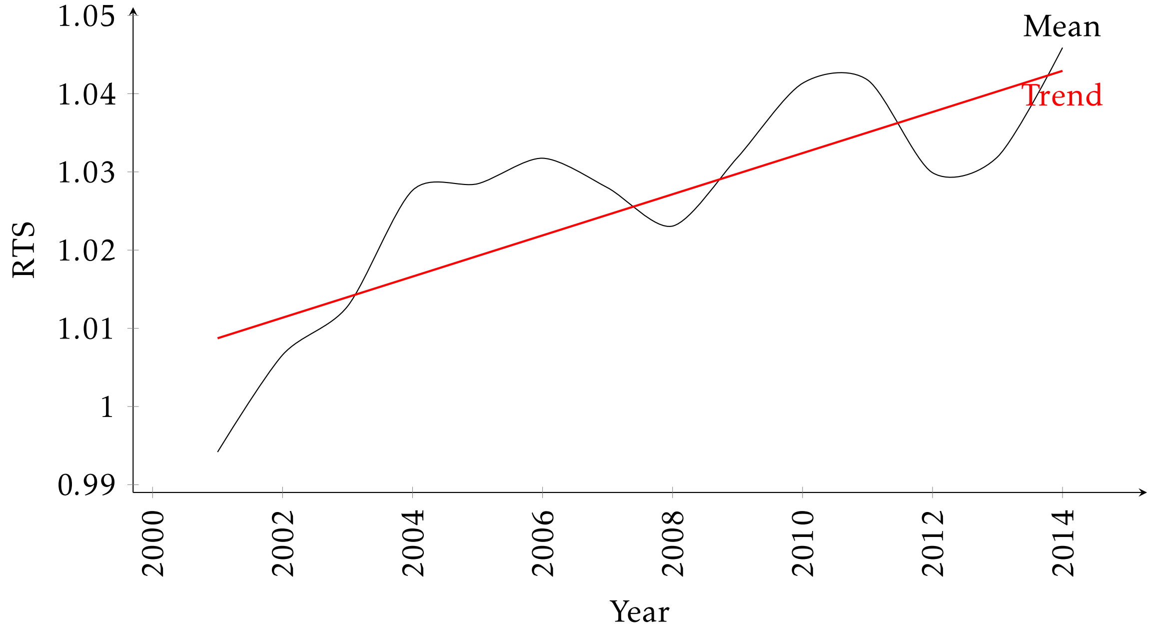
$$y_{jt} = A_{jt} k_{jt}^{\beta_k} \ell_{jt}^{\beta_\ell}$$

where y_t, k_{jt}, ℓ_{jt} are firm value-added and inputs of capital and labour. A_{jt} is a measure of firm-level technical efficiency which we do not observe. Our aim is to estimate the β_k and β_ℓ parameters which represent output elasticities. The sum of these output elasticities is returns to scale in variable inputs.

Production function estimation su ers from omitted variable bias. The bias occurs because the input variables are correlated with the unobserved firm-level technology term. There are various methods to address this problem (Olley and Pakes 1996; Levinsohn and Petrin 2003; Ackerberg, Caves, and Frazer 2015; Gandhi, Navarro, and Rivers 2020). Since we estimate Cobb-Douglas production functions, we obtain a single, time-invariant, coefficient for each input in the production function.

Figure 4 shows estimated returns to scale across firms in the UK using the estimation methodology of Gandhi, Navarro, and Rivers (2020). There is a rising trend in returns to scale, from weakly decreasing to above unity. In the appendix, we provide estimates at the industry level and for alternative estimation methodologies. All the results imply rising returns to scale.

Figure 4: UK RTS, 2001 - 2014

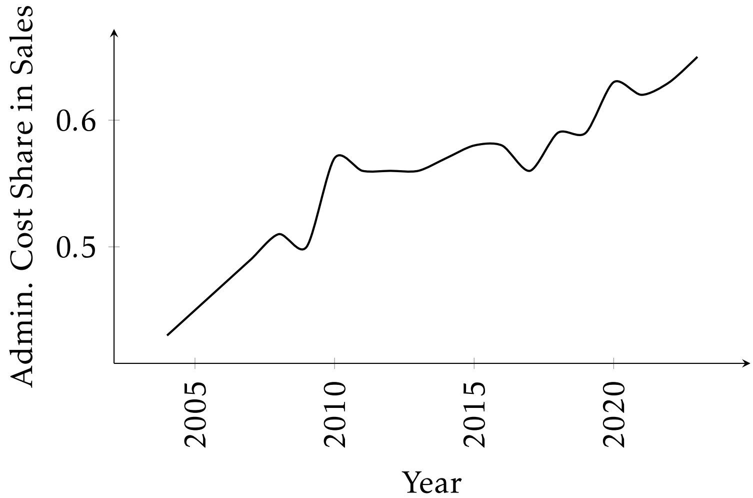


RTS are the sum of firm-level coefficients from a Cobb-Douglas, gross-output, production function estimated with the methodology of Gandhi, Navarro, and Rivers (2020). To obtain time-varying estimates of RTS, we estimate production functions over rolling windows.

2.3 Fixed Cost Share in Revenue

An alternative contributor to firm scale economies is the fixed cost share. In Figure 5, we use the administration expense share in revenue as a proxy for a companies' fixed cost share. This follows other literature such as De Loecker, Eeckhout, and Unger (2020). The figure shows rising fixed cost shares which is consistent with rising scale economies at the firm level. Administration expenses in UK company accounts are the costs incurred by a company that are not directly related to the production, manufacture or sale of goods or services. In the Appendix we discuss the data in greater detail and provide examples of administrative costs.

Figure 5: Median Fixed Cost Share in Sales, Source: BvD FAME



The plot shows the median 'Administration Expenses' share in 'Turnover' for UK firms.

3 Model

The household side of the model follows a neoclassical growth setup. The production side of the economy has firm entry and exit, monopolistic competition, and production functions that have fixed costs and returns to scale.

3.1 Households

A representative household maximizes lifetime utility subject to a budget constraint

$$\max_{\{C_t, K_{t+1}\}_{t=0}^{\infty}} \quad \sum_{t=0}^{\infty} \beta^t \frac{C_t^{1-\sigma} - 1}{1-\sigma}, \quad \beta \in (0,1),$$

$$\text{s.t.} \quad C_t + I_t = r_t K_t + w_t L^s + \Pi_t + T_t \tag{1}$$

$$I_t = K_{t+1} - (1-\delta)K_t. \tag{2}$$

Households own all firms in the economy and receive profits Π_t. T_t is a lump sum transfer from the government that will be equal to the entry fees paid by the firms. Households supply a fixed amount of labour that is not time-varying, we normalize this to one:

$$L^s = 1. \tag{3}$$

Households own the capital stock and rent it to firms at a rental rate r_t, hence the capital investment decision is part of the household problem. The household optimization problem satisfies the following condition

$$\left(\frac{C_{t+1}}{C_t}\right)^{\sigma} = \beta(r_{t+1} + (1-\delta)). \tag{4}$$

plus a transversality condition and the resource constraint.

3.2 Firms

3.2.1 Final goods producer

The final goods aggregator is

$$Y_t = N_t \left[\frac{1}{N_t} \int_0^{N_t} y_t(\imath)^{\frac{1}{\mu}} d\imath \right]^{\mu}. \tag{5}$$

There are N_t intermediate producers on the interval $\imath \in (0, N_t)$. The parameter $\mu \geq 1$ captures product substitutability.[13] The aggregator has constant returns to scale.[14]

The maximization problem of the final goods producer is

$$\Pi_t^F = \max_{y_t(\imath)} \quad Y_t - \int_0^{N_t} p_t(\imath) y_t(\imath) d\imath \tag{6}$$

$$\text{s.t.} \quad Y_t = N_t \left[\frac{1}{N_t} \int_0^{N_t} y_t(\imath)^{\frac{1}{\mu}} d\imath \right]^{\mu} \tag{7}$$

The firm is infinitesimal so firm level output does not a ect Y_t. The first-order condi-

[13] Perfectly substitutable products $\mu = 1$ are admissible when intermediate producers have a fixed cost and increasing marginal cost ($\phi > 0$ and $\nu \in (0,1)$). This is the case of perfect competition where profit maximizing intermediate producers take price as given. Under perfect competition all firms produce at the minimum on their average cost curves with perfectly-elastic, horizontal, demand curves.

[14] A typical CES production function would have the pre-multiplying term as N_t^{μ}, such that is cancels with the $1/N_t$ inside the square brackets. However, this creates increasing scale economies in aggregation. Since our interest is scale economies at the firm level, we remove this additional source of scale in aggregation.

tion with respect to $y_t(\imath)$ gives the inverse-demand for a firm

$$p_t(\imath) = \left(\frac{N_t y_t(\imath)}{Y_t}\right)^{\frac{1-\mu}{\mu}}. \tag{8}$$

3.2.2 Intermediate goods producer

The timeline for the intermediate goods producer is as follows. The firm pays cost κ to enter. It receives a draw $\jmath \in (0,1)$ from an i.i.d uniform distribution which translates to productivity $A(\jmath)$. It then decides whether to produce which incurs a fixed overhead cost. If the firm does not produce it remains inactive which we refer to as endogenous exit. All firms, active and inactive, exit at the end of one period.

The production function for a firm with productivity $\jmath$ is given by

$$y_t(\jmath) = A(\jmath)\left[k_t(\jmath)^{\alpha}\ell_t(\jmath)^{1-\alpha}\right]^{\nu}. \tag{9}$$

The parameter $0 < \alpha < 1$ captures the capital cost in total variable cost. The parameter $\nu > 0$ captures returns to scale in variable inputs. This represents returns to scale in variable inputs which captures the slope of the marginal cost curve. There are decreasing returns in variable production when $\nu \in (0,1)$, constant returns when $\nu = 1$, and increasing returns when $\nu > 1$. As $\nu : 0 \to 1$ the marginal cost curve flattens which raises returns to scale, when $\nu = 1$ the marginal cost curve is flat, and as $\nu : 1 \to \infty$ the marginal cost curve is increasingly downward sloping.[15] The labour employed to produce output is:

$$\ell_t(\jmath) = \ell_t^{\text{tot}}(\jmath) - \phi, \tag{10}$$

where $\ell_t^{\text{tot}}(\jmath)$ represents the total labour employed by the firm, and ϕ is an overhead cost.[16] Both ϕ and ν determine scale economies.

[15] We show that downward sloping MC curve must be shallower than the downward sloping demand curve to ensure a profit-maximizing equilibrium where $MR = MC$ exists.

[16] We follow related theoretical literature in using labour-denominated overhead costs (Melitz 2003; Hopenhayn, Neira, and Singhania 2022). Dhyne, Kikkawa, Komatsu, Mogstad, and Tintelnot (2022) present empirical evidence of sizable fixed overhead costs in labour for Belgian firms. An output-denominated overhead cost, as in Savagar (2021), limits the tractability of our analytical results.

The firm solves the following profit maximization problem:

$$\max_{k_t(j),\ell_t(j)} p_t(j)y_t(j) - r_t k_t(j) - w_t(\ell_t(j) + \phi) \tag{11}$$

subject to the production function (9) and inverse demand function (8). The optimality conditions imply constant factor shares in revenue:

$$\frac{r_t k_t(j)}{p_t(j)y_t(j)} = \frac{\nu}{\mu}\alpha \tag{12}$$

$$\frac{w_t \ell_t(j)}{p_t(j)y_t(j)} = \frac{\nu}{\mu}(1-\alpha). \tag{13}$$

For the second-order conditions on profit maximization to hold, a necessary condition is: $\nu < \mu$. We present the first- and second-order conditions in Appendix C.1. Additionally, we assume $\alpha\nu < 1$.[17] Therefore, we assume the following upper-bound on returns to scale in variable inputs.

Assumption 1. Increasing returns in variables inputs are limited as follows:

$$\nu < \min\left\{\frac{1}{\alpha}, \mu\right\}. \tag{14}$$

A higher markup and a lower capital cost share in variable costs allow for greater returns to scale in variable inputs.

From the factor market equilibrium conditions, the ratio $\nu/\mu = (w\ell + rk)/py$ is variable cost share in revenue. The remaining share, $1-(\nu/\mu)$, is the profit plus fixed cost share in revenue. Additionally, $\alpha = rk/(w\ell + rk)$ and $1-\alpha = w\ell/(w\ell + rk)$ are the share of capital and production labour in variable costs. Also, $\alpha\nu = \mu(rk/py)$ is the capital share in revenue scaled by the markup.

[17]This assumption is not required for profit maximization to hold. Imperfect competition ensures that firm-level revenue is concave in inputs, even if output is not concave in inputs. That is, marginal revenue products are decreasing in their respective inputs, even if marginal products are not. Specifically, $0 < \alpha\nu < 1$ ensures firm-level output is concave in capital, and aggregate output is concave in aggregate capital and not decreasing in aggregate labour.

3.2.3 Ratio of firm size

Firm output, revenue and inputs are proportional to productivity to the power of a constant $y(j)^{\frac{1}{\mu}}, p(j)y(j), k(j), \ell(j) \propto A(j)^{\frac{1}{\mu-\nu}}$. Consequently, for a given distribution of $A(j)$ across firms, changes in μ and ν a ect the distribution of labour, capital, revenue and output across firms.

The inverse demand condition and factor price equilibrium conditions imply that for any two firms, $\imath$ and $\jmath$, their relative revenue and input choices are proportional to their relative (scaled) productivity:

$$\frac{p_t(\jmath)y_t(\jmath)}{p_t(\imath)y_t(\imath)} = \frac{k_t(\jmath)}{k_t(\imath)} = \frac{\ell_t(\jmath)}{\ell_t(\imath)} = \left(\frac{A(\jmath)}{A(\imath)}\right)^{\frac{1}{\mu-\nu}}, \quad \forall \imath, \jmath. \tag{15}$$

Additionally, if we use equation (8) to substitute out p_t, we can write:

$$\frac{y_t(\jmath)}{y_t(\imath)} = \left(\frac{A(\jmath)}{A(\imath)}\right)^{\frac{\mu}{\mu-\nu}}. \tag{16}$$

3.2.4 Zero-profit firm

We assume there is a threshold productivity draw $J_t \in (0,1)$ characterised by zero profits, which yields threshold technology $\underline{A}_t$. If a firm receives a productivity draw below the threshold productivity level they would make negative profits from production. Consequently, they prefer to produce zero and make zero profits. Therefore we define profits and characterise the threshold productivity as follows:

$$\pi_t(\jmath) = p_t(\jmath)y_t(\jmath) - r_t k_t(\jmath) - w_t(\ell_t(\jmath) + \phi) \tag{17}$$

$$\pi_t(J_t) = 0. \tag{18}$$

A helpful reduced-form expression for profits combines the profit condition with equilibrium factor prices, with the zero-profit condition and with the ratio of revenues to scaled productivity:

$$\pi_t(\jmath) = \phi w_t \left[\left(\frac{A(\jmath)}{\underline{A}_t} \right)^{\frac{1}{\mu-\nu}} - 1 \right]. \tag{19}$$

3.2.5 Free Entry

All firms die after one period. A firm produces if it makes positive profits, hence firm value is given by

$$v_t(\jmath) = \max\{\pi_t(\jmath), 0\}. \tag{20}$$

We assume a free entry condition which implies that the unconditional expected value from entering equals to the entry cost κ:

$$\mathbb{E}[v_t(\jmath)] = \kappa. \tag{21}$$

The cost of entry κ is denominated in consumption units and is rebated to households in a lump-sum. Combining (20) and (21) with our reduced-form profit expression (19) yields:

$$\phi w_t (1 - J_t) \left[\left(\frac{\hat{A}_t}{\underline{A}_t} \right)^{\frac{1}{\mu-\nu}} - 1 \right] = \kappa. \tag{22}$$

This shows that profits from being active multiplied by the probability of being active $1 - J_t$ equals the entry cost. We have defined the power mean of technology, conditional on being active, as

$$\hat{A}(J_t) \equiv \mathbb{E}\left[\left(A(\jmath)^{\frac{1}{\mu-\nu}} \;\middle|\; \jmath > J_t \right) \right]^{\mu-\nu} = \left[\left(\frac{1}{1 - J_t} \int_{J_t}^{1} A(\jmath)^{\frac{1}{\mu-\nu}} \, d\jmath \right) \right]^{\mu-\nu}. \tag{23}$$

The power mean is a weighted average of firm-level productivity.[18]

[18] The term $\hat{A}(J_t)$ generalizes Melitz (eq. 7 2003, p. 1700) and Colciago and Silvestrini (eq. 31 2022, p. 10). This term is equivalent to these papers if $\nu = 1$ and the markup is expressed in terms of elasticities of substitution between goods, for example $\mu = \theta/(\theta - 1)$ where θ is the elasticity parameter. Notably, with $\nu \neq 1$, we *cannot* represent the power mean of technology $\hat{A}$ as an output-weighted harmonic average of unscaled technology draws.

3.3 Entry

Operating firms N_t are the subset of firms who decide to produce once receiving their productivity draw. Entrants E_t are all firms who pay the entry cost.

$$N_t = \int_0^{N_t} d\imath = E_t \int_{J_t}^1 d\jmath = E_t(1 - J_t). \tag{24}$$

We can interpret the productivity cut-off J_t as the probability of exit and $1 - J_t$ as the probability of surviving.

3.4 Aggregation

To obtain aggregate output and aggregate inputs, we use that the index of operating firms $(0, N_t)$ is equivalent to the measure of entering firms E_t restricted over the region of operation $(J_t, 1)$.

3.4.1 Aggregate Factor Inputs

Aggregate labour is comprised of production labour and non-production labour

$$K_t = \int_0^{N_t} k_t(\imath) d\imath = E_t \int_{J_t}^1 k_t(\jmath) d\jmath \tag{25}$$

$$L_t = \int_0^{N_t} [\ell_t(\imath) + \phi]\, d\imath = E_t \int_{J_t}^1 [\ell_t(\jmath) + \phi]\, d\jmath. \tag{26}$$

We define u_t as the fraction of aggregate labour that goes to production

$$u_t \equiv \frac{E_t \int_{J_t}^1 \ell(\jmath) d\jmath}{L_t} = \frac{\int_0^{N_t} \ell_t(\imath) d\imath}{L_t} \tag{27}$$

$$1 - u_t = \frac{E_t(1 - J_t)\phi}{L_t} = \frac{N_t \phi}{L_t}. \tag{28}$$

3.4.2 Aggregate Output

We can express aggregate output as:

$$Y_t = N_t \hat{A}_t \left[(K_t/N_t)^{\alpha} (u_t L_t/N_t)^{1-\alpha} \right]^{\nu} = N_t^{1-\nu} \hat{A}_t \left[K_t^{\alpha} (u_t L_t)^{1-\alpha} \right]^{\nu}. \tag{29}$$

The first expression shows that aggregate output is the sum across N_t homogeneous firms each with average technology $\hat{A}_t$. The aggregate output expression is homogeneous of degree one in capital, production labour and number of firms, which implies there are constant returns in these factors. If N_t is treated as a fixed factor, then the function is homogeneous of degree ν in capital and production labour. In other words, external returns to scale in aggregate capital and production labour are given by ν.

3.4.3 Aggregate Factor Market Equilibrium

The wage, rental rate on capital and zero-profit condition are

$$r_t = \alpha \frac{\nu}{\mu} \frac{Y_t}{K_t} \tag{30}$$

$$w_t = (1-\alpha) \frac{\nu}{\mu} \frac{Y_t}{u_t L_t} \tag{31}$$

$$\frac{w_t}{Y_t} \frac{N_t \phi}{L_t} = \left(1 - \frac{\nu}{\mu}\right) \left(\frac{\underline{A}_t}{\hat{A}_t}\right)^{\frac{1}{\mu-\nu}} \tag{32}$$

3.5 Government Budget Constraint and Resource Constraints

The resource constraint is

$$Y_t = C_t + I_t. \tag{33}$$

The government rebates entry fees to households. The government budget constraint equates taxes to government expenditure

$$T_t = E_t \kappa. \tag{34}$$

Profits and labour markets clear:

$$\Pi_t = \Pi_t^F \tag{35}$$

$$L_t = L^s. \tag{36}$$

Aggregate profits received by the household from owning firms equate to profits earned by the final goods producer. The profits are zero in equilibrium. Labour demanded by the firm equates to labour supplied by the household which is normalised to 1.

3.6 Equilibrium Definition

An equilibrium is a sequence of prices $\{r_t, w_t\}_{t=0}^{\infty}$; firm capital and labour demands $\{\ell_t(\jmath), k_t(\jmath)\}_{t=0}^{\infty}$; firms' operating decisions to be active or inactive, measures of entry and active firms $\{E_t, N_t\}_{t=0}^{\infty}$; consumption and capital $\{C_t, K_{t+1}\}_{t=0}^{\infty}$, such that

1. households choose C and K optimally by solving problem (1);
2. firms compete decide optimally whether to produce or remain inactive, and demand factors according to (11);
3. the free entry condition holds (21);
4. markets clear for aggregate labour (26), aggregate capital (25), goods market (33), labour market (36) and aggregate profits (35);
5. the government budget constraint is satisfied (34).

3.7 Model Characteristics

Aggregation allows us to remove individual firm heterogeneity $\jmath$ from the model. This does not mean heterogeneity is irrelevant. There would be no selection e ect without heterogeneity. But aggregation allows us to summarise all the heterogeneity in one term $\hat{A}_t$, and then solve the model. In other words, the model economy with individual heterogeneity is isomorphic to the model with homogeneous firms, each endowed with the power mean of technology $\hat{A}_t$. Before imposing a Pareto distribution on the technology draws $A(\jmath)$, we characterise some general properties of the model.

3.7.1 Aggregate Labour Utilized for Production

From (31) and (32), and using $N_t\phi/L_t = 1 - u_t$, we get the level of aggregate labour utilized in production as a function of J:

$$u_t = \left[1 + \frac{1}{1-\alpha}\left(\frac{\mu}{\nu} - 1\right)\left(\frac{\underline{A}_t}{\hat{A}_t}\right)^{\frac{1}{1-\nu}}\right]^{-1}.$$

In turn, by equation (31) this implies that the aggregate labour share w_tL_t/Y_t is only a function of J_t through the $\hat{A}/\underline{A}$ ratio.

3.7.2 Aggregate Productivity

We can rearrange aggregate output into Cobb-Douglas form which gives:

$$Y_t = \text{TFP}_t K_t^{\alpha\nu} L_t^{1-\alpha\nu} \tag{37}$$

$$\text{where,} \quad \text{TFP}_t \equiv \left(\frac{N_t}{L_t}\right)^{1-\nu}\left(1 - \frac{N_t\phi}{L_t}\right)^{(1-\alpha)\nu}\hat{A}_t \tag{38}$$

$$= \left(\frac{1-u_t}{\phi}\right)^{1-\nu} u_t^{(1-\alpha)\nu}\hat{A}_t \tag{39}$$

Aggregate total factor productivity (TFP) measures aggregate output that is not accounted for by aggregate capital and aggregate labour. TFP is not the Solow residual because the exponents of aggregate capital and labour do not correspond to aggregate factor shares.[19] It is helpful to decompose TFP into allocative efficiency and technical efficiency:

$$TFP_t = \underbrace{\Omega_t}_{\text{allocative}} \times \underbrace{\hat{A}_t}_{\text{technical}}. \tag{40}$$

[19]The term $\alpha\nu$ is the aggregate capital share in output multiplied by the markup $\alpha\nu = \mu \times rK/Y$.

We define $\hat{A}_t$ as *technical efficiency*, and we define *allocative efficiency* as:

$$\Omega_t \equiv \underbrace{\left(\frac{N_t}{L_t}\right)^{1-\nu}}_{\text{Scale effect}} \times \underbrace{\left(1 - \frac{N_t \phi}{L_t}\right)^{(1-\alpha)\nu}}_{\text{Resource duplication}}.$$

Allocative efficiency captures the negative effect of more firms duplicating fixed costs, and the scale effect of dividing aggregate labour among more firms, which will depend on returns to scale $\nu \gtreqless 1$. Technical efficiency is the generalised mean, conditional on being active, of exogenously drawn technology. It is determined by selection. Under Pareto distributed $A(j)$, technical efficiency is a linear function of the threshold productivity level $\underline{A}$.[20]

3.7.3 Scale Economies

The parameters ν and ϕ are both sources of scale economies in the model. Scale economies are measured as the ratio of average cost to marginal cost (the inverse cost elasticity). In this section, we show this from the production side by summing output elasticities. The same result can be shown from the cost function.[21]

From equations (9) and (10), the response of firm output to a change in each variable input is constant. Consequently, returns to scale in variable inputs is constant:

$$\frac{\partial \ln y_t(j)}{\partial \ln k_t(j)} = \nu\alpha, \quad \frac{\partial \ln y_t(j)}{\partial \ln \ell_t(j)} = \nu(1-\alpha), \quad \frac{\partial \ln y_t(j)}{\partial \ln k_t(j)} + \frac{\partial \ln y_t(j)}{\partial \ln \ell_t(j)} = \nu.$$

The effect of a change in total labour input is decreasing in firm size:[22]

$$\frac{\partial \ln y_t(j)}{\partial \ln \ell_t^{\text{tot}}(j)} = \nu(1-\alpha)\left(1 + \frac{\phi}{\ell_t(j)}\right) = \nu(1-\alpha) + (\mu - \nu)\left(\frac{\underline{A}_t}{A(j)}\right)^{\frac{1}{\mu-\nu}} \in (\nu(1-\alpha), \mu - \alpha\nu).$$

[20]Our TFP decomposition is similar to Jaimovich, Terry, and Vincent (2023), but they do not have a resource duplication effect from entry. They study the effect of an output subsidy on the components.

[21]Savagar (2021) shows this for a model with output denominated fixed costs.

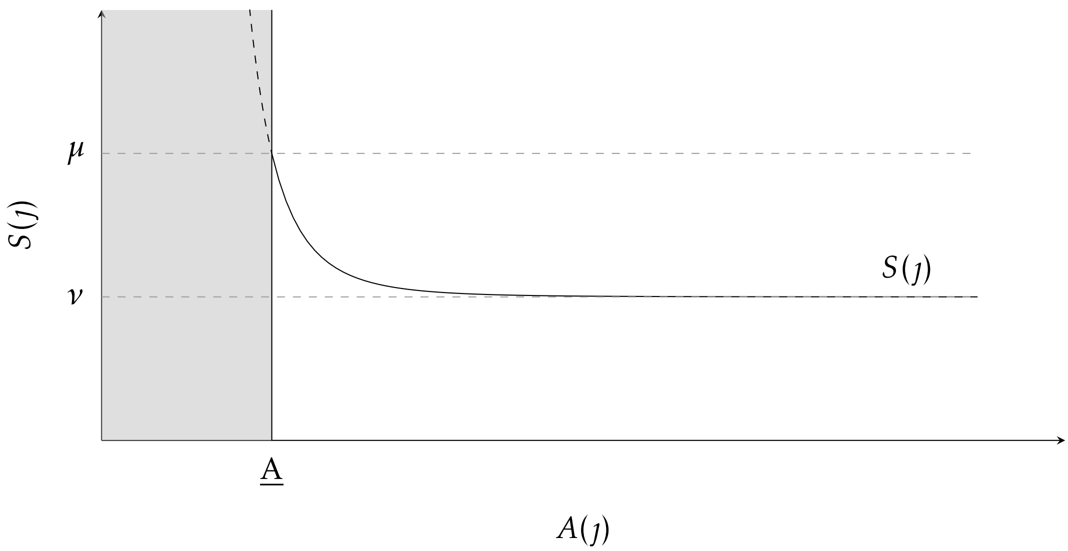
[22]For the second equality, we use the zero-profit condition $\left(1 - \frac{\nu}{\mu}\right) p_t(j) y_t(j) = w_t \phi \left(\frac{A(j)}{\underline{A}_t}\right)^{\frac{1}{\mu-\nu}}$ combined with labour demand $\frac{w_t}{p_t(j) y_t(j)} = \frac{\nu(1-\alpha)}{\mu} \frac{1}{\ell(j)}$ to yield $\nu(1-\alpha)\frac{\phi}{\ell_t(j)} = (\mu - \nu)\left(\frac{\underline{A}_t}{A(j)}\right)^{\frac{1}{\mu-\nu}}$.

Therefore, scale economies at the firm are decreasing in firm size:

$$S_t(j) \equiv \frac{\partial \ln y_t(j)}{\partial \ln k_t(j)} + \frac{\partial \ln y_t(j)}{\partial \ln \ell_t^{tot}(j)} = \nu\left(1 + (1-\alpha)\frac{\phi}{\ell_t(j)}\right) = \nu + (\mu - \nu)\left(\frac{\underline{A}_t}{A(j)}\right)^{\frac{1}{\mu-\nu}} \in (\nu, \mu). \quad (41)$$

A firm's scale economies decrease as production labour rises relative to the labour overhead, or as firm productivity rises relative to the productivity cut-off. Figure 6 plots (41) for a given $\underline{A}$. More productive firms have lower scale economies. The cut-off firm has the highest level of scale equals to the markup, and scale converges on returns to scale in variable inputs ν for high-productivity firms.

Figure 6: Firm-level Scale Economies in Steady-State



Plot shows equation (41) scale of a firm given its productivity draw. In the shaded region firms are inactive and the dashed line shows their hypothetical scale economies if they were to produce. The horizontal lines show the bounds on scale economies of active firms $S(j) \in (\nu, \mu)$. We have assumed $A(j)$ is Pareto distribution and we have set $\underline{A}$ arbitrarily.

3.8 Model with Pareto Distribution

We assume that the technology variable is Pareto distributed. Given a random variable j drawn from the uniform distribution on the unit interval $[0, 1)$, then the productivity variable $A(j)$ given by the quantile function:

$$A(j) = \frac{h}{(1-j)^{\frac{1}{\vartheta}}}. \quad (42)$$

The parameter $\vartheta > 1$ is the Pareto shape parameter and h is the scale parameter, which is the lowest value of technology, corresponding to $\jmath = 0$. We set $h = 1$. A thicker-tailed Pareto distribution occurs as $\vartheta \to 1$, which implies a higher density of high-productivity draws and a lower density of low-productivity draws. A thinner-tailed Pareto distribution occurs as $\vartheta \to \infty$ which implies a lower density of high-productivity draws and a higher density of low-productivity draws.

Under Pareto, the power mean of technology is:

$$\hat{A}_t = \left(\frac{\vartheta(\mu-\nu)}{\vartheta(\mu-\nu)-1}\right)^{\mu-\nu} \underline{\mathrm{A}}_t = \Gamma \underline{\mathrm{A}}_t \quad \text{where } \Gamma \equiv \left(\frac{\vartheta(\mu-\nu)}{\vartheta(\mu-\nu)-1}\right)^{\mu-\nu}. \tag{43}$$

The constant Γ is the unconditional expectation of scaled technology $A(\jmath)^{\frac{1}{\mu-\nu}}$. If the cuto took its minimum value $\underline{\mathrm{A}}_t = 1$, such that all participants were active and there was no selection $J_t = 0$, this represents the average technology that would arise. To ensure that scaled technology $A(\jmath)^{\frac{1}{\mu-\nu}}$ has a finite expectation, we require that the scaled Pareto shape parameter satisfies the following assumption.

$$\vartheta(\mu-\nu) > 1. \tag{44}$$

This limits the degree of fat tails in the technology distribution. The assumption is analogous to the assumption $\vartheta > 1$ for the Pareto distributed technology before it is scaled.

3.8.1 Equilibrium Conditions with Pareto Distribution

Given the constant ratio between the power mean of technology and cut-o technology in equation (43), several equilibrium conditions simplify. Labour utilized for production is constant, aggregate TFP is a linear function of cut-o technology, and wage is a log-linear function of cut-o technology:

$$u = \left(1 + \frac{\vartheta(\mu-\nu)-1}{\nu\vartheta(1-\alpha)}\right)^{-1} \qquad 1-u = \frac{\vartheta(\mu-\nu)-1}{\vartheta(\mu-\alpha\nu)-1} \tag{45}$$

$$TFP_t = \Omega \hat{A}_t, \quad \text{where } \Omega \equiv \left(\frac{1-u}{\phi}\right)^{1-\nu} u^{(1-\alpha)\nu} \text{ and } \hat{A}_t = \Gamma \underline{A}_t \tag{46}$$

$$w_t = \frac{\kappa}{\phi}\left[\vartheta(\mu-\nu)-1\right]\underline{A}_t^{\vartheta}. \tag{47}$$

The final equation determines the wage from the free entry condition. The lowest value $\underline{A}_t$ can take is 1 which is the lowest productivity draw corresponding to $J = 0$. The constant u implies that total production labour is always a fixed fraction of aggregate labour as an economy transitions over time. Labour utilized for production is invariant to the fixed cost, increasing in returns to scale, and decreasing in the markup:

$$\frac{du}{d\phi} = 0, \qquad \frac{du}{d\nu} = \frac{(1-\alpha)\vartheta(\vartheta\mu-1)}{(\vartheta(\mu-\alpha\nu)-1)^2} > 0, \qquad \frac{du}{d\mu} = -\frac{(1-\alpha)\vartheta^2\nu}{(\vartheta(\mu-\alpha\nu)-1)^2} < 0. \tag{48}$$

The constant u implies that the number of active firms is constant

$$N = \frac{1-u}{\phi} = \frac{1}{\phi}\frac{\vartheta(\mu-\nu)-1}{\vartheta(\mu-\alpha\nu)-1}. \tag{49}$$

Therefore, we can characterise the number of active firms as decreasing in the fixed cost and returns to scale, and increasing in the markup:

$$\frac{dN}{d\phi} = -\frac{N}{\phi} < 0, \qquad \frac{dN}{d\nu} = -\frac{1}{\phi}\frac{du}{d\nu} < 0, \qquad \frac{dN}{d\mu} = \frac{1}{\phi}\frac{du}{d\mu} > 0. \tag{50}$$

As the marginal cost curve becomes flatter $\nu < 1$, horizontal $\nu = 1$ and downward sloping $\nu > 1$, optimal firm size (MR=MC) increases, and more total labour goes toward production (u rises). With larger firms the number of firms declines. An increase in fixed cost ϕ does not alter the fraction of production labour in total labour, so the number of firms must decrease to keep the ratio of total fixed costs to labour fixed. An increase in the markup increases the number of firms because the demand curve becomes steeper which reduces optimal size, consequently there is more duplication of the fixed cost and the production labour share in total labour falls.[23]

An implication of constant u and N is that the aggregate labour share $w_t L_t / Y_t$ is constant:

$$s_L \equiv \frac{wL}{Y} = \frac{1}{\mu}\left(\mu - \alpha\nu - \frac{1}{\vartheta}\right).$$

The labour share is increasing in the markup, decreasing in returns to scale and invariant to the fixed cost.

The equilibrium conditions under Pareto reduce to a dynamic system in $\{K_t, C_t\}$:

$$\Omega\Gamma\Psi K_t^{\frac{\alpha\nu\vartheta}{\vartheta-1}} - C_t = K_{t+1} - (1-\delta)K_t \tag{51}$$

$$\left(\frac{C_{t+1}}{C_t}\right)^{\sigma} = \beta\left[\left(\alpha\frac{\nu}{\mu}\Omega\Gamma\Psi K_{t+1}^{\frac{\alpha\nu\vartheta}{\vartheta-1}-1} + (1-\delta)\right]\right(\tag{52}$$

where Ω, Ψ, Γ are constants.[24] We impose the following:

$$1 - \vartheta(1-\alpha\nu) < 0. \tag{53}$$

This ensures that aggregate output is concave in aggregate capital, and therefore the price of capital is decreasing in aggregate capital. From equations (44) and (53), we have limited the thickness of the Pareto tail by making two assumptions, which we summarise below.

Assumption 2. The Pareto shape parameter must satisfy

$$\frac{1}{\vartheta} < \min\{\mu - \nu, 1 - \alpha\nu\}. \tag{54}$$

[23]This is the result of excess entry of 'small' firms under monopolistic competition (Dixit and Stiglitz 1977; Mankiw and Whinston 1986).

[24]Full derivation in appendix.

3.8.2 Steady-state with Pareto Distribution

In steady state the system satisfies $K_{t+1} = K_t = K$ and $C_{t+1} = C_t = C$. This yields the following steady-state solution for capital and consumption:

$$K = \left[\frac{\alpha\nu\Omega\Gamma\Psi}{\mu r}\right]^{\frac{\vartheta-1}{\vartheta(1-\alpha\nu)-1}} \tag{55}$$

$$C = K\left(\frac{\mu r}{\alpha\nu} - \delta\right) \tag{56}$$

where $r = \frac{1}{\beta} - (1-\delta)$. The remaining steady-state variables follow by substituting the expression for K into the reduced model, which we present in the appendix. In particular, solving for the technology threshold $\underline{\mathrm{A}}$ yields:

$$\underline{\mathrm{A}} = \left[\nu^{\nu}\frac{1}{\mu}\left(\frac{\alpha}{r}\right)^{\alpha\nu}(\phi(1-\alpha))^{\nu(1-\alpha)}\,\vartheta^{\mu-1}(\mu-\nu)^{\mu-\nu}\frac{1}{\kappa^{1-\alpha\nu}}\frac{1}{(\vartheta(\mu-\nu)-1)^{\mu-\alpha\nu}}\right]^{\frac{1}{\vartheta(1-\alpha\nu)-1}}. \tag{57}$$

4 Theoretical Analysis

Changes in aggregate productivity occur through an allocation component $d\ln\Omega$ and a technical efficiency component $d\ln\hat{A}$:

$$d\ln TFP = d\ln\Omega + d\ln\hat{A}$$

4.1 The E ect of Entry Cost on Aggregate Productivity

The entry cost κ does not a ect allocative efficiency Ω, but a ects technical efficiency $\hat{A}$. If the entry cost increases, then technical efficiency decreases because the threshold technology level falls, thus weakening selection.[25] Selection weakens as the entry cost increases because, by the free-entry condition, the expected value of the firm must increase. The expected value increases if the threshold productivity declines.

[25] Barseghyan and DiCecio (2011) study this in a perfectly competitive economy, where the entry cost is in terms of output κ/Y. They find empirical evidence that higher entry costs decrease aggregate TFP across countries.

4.2 The E ect of Fixed Costs on Aggregate Productivity

Changes in fixed costs a ect aggregate TFP through an allocation component and a technology component:

$$\frac{d\ln TFP}{d\ln\phi} = \frac{d\ln\Omega}{d\ln\phi} + \frac{d\ln\hat{A}}{d\ln\phi}$$

Under Pareto, technical efficiency depends on the technology threshold $\underline{A}$ only since the constant Γ is invariant to ϕ, therefore:

$$\frac{d\ln\hat{A}}{d\ln\phi} = \frac{d\ln\Gamma}{d\ln\phi} + \frac{d\ln\underline{A}}{d\ln\phi} = 0 + \frac{\nu(1-\alpha)}{\vartheta(1-\alpha\nu)-1} > 0.$$

The technology threshold is increasing in the overhead cost if $\vartheta(1-\alpha\nu)-1 > 0$. This is the condition for the rental rate r to be decreasing in aggregate capital.

The allocation e ect depends on the degree of returns to scale in variable production:

$$\frac{d\ln\Omega}{d\ln\phi} = -(1-\nu).$$

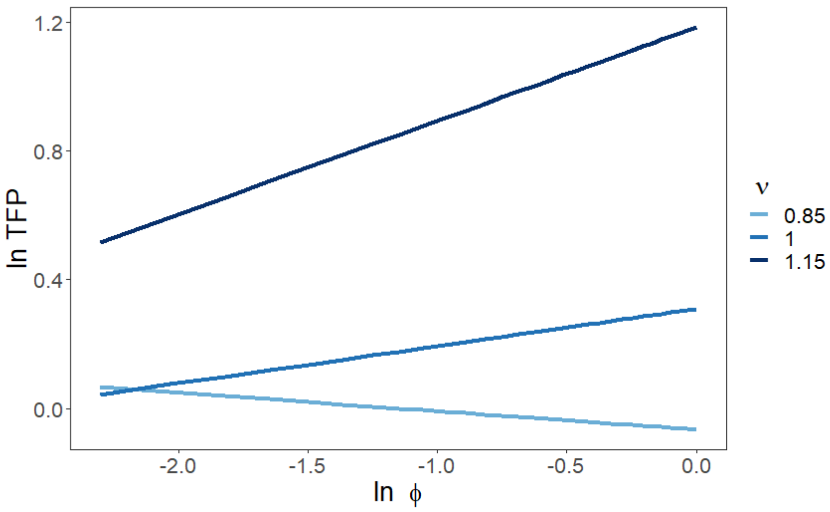
The result is independent of the Pareto distribution assumption. We can interpret the allocation e ect through the number of firms. Note that $\Omega = \left(\frac{1-u}{\phi}\right)^{1-\nu} u^{(1-\alpha)\nu} = N^{1-\nu}u^{(1-\alpha)\nu}$ and u is independent of ϕ. An increase in ϕ, decreases the number of active firms. With increasing returns ($\nu > 1$), allocative efficiency is improved by having fewer firms, as they benefit more from the increasing returns. On the other hand, with decreasing returns ($\nu < 1$), then having fewer firms is detrimental to allocative efficiency, as the e ect of decreasing returns is accentuated. Lastly, with constant returns ($\nu = 1$), the number of firms has no e ect on allocative efficiency.

Combining the allocative and technical efficiency e ects, shows that the response of aggregate TFP to a change in fixed costs will depend on the level of returns to scale in variable inputs ν.

$$\frac{d\ln TFP}{d\ln\phi} = -(1-\nu) + \frac{\nu(1-\alpha)}{\vartheta(1-\alpha\nu)-1} \tag{58}$$

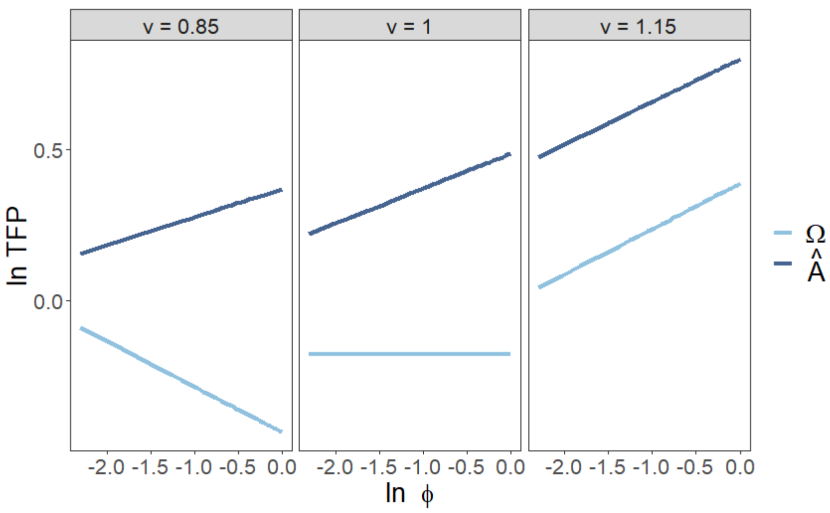
Figure 7 simulates equation (58) for di erent values of ν based on our benchmark calibration (Table 1).

Figure 7: E ect of ϕ on TFP for di erent ν



In Figure 8, we decompose the three cases from Figure 7.[26] Technical efficiency always rises as the fixed cost increases, while the allocative efficiency component is determined by $\nu \gtreqless 1$, as previously discussed.

Figure 8: E ect of $\ln \phi$ on TFP decomposed into $\hat{A}$ and Ω for di erent ν



[26]The e ect of ϕ on TFP is the same regardless of μ.

4.3 The Effect of Returns to Scale on Aggregate Productivity

The nonlinearity of Equation (57) in ν makes it difficult to obtain a closed-form expression for the influence of ν on TFP. Therefore, we present simulations to illustrate this effect. The model is calibrated as in Table 1. We set ν and μ to the middle of the range presented as a baseline, but allow the evolution of both variables as estimated from the microdata and external sources.

Calibration

Table 1: Parameter Values for Comparative Statics

	Parameter	Value	Target
β	Discount rate	0.96	Real interest rate
δ	Depreciation rate	0.08	Office for National Statistics
ν	Variable RTS	0.99 - 1.05	ABS (authors' estimates)
μ	Markup	1.21 - 1.28	CMA (2022)
α	Capital share	0.25	ABS (authors' calculations)
ϑ	Pareto shape	10	Match firms per worker
κ	Entry cost	0.017	Model-implied maximum given range of ν, μ
ϕ	Overhead cost	0.85	Match share inactive firms

We set the discount factor β to match the average real interest rate of 2.08 percent over the period. To do this, we use the equation for steady-state interest rate $r = \frac{1}{\beta} + 1 - \delta$.[27] The depreciation rate δ is determined by a weighted-average from ONS data. Our estimates of the returns to scale ν come from our estimates of the production function using the estimation of Gandhi, Navarro, and Rivers (2020). Markup estimates are from CMA (2022). They use a different dataset and estimation strategy. These markup estimates are consistent with other studies that show rising markups over this time period (ONS 2022; Hwang, Savagar, and Kariel 2022). Our results are not sensitive to these markup estimates. In the model $\frac{\alpha\nu}{\mu}$ is the capital share in revenue and $\frac{(1-\alpha)\nu}{\mu}$ is

[27] Data on UK long-term government bond and inflation used to compute the real interest rate from FRED database: IRLTLT01GBM156N and FPCPITOTLZGGBR.

the production labour share in revenue. Given our ν and μ estimates, we set $\alpha = 0.25$ to match a capital share of 20%.[28]

The entry cost parameter κ and the fixed cost parameter ϕ must satisfy restrictions such that $J_t \in (0,1)$. The model places an upper limit on κ for values of μ, ν. We set κ at this maximum. We choose ϕ to target the share of 'inactive' firms J_t, to match the share of firms that do not produce but 're-activate' within a two year window. [29] In the UK the average share of 'inactive' firms between 2016 - 2020 was 10%.[30] We also check our calibration of κ and ϕ by looking at the ratio $\kappa/\phi w$. Barseghyan and DiCecio (2011) report a range of values from industry studies. In most industries, the ratio is less than one, so entry costs are less than overhead costs. The average they report is 0.82. Our experiments vary ν, ϕ, μ parameters, so the entry-to-overhead cost ratio will vary as we change these values, but the outcome always remains below 1.

Our theory imposes restrictions on the Pareto shape parameter ϑ. First, $\vartheta > \frac{1}{\mu-\nu}$ which ensures scaled productivity is Pareto distributed and the first moment exists, and second, $\vartheta > \frac{1}{1-\alpha\nu}$ which ensures aggregate output is concave in aggregate capital, so that the interest rate is decreasing in aggregate capital.[31] Our calibrated markup minus our estimated returns to scale $\mu - \nu$ is between 0.198 and 0.234 from 2001 - 2014. Therefore, our restrictions imply that we must set $\vartheta > 5$, similar to Hopenhayn (2014) who sets the Pareto shape between 5 and 10.

We use the number of firms per worker N/L to calibrate the Pareto shape ϑ. Our model yields $\phi N/L = \left(\frac{\vartheta(\mu-\nu)-1}{\vartheta(\mu-\alpha\nu)-1}\right)$. We set the parameters ϕ, α as calibrated. We plug in the estimates for ν, μ. We obtain the number of firms from business population

[28] The ratio ν/μ is the revenue elasticity, which is typically set to 0.85 in US studies (Restuccia and Rogerson 2008; Barseghyan and DiCecio 2011). Hopenhayn (2014) discusses this common calibration. Our estimates for ν divided by our calibrated markup μ yield a ratio from 0.81 to 0.84 between 2001 and 2014.

[29] This is a standard approach by the ONS and OECD to ensure accurate measures of firm deaths. https://www.ons.gov.uk/businessindustryandtrade/business/activitysizeandlocation/datasets/businessdemographyreferencetable

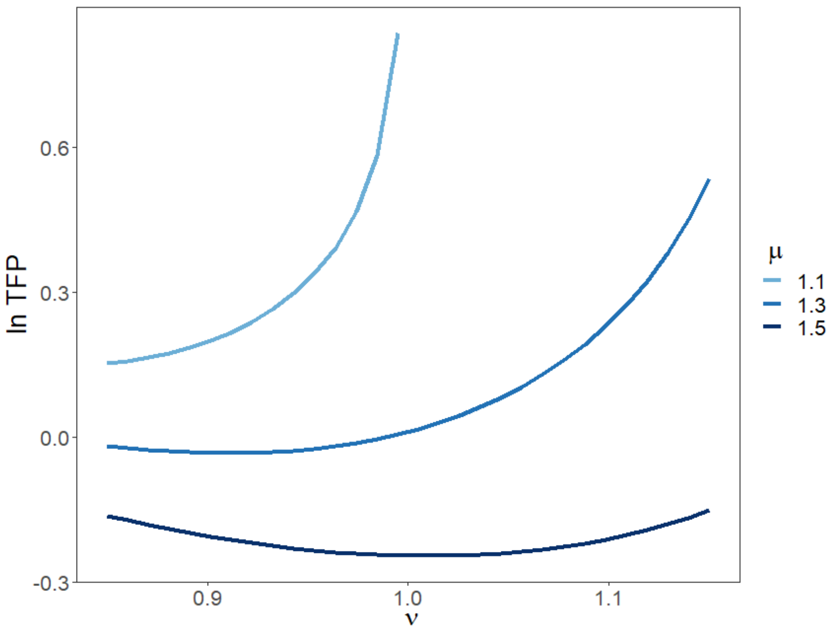
[30] This is the only time frame for which the detail is available.

[31] The first restriction implies that scaled technology, $A(\jmath)^{\frac{1}{\mu-\nu}} = (1-\jmath)^{-\frac{1}{\vartheta(\mu-\nu)}}$, is Pareto distributed. In some experiments, we take $\nu \to \mu$ from below, and this requires us to raise the value of ϑ. The relevant value for us is the scaled Pareto parameter $\vartheta(\mu-\nu)$, since labour is distributed proportionally to this term.

estimates[32] and the number of people employed from the ONS.[33] Combining the data on N/L, which rises from 0.126 to 0.170, we can back-out a series for ϑ. This yields ϑ averaging 8.6 between 2001 - 2004, rising to 10.4 at the end of the sample from 2011 - 2014. It averages 10 over the sample, so we choose this calibration, which is also the upper bound from Hopenhayn (2014). If we allow ϑ to rise in our quantitative exercise, the results are do not change substantively.

Figure 9 shows the e ect of ν on aggregate productivity for di erent values of the markup μ. We observe that aggregate productivity rises unambiguously in ν in a low markup economy, but not when the markup is higher. Both the level and the slope of the relationship is falling in μ.

Figure 9: E ect of variable RTS on ln TFP for di erent levels of the markup



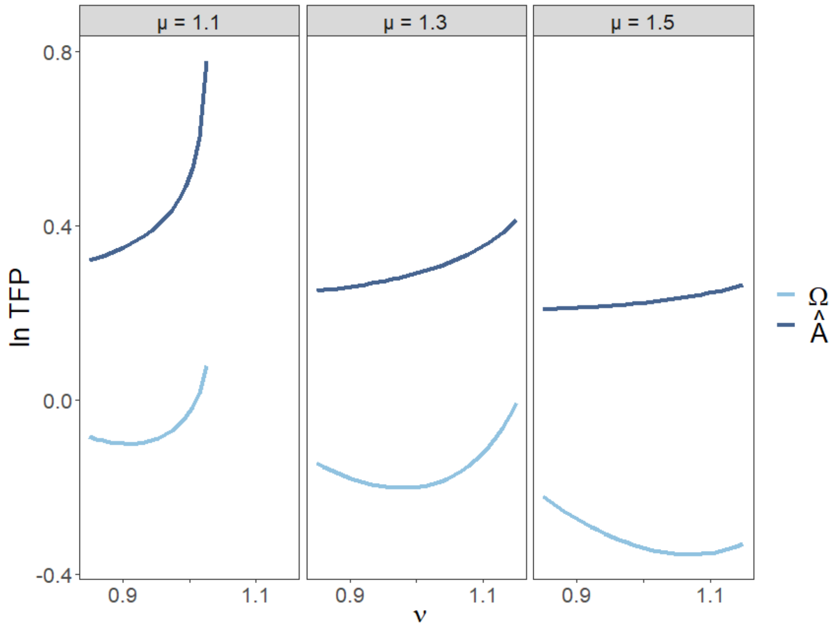
ln TFP for calibrated model for a range of ν and μ.

In Figure 10 we provide a decomposition into technical efficiency and allocative efficiency for each of these markup cases. We observe that the weakening passthrough of returns to scale to TFP occurs because of weakening technical efficiency (i.e. less selection), and worsening allocative efficiency.

[32] https://www.gov.uk/government/statistics/business-population-estimates-2022.

[33] MGRZ series, Labour Market Statistics: https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/employmentandemployeetypes/timeseries/mgrz/lms.

Figure 10: Effect of ν on TFP decomposed into $\hat{A}$ and Ω for different μ



As returns to scale ν increase, technical efficiency $\hat{A}$ increases. This implies stronger selection of high $A(j)$ firms. However, the effect is weaker as market power increases. Hence, in high-markup economies, there is weaker selection of high productivity firms as returns to scale increase.

Returns to scale ν have a U-shaped relationship with allocative efficiency. This occurs because an increase in ν decreases the number of firms. With decreasing returns ($\nu < 1$), fewer firms harm allocative efficiency. However, with increasing returns ($\nu > 1$), fewer firms improve allocative efficiency. As market power increases, the minimum point moves right, causing a wider range of declining allocative efficiency. This occurs because higher markups increase the number of firms. Hence, the benefits of growing returns to scale for allocative efficiency are counteracted by higher markups, reducing the size of firms and limiting their ability to benefit from increasing returns.[34]

[34] In Appendix C.3 we present aggregate output as a function of N_t, which shows these channels formally.

4.4 The Effect of Returns to Scale and Fixed Costs on Aggregate Productivity

To summarise, the impact of changing ν and changing ϕ on aggregate productivity, operates through their effect on the number of active firms N_t, which, in turn, reflects the level of profits for a given productivity draw. Higher returns to scale or fixed costs both reduce profits which reduces the number of active firms. Whereas, a higher markup increases profits which increases the number of active firms. The number of active firms is important for aggregate output because it determines technical efficiency through selection (fewer active firms means more selection), and it affects allocative efficiency through both a scale effect and resource duplication effect.

An important difference between the effects of higher fixed costs and higher returns to scale on aggregate productivity is that the impact of returns to scale is affected by the markup, whereas fixed costs affect aggregate productivity the same regardless of the level of markup.[35] As returns to scale increase their effect of reducing the number of firms and enhancing productivity is strongly mitigated in higher markup environments. This is primarily because the selection effect is weakened, limiting gains in technical efficiency, whilst there is also a drag from allocative efficiency too.

Overall, we conclude that both higher fixed costs or higher returns to scale tend to increase aggregate productivity. However, this unambiguous outcome depends on the presence of increasing returns $\nu > 1$ in levels, which is what we find in our estimation on UK data. In the presence of decreasing returns $\nu < 1$, the effects of higher fixed costs or higher returns to scale are more ambiguous. This is because as these parameters increase, which reduces the number of active firms, aggregate resources are concentrated on a smaller number of firms and these firms are subject to decreasing returns. In an environment of decreasing returns, it is better for aggregate output to spread aggregate resources across more active firms.

[35] As our quantitative exercise will show, higher markups still have a direct effect of increasing the number of firms and reducing productivity, but they do not enhance or diminish the effect of greater fixed costs.

5 Quantitative Application

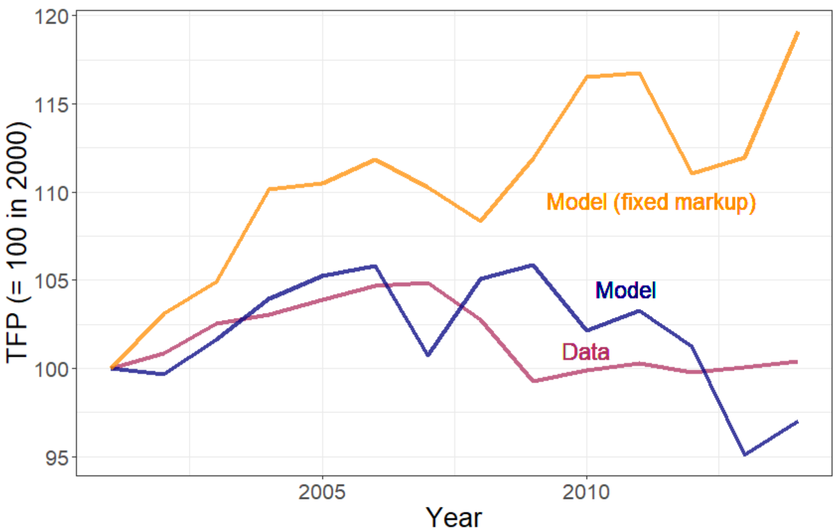
In Section 4, we examined the impact on aggregate productivity of the parameters of the production function that cause scale economies. We concluded that either higher fixed costs or higher returns to scale tend to increase aggregate productivity, though this relies on increasing returns $\nu > 1$ to be unambiguous. We now analyse the quantitative plausibility of scale economies alongside stagnating productivity, which has occured in the US and UK in recent years. We find that changing returns to scale in variable inputs alongside rising markups explains the data well, but rising fixed costs alongside rising markups cannot explain the data as well.

5.1 Rising Returns to Scale

We calibrate the parameter ν to our annual estimates from 2001 to 2014, while the parameter μ is set to annual estimates from CMA 2022. We set $\phi = 0.135$ such that the share of inactive firms is empirically plausible in our benchmark calibration.

Figure 11 compares the trends in TFP in the data and our model. It reveals a rise in both series prior to the Financial Crisis, followed by a sharp decline in the data and a more gradual decrease in the model. Fixing the markup to its 2001 value highlights the significant impact of rising returns to scale on aggregate productivity. If market power had remained constant, higher returns to scale would have boosted aggregate productivity by over 20% between 2001 and 2014. However, when we incorporate the simultaneous increase in markups and returns to scale, our estimated productivity trend aligns more closely with observed data.

Figure 11: TFP Growth: Model vs Data



We give the model estimates of μ and estimates of ν and solve in each year for steady-state to obtain the model-implied TFP. The TFP data series is from the Penn World Table 10.01 (Feenstra, Inklaar, and Timmer 2015), accessed from FRED: Total Factor Productivity at Constant National Prices for United Kingdom (RTFPNAGBA632NRUG).

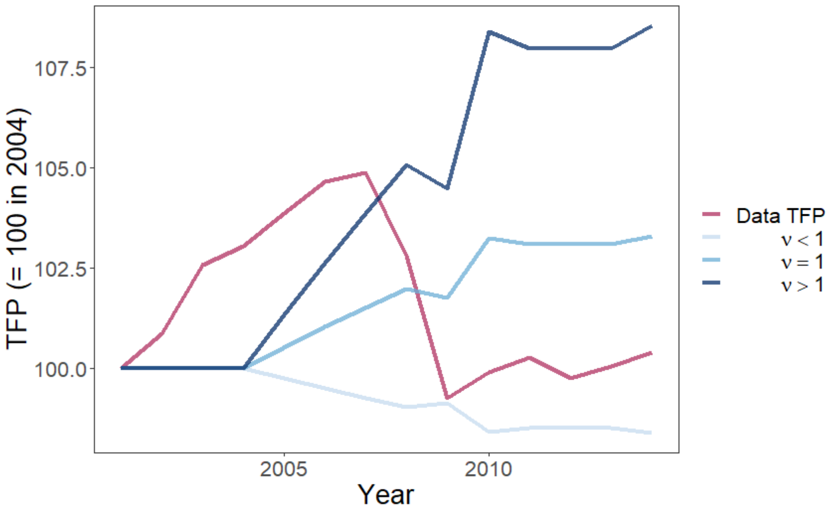
5.2 Rising Overhead Costs

The rise in both ν and μ in the UK explains aggregate productivity growth well. However, our empirical evidence shows that payments to administration costs as a share of sales has increased for the median firm. We consider this data series as a proxy for $w_t\phi/Y_t$ in the model.[36]

In Figure 12 we calibrate ϕ to match our estimates of this ratio. The results highlight the opposing response of aggregate TFP conditional on the level of ν that we discussed in our theoretical analysis. Therefore the level of returns to scale in variable production is crucial for the implied e ect of changing overhead costs. In our estimates, ν is greater than one, which implies productivity should have risen 10% over the period.

[36] Since changing ϕ has general equilibrium e ects on w_t and Y_t, increasing this ratio does not necessarily mean ϕ increases each period. This is relevant because our theoretical analysis focuses on changing ϕ, not the ratio. However, in practice for our calibration, ϕ and the ratio move together.

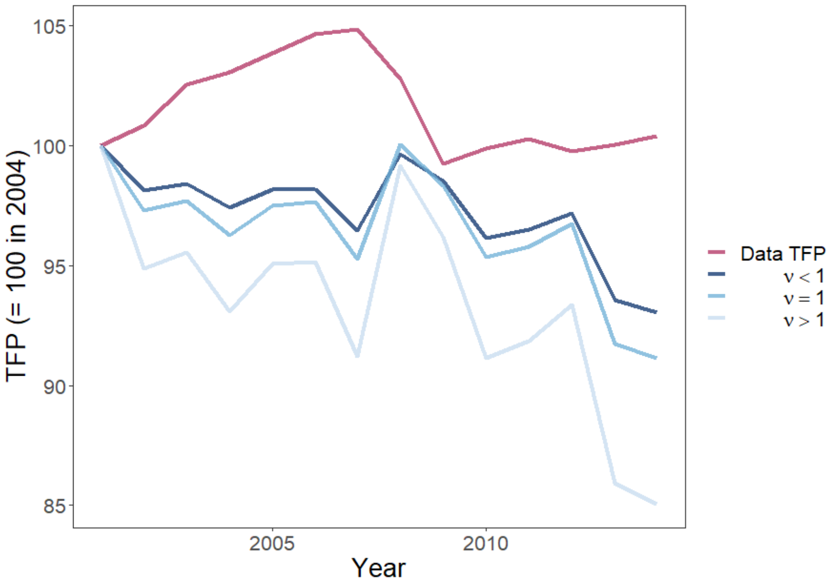
Figure 12: TFP Growth: Model (fixed ν and μ, with variable ϕ) vs Data



We fix μ to its 2001 level and calibrate ϕ to match the overhead share in BvD data. We solve the model steady-state in each year to obtain the model-implied TFP. The TFP data series is from the Penn World Table 10.01 (Feenstra, Inklaar, and Timmer 2015), accessed from FRED: Total Factor Productivity at Constant National Prices for United Kingdom (RTFPNAGBA632NRUG).

In Figure 13, we also re-calibrate μ each year to match CMA (2022) estimates. In this case, aggregate TFP growth underperforms TFP growth in the data, regardless of returns to scale in variable production. Therefore, the markup e ect dominates the fixed cost e ect and we do not observe opposing dynamics for productivity conditional on $\nu \gtreqless 1$.

Figure 13: TFP Growth: Model (fixed ν, with variable μ and ϕ) vs Data



We give the model estimates of μ and calibrate ϕ to match the overhead share in BvD data. We solve the model steady state in each year to obtain the model-implied TFP. The TFP data series is from the Penn World Table 10.01 (Feenstra, Inklaar, and Timmer 2015), accessed from FRED: Total Factor Productivity at Constant National Prices for United Kingdom (RTFPNAGBA632NRUG).

5.3 Rising Returns to Scale Versus Rising Fixed Costs

Overall, the model with changing ν and μ replicates the TFP data better than the model with changing ϕ and μ. And, in both cases treating the markup μ as fixed implies the UK economy should have experienced large productivity increases. Our theory explains that this is because both e ects enhance selection, and reduce the number of active firms, and a reduced number of firms has a further positive e ect on aggregate productivity if there are increasing returns to scale at the firm-level because aggregate resources are concentrated on fewer firms, and those firms benefit from the returns to scale. Higher returns to scale alongside higher markups appear to be a better candidate than higher fixed costs alongside higher markups because given the rise in markups always leads to a strong negative e ect on productivity, the higher returns to scale have a stronger positive e ect on aggregate productivity than higher fixed costs. The stronger positive e ect on productivity is necessary to buoy aggregate productivity to an empirically plausible level. Fixed costs alone cannot raise productivity sufficiently to offset the negative markup effect.

6 Conclusion

This paper investigates the relationship between firm-level scale economies and aggregate productivity. We find evidence that returns to scale and fixed costs, which both determine scale economies, have increased since 1998 in the UK. To explore this relationship, we develop a theoretical framework linking firm-level fixed costs and returns to scale to aggregate productivity. We demonstrate that scale economies can stem from either fixed costs or returns to scale in variable inputs, each with distinct implications for aggregate productivity due to their effects on firm selection and resource allocation.

Our model simulations reveal that while changing returns to scale offer a more plausible explanation for rising scale economies than rising fixed costs, these changes should have significantly boosted aggregate productivity. This predicted growth is not reflected in the data. We resolve this discrepancy by highlighting evidence of increased markups within the UK over the same period, a factor that counteracts the productivity-enhancing effects of scale economies.

In conclusion, our findings suggest that the combined effects of higher scale economies and increased market power help explain the stagnation of productivity growth seen during this period.

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 (2023). "A theory of falling growth and rising rents". In: *Review of Economic Studies* 90.6, pp. 2675–2702.

Asturias, Jose, Sewon Hur, Timothy J. Kehoe, and Kim J. Ruhl (2022). "Firm Entry and Exit and Aggregate Growth". In: *American Economic Journal: Macroeconomics*.

Atkeson, Andrew and Patrick J Kehoe (2005). "Modeling and measuring organization capital". In: *Journal of political Economy* 113.5, pp. 1026–1053.

Baqaee, David and Emmanuel Farhi (May 2020). *Entry vs. Rents: Aggregation with Economies of Scale*. Working Paper 27140. National Bureau of Economic Research.

Baqaee, David, Emmanuel Farhi, and Kunal Sangani (June 2023). "The Darwinian Returns to Scale". In: *The Review of Economic Studies*.

Barnett, Alina, Sandra Batten, Adrian Chiu, Jeremy Franklin, and Maria Sebastia-Barriel (2014). "The UK productivity puzzle". In: *Bank of England Quarterly Bulletin*, Q2.

Barseghyan, Levon and Riccardo DiCecio (Oct. 2011). "Entry costs, industry structure, and cross-country income and TFP di erences". In: *Journal of Economic Theory* 146.5, pp. 1828–1851.

— (2016). "Externalities, endogenous productivity, and poverty traps". In: *European Economic Review* 85.C, pp. 112–126.

Bartelsman, Eric, John Haltiwanger, and Stefano Scarpetta (2013). "Cross-country differences in productivity: The role of allocation and selection". In: *American economic review* 103.1, pp. 305–334.

Basu, Susanto (2008). "Returns to Scale Measurement". In: *The New Palgrave Dictionary of Economics: Volume 1 – 8*. Ed. by Steven N. Durlauf and Lawrence E. Blume. London: Palgrave Macmillan UK, pp. 5559–5562.

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.

Bilbiie, Florin O. and Marc J Melitz (Dec. 2020). *Aggregate-Demand Amplification of Supply Disruptions: The Entry-Exit Multiplier*. Working Paper 28258. National Bureau of Economic Research.

Bloom, Nicholas, Luis Garicano, Ra aella Sadun, and John Van Reenen (2014). “The distinct e ects of information technology and communication technology on firm organization”. In: *Management Science* 60.12, pp. 2859–2885.

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

Chiavari, Andrea (Oct. 2022). *Customer Accumulation, Returns to Scale, and Secular Trends*. Tech. rep. Working paper (Oct 2022).

Church, Je rey R. and Roger Ware (2000). *Industrial Organization: A Strategic Approach*. Irwin McGraw-Hill.

CMA (Apr. 2022). *The State of UK Competition*. Tech. rep.

Colciago, Andrea and Riccardo Silvestrini (2022). “Monetary policy, productivity, and market concentration”. In: *European Economic Review* 142, p. 103999.

Conlon, Christopher, Nathan H. Miller, Tsolmon Otgon, and Yi Yao (May 2023). “Rising Markups, Rising Prices?” In: *AEA Papers and Proceedings* 113, pp. 279–83.

Davis, P. and E. Garcés (2009). *Quantitative Techniques for Competition and Antitrust Analysis*. Princeton University Press.

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, Jan Eeckhout, and Gabriel Unger (Jan. 2020). “The Rise of Market Power and the Macroeconomic Implications*”. In: *The Quarterly Journal of Economics* 135.2, pp. 561–644.

De Ridder, Maarten (Jan. 2024). “Market Power and Innovation in the Intangible Economy”. In: *American Economic Review* 114.1, pp. 199–251.

Dhyne, Emmanuel, Ayumu Ken Kikkawa, Toshiaki Komatsu, Magne Mogstad, and Felix Tintelnot (Sept. 2022). *Foreign Demand Shocks to Production Networks: Firm Responses and Worker Impacts*. Working Paper 30447. National Bureau of Economic Research.

Dixit, Avinash K. and Joseph E. Stiglitz (1977). “Monopolistic Competition and Optimum Product Diversity”. English. In: *The American Economic Review* 67.3, pp. 297–308.

Edmond, Chris, Virgiliu Midrigan, and Daniel Yi Xu (May 2021). *How Costly Are Markups?* NBER Working Papers 24800. National Bureau of Economic Research, Inc.

Ethier, Wilfred J (1982). “National and international returns to scale in the modern theory of international trade”. In: *The American Economic Review* 72.3, pp. 389–405.

Feenstra, Robert C., Robert Inklaar, and Marcel P. Timmer (Oct. 2015). “The Next Generation of the Penn World Table”. In: *American Economic Review* 105.10, pp. 3150–82.

Ganapati, Sharat (2021). “The Modern Wholesaler: Global Sourcing, Domestic Distribution, and Scale Economies”. In:

Gandhi, Amit, Salvador Navarro, and David A. Rivers (2020). “On the Identification of Gross Output Production Functions”. In: *Journal of Political Economy* 128.8, pp. 2973–3016.

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)*.

Ghironi, Fabio and Marc J Melitz (2005). “International trade and macroeconomic dynamics with heterogeneous firms”. In: *The Quarterly Journal of Economics* 120.3, pp. 865–915.

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

Goodridge, Peter, Jonathan Haskel, and Gavin Wallis (2016). "Accounting for the UK Productivity Puzzle: A Decomposition and Predictions". In: *Economica*.

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.

Haskel, Jonathan and Stian Westlake (2017). *Capitalism without Capital: The Rise of the Intangible Economy*. Princeton University Press.

Hopenhayn, Hugo (2014). "Firms, misallocation, and aggregate productivity: A review". In: *Annual Review of Economics* 6.1, pp. 735–770.

Hopenhayn, Hugo, Julian Neira, and Rish Singhania (2022). "From Population Growth to Firm Demographics: Implications for Concentration, Entrepreneurship and the Labor Share". In: *Econometrica* 90.4.

Hopenhayn, Hugo and Richard Rogerson (1993). "Job Turnover and Policy Evaluation: A General Equilibrium Analysis". In: *Journal of Political Economy* 101.5, pp. 915–938.

Hwang, Kyung In, Anthony Savagar, and Joel Kariel (2022). *Market Power in the UK*. Working Papers.

Jaimovich, Nir, Stephen J Terry, and Nicolas Vincent (June 2023). *The Empirical Distribution of Firm Dynamics and Its Macro Implications*. Working Paper 31337. National Bureau of Economic Research.

Kim, Daisoon (2021). "Economies of scale and international business cycles". In: *Journal of International Economics* 131, p. 103459.

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

Krugman, Paul (1991). "Increasing returns and economic geography". In: *Journal of political economy* 99.3, pp. 483–499.

Lashkari, Danial, Arthur Bauer, and Jocelyn Boussard (2024). "Information technology and returns to scale". In: *American Economic Review*.

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.

Lucas, Robert E. (1978). "On the Size Distribution of Business Firms". In: *The Bell Journal of Economics* 9.2, pp. 508–523.

Luttmer, Erzo G. J. (Aug. 2007). "Selection, Growth, and the Size Distribution of Firms". In: *The Quarterly Journal of Economics* 122.3, pp. 1103–1144.

Mankiw, N Gregory and Michael D Whinston (1986). "Free entry and social inefficiency". In: *The RAND Journal of Economics*, pp. 48–58.

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.

ONS (2022). *Estimates of markups, market power and business dynamism from the Annual Business Survey, Great Britain: 1997 to 2019*. Tech. rep. Office for National Statistics website.

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

Panzar, John C. (1989). "Technological determinants of firm and industry structure". In: vol. 1. Handbook of Industrial Organization. Elsevier, pp. 3–59.

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.

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

Savagar, Anthony (2021). "Measured productivity with endogenous markups and economic profits". In: *Journal of Economic Dynamics and Control* 133, p. 104232.

Silberberg, E and W C Suen (2000). *The Structure of Economics: A Mathematical Analysis 3rd Edition*. McGraw-Hill.

Syverson, Chad (2019). "Macroeconomics and market power: Context, implications, and open questions". In: *Journal of Economic Perspectives* 33.3, pp. 23–43.

Appendix

A Graphical Illustration of Scale Economies (Cost based)

It is helpful to consider the three types of cost curve scenarios faced by firms in our model.

Figures 14, 15 and 16 show a firm's cost curves for the case where there is a fixed cost and increasing, constant or decreasing marginal costs. The diagrams show average total cost (ATC), average variable cost (AVC), average fixed cost (AFC) and marginal cost (MC) as firm output varies. Specifically, total cost is the sum variable cost and a fixed cost: $TC = VC + FC$, and averages are the components when divided by output y. The demand curve ($p(y)$) and marginal revenue (MR) curve ($\frac{d\,p(y)y}{dy}$) are not shown. We can imagine them as horizontal in the perfectly competitive case and downward sloping with imperfect competition, for example, due to product di erentiation. The first case (Figure 14) allows for a perfectly competitive equilibrium when the demand curve is horizontal and firms produce at minimum average cost. The second and third cases (Figure 15 and 16) require imperfect competition. The demand curve must be downward sloping for $MR = MC$ to occur.

Figure 14 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 scale. To the left-hand side of the MES the firm has economies of scale and to the right-hand side the firm has diseconomies of scale.

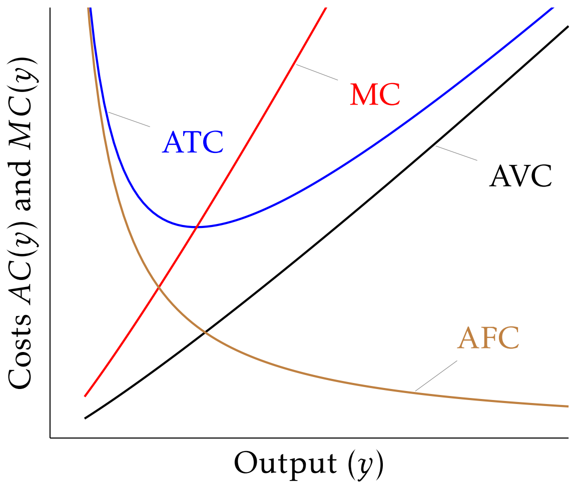


Figure 14: Fixed Cost with Increasing MC, U-Shaped AC Curve

Figure 15 has a constant marginal cost curve and a fixed cost, so there are globally decreasing returns and ATC=MC in the limit. In this case there must be a downward sloping demand curve for an equilibrium where $MR = MC$ to exist. Any degree of slope in the demand curve is sufficient to give an equilibrium, unlike in the next example example which requires a sufficiently steep demand curve (or a sufficiently shallow decreasing marginal cost).

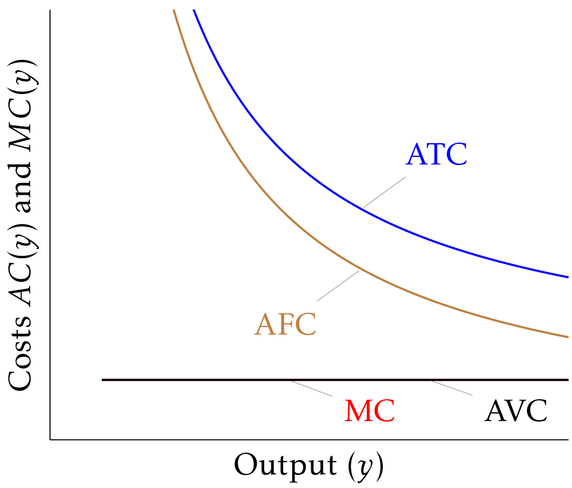


Figure 15: Fixed Cost with Constant MC, Globally Decreasing Returns

Figure 16 has a decreasing marginal cost and a fixed cost so there are global diseconomies of scale. In this case there must be a downward sloping demand curve for an equilibrium where $MR = MC$ to exist. The demand curve must be steeper than the downward-sloping marginal cost curve to ensure this occurs.

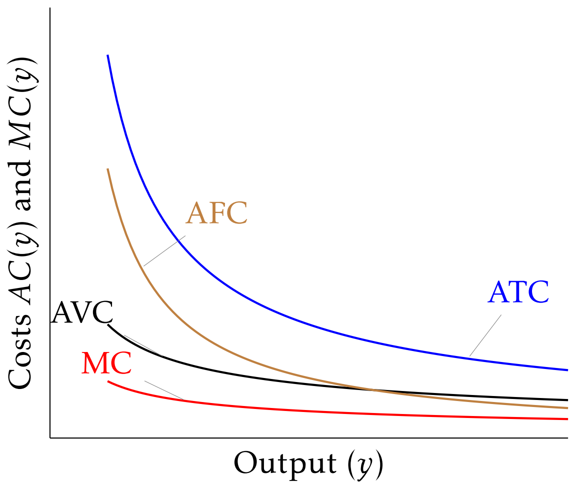


Figure 16: Fixed Cost with Decreasing MC, Globally Decreasing Returns

B Pareto Distributed Productivity

We obtain a measure of productivity $A(\jmath)$ from a random draw on the unit interval $\jmath \in [0,1]$ using inverse transform sampling. The Pareto CDF is given by

$$F(A;\vartheta) = 1 - \left(\frac{h}{A}\right)^{\vartheta}; \quad A \geq h > 0 \quad \text{and} \quad \vartheta > 0.$$

If $\mathcal{J} \sim Uniform(0,1]$, then for $\jmath \in \mathcal{J}$, we have

$$1 - \left(\frac{h}{A}\right)^{\vartheta} = \jmath$$

Therefore, the *quantile function* is

$$A(\jmath) = h(1-\jmath)^{-\frac{1}{\vartheta}}.$$

Typically we set the scale parameter, which is the minimum possible value of A, to $h = 1$. Calibrations of the shape parameter (tail index) vary, for example $\vartheta = 1.15$ in Barseghyan and DiCecio (2011) and $\vartheta = 1.06$ in Luttmer (2007) and $\vartheta = 6.10$ in Asturias, Hur, T. J. Kehoe, and Ruhl (2022). These estimates are set to match the firm size distribution in terms of employment since in these models $A(\jmath)$ is proportional to employment, though, as below, scaling can a ect this.

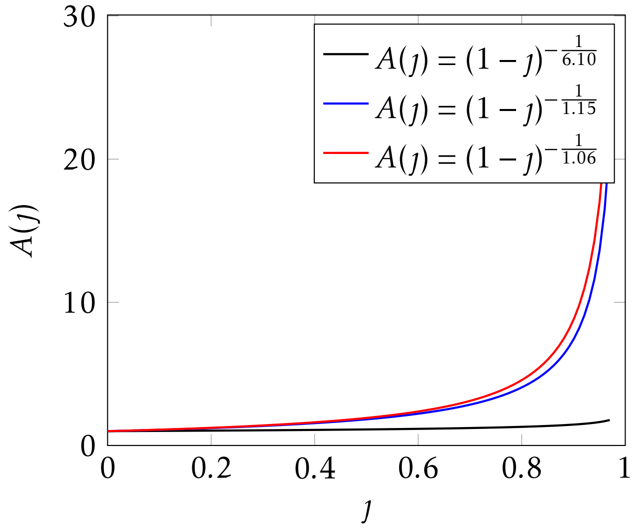


Figure 17: Productivity with Pareto Distribution, $h = 1, \vartheta = \{1.06, 1.15\}$. Domain $\jmath \in (0 : 0.97)$

Figure 18 plots scaled technology $A(\jmath)^{\frac{1}{\mu-\nu}}$ for di erent calibrations of $\nu = \{0.95, 1.00, 1.05\}$ given fixed values of $\mu = 1.1$ and $\vartheta = 50$. The Pareto shape parameter must be large such that $(\mu-\nu)\vartheta > 1$. The distribution of scaled technology is proportional to the distribution of labour, capital and revenue. We require $(\mu-\nu)\vartheta > 1$ so that the expected value of scaled technology is finite, and consequently the expected value of labour per firm, capital per firm and revenue per firm is not infinite.

We observe that a higher ν leads to a greater scaled technology for any given $\jmath$ draw. Since a higher ν decreases the tail index for scaled technology, it causes a lower density of firms to have low-productivity draws and a greater density of firms to have high-productivity draws. Therefore, it thickens the tail of the probability density function. Since employment, capital, and revenue are proportional to this, it also means the distribution of firms is denser towards large firms in terms of labour, capital and employment, and with a lower density of small firms.

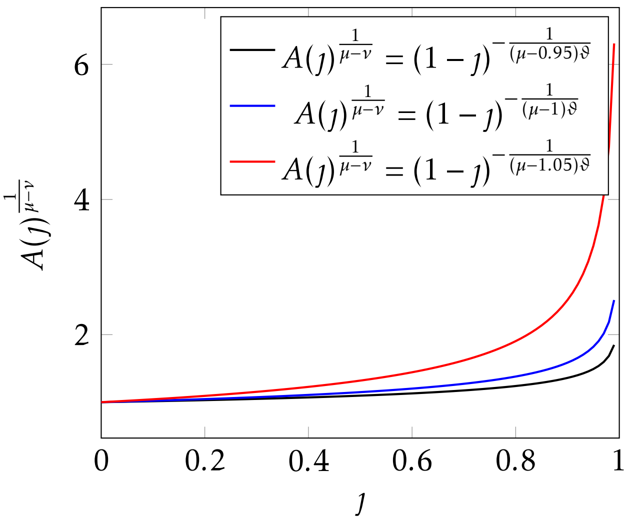


Figure 18: Scaled Technology with Pareto Distribution, $h = 1, \vartheta = 50$ and $\mu = 1.1,\ \nu = \{0.95, 1.00, 1.05\}$. Domain $\jmath \in (0 : 0.95)$.

C Additional Model Derivations

C.1 Profit Maximization Problem

First-Order Conditions

We drop time subscripts t and firm-specific notation $\jmath$. Fixed parameters are $\{\nu, \mu, \alpha, \phi\}$ and endogenous variables are $\{N, Y, A, k, \ell, r, w\}$. The revenue function is

$$py = N^{\frac{1-\mu}{\mu}} Y^{\frac{\mu-1}{\mu}} y^{\frac{1}{\mu}} = N^{\frac{1-\mu}{\mu}} Y^{\frac{\mu-1}{\mu}} A^{\frac{1}{\mu}} k^{\frac{\alpha\nu}{\mu}} \ell^{\frac{(1-\alpha)\nu}{\mu}}.$$

The variables $\{N, Y, A, w, r\}$ are taken as given by the firm. The firm maximizes revenue less costs:

$$\max_{k,\ell} \quad p(k,\ell)y(k,\ell) - rk - w(\ell + \phi).$$

The first-order conditions of the maximization problem state that the marginal revenue product of labour (MRPL) and marginal revenue product of capital (MRPK) – *i.e.* the revenue derivatives with respect to labour and capital – equal to wage and rental rate at optimal choices:

$$MRPL = \frac{\nu(1-\alpha)}{\mu}\frac{p(k^*,\ell^*)y(k^*,\ell^*)}{\ell^*} = w$$
$$MRPK = \frac{\nu\alpha}{\mu}\frac{p(k^*,\ell^*)y(k^*,\ell^*)}{k^*} = r.$$

Since $0 < \alpha < 1$, $\mu \geq 1$, $\nu > 0$ the marginal revenue products are positive. Asterisk notation denotes the profit-maximizing levels of capital and labour.

Second-Order Conditions

The second-order conditions for maximization require that, at the optimal point $\{k^*,\ell^*\}$, the objective function is decreasing in capital and labour and the determinant of the Hessian of the objective function is positive. This implies that $MRPL_\ell < 0$ and $MRPK_k < 0$ where subscripts denote derivatives. And, $MRPL_\ell MRPK_k - MRPL_k^2 > 0$. First note:

$$MRPL_k = MRPK_\ell = \frac{\nu\alpha}{\mu}\frac{MRPL}{k^*} = \frac{\nu(1-\alpha)}{\mu}\frac{MRPK}{\ell^*}.$$

Therefore the following conditions must be satisfied:

$$MRPL_\ell = \left(\frac{\nu(1-\alpha)}{\mu} - 1\right)\left(\frac{MRPL}{\ell^*}\right) < 0$$
$$MRPK_k = \left(\frac{\nu\alpha}{\mu} - 1\right)\left(\frac{MRPK}{k^*}\right) < 0$$
$$MRPL_\ell MRPK_k - MRPL_k^2 = \frac{MRPL \times MRPK}{k^*\ell^*}\left(1 - \frac{\nu}{\mu}\right) > 0$$

These conditions hold if $\nu < \mu$.

C.2 Reduced-form Aggregate Output

We can show that aggregate output reduces to a Cobb-Douglas function of capital and labour scaled by a power mean measure of technology.

$$Y_t = N_t\left[\frac{1}{N_t}\int_0^{N_t} y_t(\imath)^{\frac{1}{\mu}}d\imath\right]^\mu = N_t\left[\frac{E_t}{N_t}\int_{J_t}^1 y_t(\jmath)^{\frac{1}{\mu}}d\jmath\right]^\mu = N_t\left[\frac{1}{1-J_t}\int_{J_t}^1 y_t(\jmath)^{\frac{1}{\mu}}d\jmath\right]^\mu \tag{59}$$

Next, we use the technique of expressing firm-level variables relative to the threshold firm variable, which in turn can be summarised by relative productivity. Here, we re-write as the ratio of firm output $y_t(\jmath)$ to threshold firm output $y_t(J_t)$, where threshold firm output is a constant over $\jmath$:

$$Y_t = N_t y_t(J_t)\left[\left(\frac{1}{1-J_t}\int_{J_t}^{1}\left[\frac{y_t(\jmath)}{y_t(J_t)}\right]^{\frac{1}{\mu}} d\jmath\right)\right]^{\mu} \tag{60}$$

Use the result that

$$\left[\frac{y_t(\jmath)}{y_t(J_t)}\right]^{\frac{1}{\mu}} = \frac{p_t(\jmath)y_t(\jmath)}{p_t(J_t)y_t(J_t)} = \left(\frac{A(\jmath)}{\underline{A}_t}\right)^{\frac{1}{\mu-\nu}} \tag{61}$$

Hence

$$Y_t = N_t y_t(J_t)\left[\frac{1}{1-J_t}\int_{J_t}^{1}\left(\frac{A(\jmath)}{\underline{A}_t}\right)^{\frac{1}{\mu-\nu}} d\jmath\right]^{\mu} = N_t y_t(J_t)\left(\frac{\hat{A}_t}{\underline{A}_t}\right)^{\frac{\mu}{\mu-\nu}} \tag{62}$$

This shows that aggregate output depends on the number of active firms, the size of the threshold firm and the ratio of average technology to threshold technology.[37] Substituting in $y_t(J_t) = \underline{A}_t\left[k_t(J_t)^{\alpha}\ell_t(J_t)^{1-\alpha}\right]^{\nu}$ yields:

$$Y_t = N_t \hat{A}_t^{\frac{\mu}{\mu-\nu}}\underline{A}_t^{\frac{-\nu}{\mu-\nu}}\left[\left(k_t(J_t)^{\alpha}\ell_t(J_t)^{1-\alpha}\right]^{\nu} \tag{63}$$

The next step again applies the technique of representing firm-level variables relative to the the threshold-firm. This allows us to replace $k_t(J_t)$ and $\ell_t(J_t)$ in terms of aggregates.

$$K_t = E_t\int_{J_t}^{1} k_t(\jmath)\, d\jmath = \frac{N_t}{1-J_t}\int_{J_t}^{1} k_t(\jmath)\, d\jmath = \frac{N_t k_t(J_t)}{1-J_t}\int_{J_t}^{1}\frac{k_t(\jmath)}{k_t(J_t)}\, d\jmath \tag{64}$$

$$= \frac{N_t k_t(J_t)}{1-J_t}\int_{J_t}^{1}\left(\frac{A_t(\jmath)}{\underline{A}_t}\right)^{\frac{1}{\mu-\nu}} d\jmath = N_t k_t(J_t)\left(\frac{\hat{A}_t}{\underline{A}_t}\right)^{\frac{1}{\mu-\nu}} \tag{65}$$

$$L_t = E_t\int_{J_t}^{1} \ell_t(\jmath) + \phi\, d\jmath = \frac{N_t}{1-J_t}\int_{J_t}^{1} \ell_t(\jmath) + \phi\, d\jmath = \frac{N_t\ell_t(J_t)}{1-J_t}\int_{J_t}^{1}\frac{\ell_t(\jmath)}{\ell_t(J_t)} + \frac{\phi}{\ell_t(J_t)}\, d\jmath \tag{66}$$

$$= \frac{N_t\ell_t(J_t)}{1-J_t}\int_{J_t}^{1}\left(\frac{A_t(\jmath)}{\underline{A}_t}\right)^{\frac{1}{\mu-\nu}} + \frac{\phi}{\ell_t(J_t)}\, d\jmath = N_t\ell_t(J_t)\left(\frac{\hat{A}_t}{\underline{A}_t}\right)^{\frac{1}{\mu-\nu}} + N_t\phi \tag{67}$$

[37] Gao and Kehrig (2021) present an analogous result for the partial equilibrium case with perfect competition ($\mu = 1$.

Therefore we can express threshold firm capital and labour as

$$k(J_t) = \left(\frac{A_t}{\hat{A}_t}\right)^{\frac{1}{\mu-\nu}} \frac{K_t}{N_t} \tag{68}$$

$$\ell(J_t) = \left(\frac{A_t}{\hat{A}_t}\right)^{\frac{1}{\mu-\nu}} \frac{u_t L_t}{N_t}, \quad \text{where } u_t \equiv 1 - \frac{N_t \phi}{L_t}. \tag{69}$$

Finally, substituting these two expressions into our reduced-form expression for output yields:

$$Y_t = N_t^{1-\nu} \hat{A}_t \left[K_t^{\alpha} \left(u_t L_t\right)^{1-\alpha}\right]^{\nu}. \tag{70}$$

C.3 The Number of Firms and Aggregate Output

The number of firms is a crucial determinant of aggregate output. To see this formally, we can present an alternative expression for average technology in terms of active firms N_t:

$$\hat{A}(N_t) = \left[\frac{E_t}{N_t}\int_{J_t}^{1} A(\jmath)^{\frac{1}{\mu-\nu}}\, d\jmath\right]^{\mu-\nu} = \left[\frac{1}{N_t}\int_{0}^{N_t} A(\imath)^{\frac{1}{\mu-\nu}}\, d\imath\right]^{\mu-\nu}.$$

An alternative expression for aggregate output is in terms of N_t:

$$\begin{aligned}
Y_t &= N_t^{1-\nu} \hat{A}_t \left[K_t^{\alpha} \left(L_t - N_t \phi\right)^{1-\alpha}\right]^{\nu} \\
&= \frac{1}{N_t^{\mu-1}} \left[\int_{0}^{N_t} A(\imath)^{\frac{1}{\mu-\nu}}\, d\imath\right]^{\mu-\nu} \left[K_t^{\alpha} \left(L_t - N_t \phi\right)^{1-\alpha}\right]^{\nu} \\
&= \underbrace{\frac{L_t^{\nu-1}}{N_t^{\mu-1}} \left[\int_{0}^{N_t} A(\imath)^{\frac{1}{\mu-\nu}}\, d\imath\right]^{\mu-\nu} \left(1 - \frac{N_t \phi}{L_t}\right)^{(1-\alpha)\nu}}_{\text{TFP}} K_t^{\alpha\nu} L_t^{1-\alpha\nu}.
\end{aligned}$$

This shows that the number of active firms N_t is an important determinant of aggregate output. It enters the expression in three places with ambiguous effects. First, it appears in the premultiplying denominator, raised to the power $\mu - 1$, which has a negative effect since $\mu > 1$. Second, it appears in the limit of the integral which has a positive effect. Third, it appears in the aggregation of fixed labour overhead costs, which has a negative effect. In addition to the effects through N_t, the parameters ν, ϕ, μ have direct effects as they appear directly in the TFP expression.

D Fixed Cost Share Data

We use the administration expenses share in turnover to proxy the fixed cost share for UK firms. Figure 5 shows the median administration expenses share in turnover for UK firms from 2004 to 2023.

Administrative Expenses

In UK company accounts, 'Administrative Expenses' are defined as expenses an organization incurs that are not directly related to a specific function such as manufacturing, production, or sales. These expenses can include things like: rent, utilities, insurance, wages and benefits for administrative staff, depreciation on office furniture and equipment, professional fees (e.g., accounting and legal fees), and travel expenses. They are necessary for the day-to-day operation of a business, but they do not directly contribute to the generation of revenue. Expenses related to the generation of revenue fall under cost of goods sold (COGs). Administration expenses are typically reported on a company's income statement, below the cost of goods sold (COGS) line.

FAME data

We use the Bureau van Dijk FAME dataset, a UK version of Orbis, to obtain firm financial information. The dataset records the annual financial statements of all incorporated companies in the UK. Over the entire period, there are 16,426,460 company entries. We restrict our analysis to companies that have at least one entry in administration expenses for any year between 2004 and 2023. The company does not need to be active today; it could have dissolved. This restriction reduces the number of companies to 680,763. The companies removed in this step have no administration expenses recorded over the sample period. This occurs because smaller companies can submit micro-entity accounts which do not include this information. Medium and large companies submit 'full accounts' which do record this information. Due to download restrictions, we take a random sample of 250,000 companies, and we keep this same sample of firms every year. Since a firm only needs to have an administration expense in one year, there will be many blanks in any given year for any given company, either because it is inactive or because administration expenses were not recorded because it is a micro-entity. In the end, there are approximately 50,000 firms each year that have an entry in both administration expenses and turnover.

E ARD Data

We use the Annual Respondents Database (ARD) or the time-series version known as ARDx. The ARD is based on the Annual Business Survey (ABS). The ABS is an annual survey of firms in the UK economy. It is a core ONS product used in the construction of national accounts. The ARD adds information from other business surveys to the ABS data.[38] Firms are legally obligated to respond to the survey. The survey forms a firm-level panel that covers all large firms and a representative sample of small firms by geography, size and sector. Large firms are surveyed annually, while small firms are surveyed for a fixed number of years. The ARDx Methodology and ABS Methodology provide more detail.

E.1 Capital Construction

The Perpetual Inventory Method (PIM) allows the construction of firm-level capital stocks when such data are unavailable, but investment data is present. The method here follows Martin (2002) and Hwang, Savagar, and Kariel (2022). The PIM is constructed using the following equation:

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

[38] Specifically, the ARD brings together the ABS and the Business Register and Employment Survey (BRES), and prior to 2009 it brought together the two parts of the Annual Business Inquiry (ABI).

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 company, 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. The capital stock is then constructed for all future years with the above equation, with the missing investment data interpolated. The depreciation rate is taken to be 18.195%, which is a weighted average of the ONS depreciation rates for the three di erent capital categories: Building, Vehicles, Other.

E.2 Deflating

We convert firm gross output and value added into real values using the ONS industry deflators. Material inputs are deflated with the ONS producer price inflation data. The capital stock is deflated with the ONS gross fixed capital formation deflator.

E.3 Cleaning

For the purpose of our production function estimation, we exclude sectors: Agriculture, Public Sector, Finance & Insurance, Education, and Health. Standard Industrial Classification (SIC) 2007 codes: A, K, O, P, Q. These sectors were excluded from the survey after 2012. K,O,P were fully excluded and A,Q had various subsectors excluded. 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 divide the number by 1000 to match with pre-2007 codes. To avoid outliers, which may represent recording errors in the surveys, we winsorize firms with the top and bottom 0.1% of factor shares in revenue (M/Y, K/Y, L/Y) in each year. Table 2 contains number of firms at each stage of the data cleaning process, along with the final number of observations for estimation.

Table 2: Data Cleaning: Firms Dropped

	# Firms
All ARD firm-year obs	854,732
Drop if no 2-digit sector	852,424
Drop if < 100 firms in sector	852,331
Drop sectors A,K,O,P,Q	761,348
Take logs of regression variables	539,368
Drop outlier factor shares	527,813

E.4 Summary Statistics

Table 3 presents aggregate descriptive statistics of the variables used in our regression analysis.

Table 3: Descriptive Statistics of Regression Variables for Full Sample

	Mean	SD	p10	p50	p90	No. Obs
Revenue	39,736	675,183	92	1,458	42,797	527,813
Labour	224	2,213	2	20	349	527,813
Capital	7,696	150,007	22	351	7,915	527,813
Materials	29,651	636,176	32	703	26,255	527,813
Materials Share	0.55	-	0.17	0.58	0.87	527,813
Labour Share	0.26	-	0.04	0.23	0.52	527,813
Capital Share	0.27	-	0.06	0.19	0.60	527,813

Table 4 presents descriptive statistics by broad industry group.

Table 4: Descriptive Statistics of Regression Variables by Broad Sector

	Mean	SD	p10	p50	p90	No. Obs
Manufacturing						
Revenue	36,005	235,437	336	4,294	58,896	125,737
Labour	192	576	8	54	431	125,737
Capital	10,362	75,776	148	1,498	16,154	125,737
Materials	24,954	178,528	122	2,400	38,999	125,737
Materials Share	0.57	-	0.30	0.58	0.81	125,737
Labour Share	0.28	-	0.11	0.27	0.47	125,737
Construction						
Revenue	17,812	108,789	111	1,414	48,782	51,784
Labour	103	395	2	11	214	51,784
Capital	2,309	41,523	11	104	2,210	51,784
Materials	12,467	89,027	18	343	16,896	51,784
Materials Share	0.51	-	0.17	0.52	0.81	51,784
Labour Share	0.25	-	0.00	0.24	0.49	51,784
Trade, Wholesale, Transport						
Revenue	62,673	1,102,305	111	1,414	48,782	182,814
Labour	256	3,404	2	14	244	182,814
Capital	7,092	103,075	20	245	5,667	182,814
Materials	52,666	1,044,112	61	929	26,219	182,814
Materials Share	0.69	-	0.37	0.74	0.92	182,814
Labour Share	0.16	-	0.02	0.13	0.35	182,814
Services						
Revenue	25,276	284,335	65	728	28,673	179,028
Labour	249	1,627	2	17	403	179,028
Capital	8,821	228,905	20	218	5,435	179,028
Materials	14,417	209,297	15	242	11,263	179,028
Materials Share	0.41	-	0.09	0.38	0.77	179,028
Labour Share	0.34	-	0.06	0.32	0.68	179,028

F Returns to Scale Estimates

Table 5 presents estimates of average returns to scale in variable inputs for the whole economy and macro sectors, across di erent estimation methods. This follows from pooling all firms and years together and running a single regression with each estimation technique. Estimates of average returns to scale in the UK from 1998 - 2014 are close in magnitude given the methodological di erences and underlying assumptions on firm behaviour. The estimates suggest returns to scale exceed one. Returns to scale is greatest in the UK in Manufacturing, and lowest in Services. The estimates following ACF and GNR show a clear split between returns to scale in Manufacturing and Construction compared to Wholesale/Trade/Transport and Services: the former sectors have higher returns to scale than the latter. This is less clear with OP and LP estimates, although these methods indicate that Manufacturing has greater returns to scale than Services.

Table 5: Elasticity Estimates: Cobb-Douglas production function, 1998 - 2014

	Olley and Pakes (1996)	Levinsohn and Petrin (2003)	Ackerberg, Caves, and Frazer (2015)	Gandhi, Navarro, and Rivers (2020)
Economy-Wide				
β_l	0.497	0.635	0.545	0.329
β_k	0.521	0.501	0.505	0.181
β_m	-	-	-	0.514
RTS	1.018	1.137	1.051	1.024
N	303,069	449,484	527,813	527,813
Manufacturing				
β_l	0.681	0.573	0.789	0.297
β_k	0.571	0.547	0.421	0.148
β_m	-	-	-	0.590
RTS	1.252	1.121	1.143	1.034
N	95,424	123,552	120,712	120,712
Construction				
β_l	0.574	0.473	0.826	0.328
β_k	0.451	0.332	0.388	0.224
β_m	-	-	-	0.493
RTS	1.025	0.805	1.192	1.044
N	22,123	50,172	51,784	51,784
Wholesale/Trade/Transport				
β_l	0.631	0.592	0.669	0.198
β_k	0.396	0.417	0.343	0.130
β_m	-	-	-	0.688
RTS	1.027	1.009	0.926	1.016
N	74,988	129,043	181,985	181,985
Services				
β_l	0.618	0.598	0.681	0.446
β_k	0.402	0.339	0.384	0.215
β_m	-	-	-	0.354
RTS	1.021	0.938	1.067	1.015
N	77,209	146,717	173,332	173,332

Table 6: Returns to Scale by 2-digit SIC, Ackerberg, Caves, and Frazer (2015)

SIC	ACF	N	SIC	ACF	N
10	1.03	12,495	51	1.05	807
11	1.19	1,724	52	1.14	8,103
13	1.19	4,981	53	1.27	489
14	-	3,355	55	1.52	8,549
15	1.02	841	56	1.39	25,219
16	1.24	3,478	58	1.13	6,802
17	1.34	4,184	59	1.15	2,547
18	1.16	7,521	60	1.29	693
19	1.10	506	61	1.06	1,062
20	1.42	5,733	62	-	9,061
21	1.16	986	63	1.07	1,224
22	1.17	7,776	69	0.54	10,295
23	-	5,616	70	1.00	10,274
24	-	4,776	71	-	11,953
25	-	15,597	72	0.89	2,323
26	1.09	7,648	73	0.94	5,168
27	1.12	4,913	74	1.01	4,769
28	1.29	10,899	75	1.32	1,482
29	-	1,633	77	1.01	6,195
30	1.44	1,973	78	1.09	9,842
31	-	4,060	79	1.15	4,136
32	1.35	5,020	80	1.08	1,926
33	1.24	4,997	81	1.14	6,472
41	1.10	12,216	82	0.98	9,624
42	0.92	12,554	90	0.85	3,111
43	1.33	27,014	91	-	1,722
45	1.02	24,639	92	1.13	1,248
46	0.76	68,969	93	1.07	7,853
47	1.01	66,171	94	1.26	6,086
49	1.21	11,501	95	-	1,889
50	0.97	1,306	96	1.06	11,807

Omitted sectors have estimated coefficients on labour, capital that are negative or greater than one.

F.1 Returns to Scale over Time Pooling all Sectors

Table 7 presents these estimates of returns to scale following Ackerberg, Caves, and Frazer (2015). Economy-wide, there is evidence of a rise in scale economies over time. Estimates in the late 1990s suggest returns to scale below unity – implying decreasing returns – but by the 2010s we find returns to scale above one – implying increasing returns. Most relevant for our analysis is the increasing trend, rather than the levels.

Table 7: Changing Returns to Scale, 1998 - 2014.

	1998 - 2001	2002 - 2005	2006 - 2009	2010 - 2014
Economy-Wide				
β_l	0.42	0.61	0.66	0.72
β_k	0.57	0.47	0.39	0.34
RTS	0.99	1.08	1.05	1.06
N	153,874	144,465	108,619	120,855

Ackerberg, Caves, and Frazer (2015) estimation with a value-added Cobb-Douglas production function.

F.2 Returns to Scale over Time by Broad Sector

Returns to scale have increased across broad macroeconomic sectors. Table 8 presents returns to scale in each sub-period, for each macro sector. Average returns to scale is higher in each sector when estimated between 2010 - 2014, compared to 1998 - 2001. The greatest rise in returns to scale is found in Construction and Services, from 0.91 and 1.02 to 1.29 and 1.11 respectively.

Table 8: Changing Returns to Scale Across Broad Macro Sectors, 1998 - 2014.

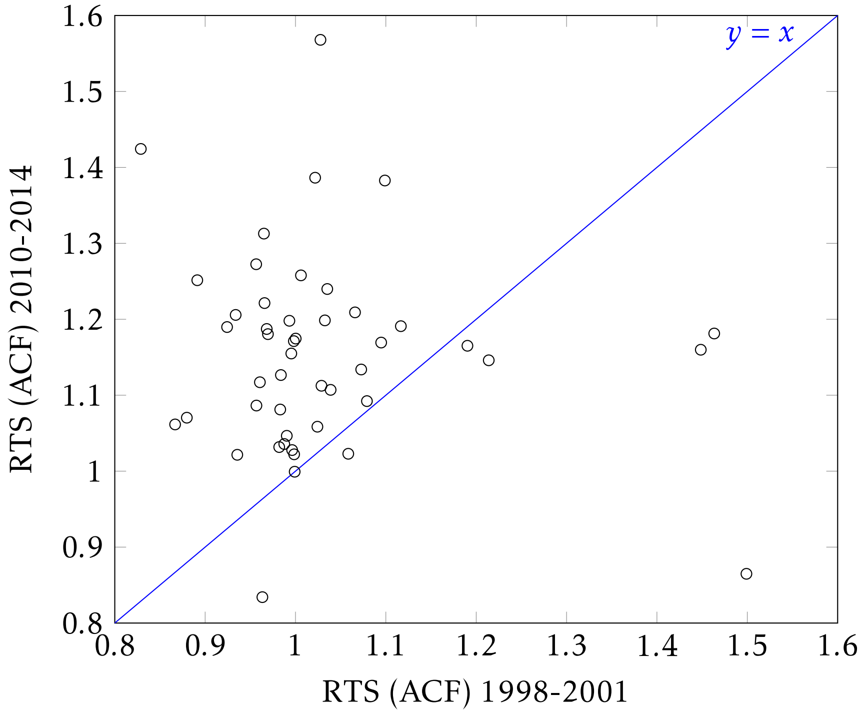
	1998 - 2001	2002 - 2005	2006 - 2009	2010 - 2014
Manufacturing				
β_l	0.58	0.74	0.74	0.78
β_k	0.54	0.50	0.35	0.36
RTS	1.11	1.24	1.10	1.15
N	41,572	36,074	24,280	21,626
Construction				
β_l	0.71	0.63	0.85	0.67
β_k	0.21	0.24	0.23	0.62
RTS	0.91	0.87	1.08	1.29
N	13,050	13,180	9,797	14,145
Wholesale, Trade, Transport				
β_l	0.68	0.61	0.62	0.73
β_k	0.45	0.43	0.38	0.44
RTS	1.13	1.04	1.00	1.17
N	32,792	31,360	27,476	37,415
Services				
β_l	0.58	0.61	0.64	0.81
β_k	0.43	0.42	0.38	0.30
RTS	1.02	1.03	1.02	1.11
N	34,698	34,241	32,070	45,708

Ackerberg, Caves, and Frazer (2015) estimation with a value-added Cobb-Douglas production function.

F.3 Returns to Scale over Time by 2-Digit Industry

The rise in returns to scale over time is more apparent when we estimate at the 2-digit industry level. Figure 19 plots a comparison of returns to scale in 1998 - 2001, compared to 2010 - 2014, across sectors, using the ACF estimation. We remove industries where estimated factor elasticities are below zero or above one. Most industries experienced an increase in returns to scale, as the majority of points sit above the 45 degree line.

Figure 19: Changing Returns to Scale by 2-digit SIC, ACF Estimation.



Comparison of returns to scale at 2-digit SIC level, from 1998 - 2001 to 2010 - 2014. Line is 45 degree line: points above that line are consistent with a rise in returns to scale.

G Reduced Model Under Pareto

Given the equations for u, TFP_t, and w_t that simplify under Pareto, the remaining model equations are:

$$Y_t - C_t = K_{t+1} - (1-\delta)K_t$$

$$\left(\frac{C_{t+1}}{C_t}\right)^{\sigma} = \beta\left[r_{t+1} + (1-\delta)\right]$$

$$Y_t = \text{TFP}_t K_t^{\alpha\nu}$$

$$r_t = \frac{\nu}{\mu}\alpha\frac{Y_t}{K_t}$$

$$w_t = \frac{\nu}{\mu}(1-\alpha)\frac{Y_t}{u}$$

Therefore, we have reduced the model to seven equations in seven variables C_t, K_t, Y_t, $r_t, w_t, TFP_t, \underline{A}_t$, and u is a constant. We can further reduce the equilibrium conditions to two dynamic equations in two variables $\{C_t, K_t\}$. First, if we equate wages and substitute out Y_t, we get $\underline{A}_t$ as a function of K_t:[39]

$$\underline{A}_t = \Psi K_t^{\frac{\alpha\nu}{\vartheta-1}}, \quad \text{where } \Psi \equiv \left(\frac{\phi}{\kappa\left[\vartheta(\mu-\nu)-1\right]}(1-\alpha)\frac{\nu}{\mu}\frac{\Omega\Gamma}{u}\right)^{\frac{1}{\vartheta-1}}. \tag{71}$$

In turn, TFP, wage, rental rate and aggregate output are functions of capital:

$$TFP_t = \Omega\Gamma\Psi K_t^{\frac{\alpha\nu}{\vartheta-1}} \tag{72}$$

$$w_t = \frac{\kappa}{\phi}\left[\vartheta(\mu-\nu)-1\right]\Psi^{\vartheta}K_t^{\frac{\alpha\nu\vartheta}{\vartheta-1}} \tag{73}$$

$$r_t = \alpha\frac{\nu}{\mu}\Omega\Gamma\Psi K_t^{\frac{\alpha\nu\vartheta}{\vartheta-1}-1} \tag{74}$$

$$Y_t = \Omega\Gamma\Psi K_t^{\frac{\alpha\nu\vartheta}{\vartheta-1}}. \tag{75}$$

Substituting the rental rate and aggregate output into the two dynamic equations yields a two-dimensional system, as presented in the paper.

Threshold technology, TFP, wage and aggregate output are increasing in aggregate capital. The rental rate is ambiguously related to capital:

$$\frac{d\ln r_t}{d\ln K_t} = \frac{1-\vartheta(1-\alpha\nu)}{\vartheta-1} \gtreqless 0 \iff 1-\vartheta(1-\alpha\nu) \gtreqless 0.$$

To understand this ambiguity, consider that $r_t = \alpha\frac{\nu}{\mu}Y_t/K_t$. Since $Y_t = TFP_t K_t^{\alpha\nu} =$

[39]Equating wages with Y_t substituted out yields

$$(1-\alpha)\frac{\nu}{\mu}\frac{\Omega\Gamma\underline{A}_t K_t^{\alpha\nu}}{u} = \frac{\kappa}{\phi}\left[\vartheta(\mu-\nu)-1\right]\underline{A}_t^{\varphi}.$$

For a given level of capital, $\underline{A}_t$ adjusts such that the wage markets equate. This relationship gives the intuition for why an increase in capital increases selection. We begin with capital as it is a state variable, determined directly in steady state. From the left-hand wage equation, an increase in capital increases the wage given $\underline{A}_t$ on the left held constant. On the right, which represents wage from the free entry condition, $\underline{A}_t$ must increase – since $\vartheta > 1$, increasing $\underline{A}_t$ on the right-hand side has a stronger wage enhancing e ect than increasing $\underline{A}_t$ on the left-hand side. To summarise, an increase in K, increase w in the factor market equilibrium, therefore $\underline{A}_t$ must increase to raise wage in the free entry condition (i.e. a higher wage means only more productive firms survive). We can think of this relationship as two wage curves $\ln w = \ln\underline{A}$ and $\ln w = \vartheta\ln\underline{A}$, since $\vartheta > 1$ wage is more sensitive to selection in the free entry condition than in the factor market condition.

$\Omega\Gamma\Psi K_t^{\frac{\alpha\nu}{\vartheta-1}} \times K_t^{\alpha\nu}$, therefore $Y_t/K_t = \Omega\Gamma\Psi K_t^{\frac{\alpha\nu}{\vartheta-1}} \times K_t^{\alpha\nu-1}$ where $\alpha\nu - 1 < 0$ by assumption. Heterogeneity ϑ matters through the TFP_t component. If ϑ decreases, Pareto tails become fatter and there is a greater density of high technology draws i.e. more heterogeneity. This strengthens the TFP response to aggregate capital, and consequently aggregate output responds more to aggregate capital, such that aggregate output could increase at an increasing rate in capital. If ϑ increases, Pareto tails become thinner and there is a greater density of low technology draws i.e. no heterogeneity, TFP responds less to capital, and r_t will decrease in K_t.

H Entry Cost Restriction in Steady State

The threshold productivity cannot be lower than the minimum productivity which we have normalized to one ($\underline{A} \geq A_{min} = 1$). Therefore, we require that

$$1 \leq \left[\nu^{\nu} \frac{1}{\mu} \left(\frac{\alpha}{r} \right)^{\alpha\nu} (\phi(1-\alpha))^{\nu(1-\alpha)} \vartheta^{\mu-1} (\mu-\nu)^{\mu-\nu} \frac{1}{\kappa^{1-\alpha\nu}} \frac{1}{(\vartheta(\mu-\nu)-1)^{\mu-\alpha\nu}} \right]^{\frac{1}{\vartheta(1-\alpha\nu)-1}}.$$

Consequently, we can constrain the entry cost parameter such that it satisfies

$$\kappa \leq \left[\nu^{\nu} \frac{1}{\mu} \left(\frac{\alpha}{r} \right)^{\alpha\nu} (\phi(1-\alpha))^{\nu(1-\alpha)} \vartheta^{\mu-1} (\mu-\nu)^{\mu-\nu} \frac{1}{(\vartheta(\mu-\nu)-1)^{\mu-\alpha\nu}} \right]^{\frac{1}{1-\alpha\nu}}.$$

If κ satisfies this with equality, then $\underline{A} = 1$ (and $J = 0$), therefore all entrants are active $N = E$.