Essay in Macroeconomics and Firm Dynamics

Andrea Chiavari

TESI DOCTORAL UPF / year 2022

THESIS SUPERVISOR
Isaac Baley i Edouard Schaal
Department Departament d'Economia i Empresa

To my family

Acknowledgements

I would like to thank my advisors Isaac Baley and Edouard Schaal for their guidance, encouragement, and patience throughout the PhD; without them, this thesis would have not been possible. It has been a privilege to learn from such kind, engaging, and open-minded scholars. I hope to pass on the favor in the future.

I would also thank Andrea Caggese and Jan Eeckhout that without having any formal obligation toward me spent endless time discussing and improving my works.

I was fortunate to learn from the many seminars and meetings with CREi and UPF faculty. I am particularly grateful to Vladimir Asriyan, Davide Debortoli, Julian di Giovanni, Luca Fornaro, Jordi Galí, Manuel Garcia Santana, Priit Jeenas, David Nagy, Giacomo Ponzetto, Victoria Vanasco, Jaume Ventura, and all the participants of the CREi Macro Lunch for their helpful comments and suggestions. Special thanks to Libertad González for her help in preparing the market. Marta Araque and Laura Augusti provided amazing support throughout the whole PhD, particularly over the market and during the writing of this thesis; *moltes gràcies*.

The PhD is a long journey and cannot be done without friends. To Dante Donati, Andrea Fabiani, Sampreet Goraya, and Akhil Ilango: thank you for all the time spent together and the amazing memories. It made the PhD a much better experience. To Gian & Rebe, *semplicemente grazie di tutto, siete stati casa*. A special thanks goes also to Erfan Ghofrani and Moritz Leitner, you have been amazing friends over these years. To Giulia Briselli, Linn Mattisson, Matthias Roesti, Pia Schilling who came to visit UPF during my market and made the experience much lighter: thank you.

I met many more great people. Among them are Laurenz Baertsch, Giacomo Carlini, Madalen Castells Jauregui, Shubhdeep Deb, Luigi Falasconi, Yameng Fan, Jairo Flores, Giorgio Gulino, Andras Jagadits, Ilja Kantorovich, Cristiano Mantovani, Zoel Martinez, Maria Ptashkina, Milan Quentel, Danila Smirnov, Evangelia Spantidaki, Andrea Sy, Sofia Tromlerová.

To Mattia Banti, Matteo Grassi, Paolo Milella, Raffaella Monello, Andrea Pasteris, Vanessa Spadaro, and Ferdinando Vitiello: thank you for your long-lasting friendship and for having always provided a peaceful and recharging retreat in Milano.

This thesis is dedicated to my family, which always made me in the position to pursue my path in life. It would have not been possible without their sacrifices. Last but not least, I thank Marta Morazzoni, with whom I share the best and the worst moments of this journey. I am infinitely thankful for playing such an important role in this period of my life.

Abstract

This thesis documents novel firm-level facts and shows their implication for aggregate phenomena.

The first chapter documents an increase in returns to scale in production. It proposes a novel quantitative firm dynamics model with search frictions in the product market, where firms compete to build their demand, which shows that this transformation in firm-level production processes has sizeable implications for the aggregate economy.

The second chapter shows that firm production processes are becoming more intangible intensive and that this capital is costly to accumulate. Using a quantitative firm dynamics model shows that this shift toward intangible capital can explain an important part of the decline in labor share and allocative efficiency.

The third chapter shows that conditional to a rise in interest rates dominant firms have more countercyclical markups. Using a heterogeneous firms New Keynesian model we show that incomplete pass-through can rationalize this fact and it produces amplification.

Resum

Aquesta tesi documenta fets nous a nivell d'empresa i mostra la seva implicació per als fenòmens agregats.

El primer capítol documenta un augment dels rendiments a escala en la producció. Proposa un nou model de dinàmica quantitativa de l'empresa amb friccions de cerca al mercat de productes, on les empreses competeixen per construir la seva demanda, la qual cosa demostra que aquesta transformació en els processos de producció a nivell d'empresa té implicacions importants per a l'economia agregada.

El segon capítol mostra que els processos de producció de l'empresa són cada cop més intensius intangibles i que aquest capital és costós d'acumular. L'ús d'un model quantitatiu de dinàmica de l'empresa mostra que aquest canvi cap al capital intangible pot explicar una part important de la disminució de la quota de treball i de l'eficiència de l'assignació.

El tercer capítol mostra que, condicionada a un augment dels tipus d'interès, les empreses dominants tenen més marges contracíclics. Utilitzant un model neokeynesià d'empreses heterogènies, demostrem que la transmissió incompleta pot racionalitzar aquest fet i produeix una amplificació.

Preface

This dissertation consists of three essays that investigate the role played by changes in firm-level heterogeneity to aggregate transformations. I employ micro-level datasets in conjunction with cutting-edge econometric techniques to document stylized facts and examine underlying mechanisms using quantitative models of producer heterogeneity and firm dynamics.

This first chapter studies the macroeconomic implications of the rise in firm-level scale economies. My empirical finding is that the average firm-level returns to scale increased within all US sectors, going from 1 to 1.05 between 1980 and 2014. Simultaneously, business dynamism declined, markups rose, and firms devoted increasing resources to customer acquisition, suggesting their active involvement in building and exploiting scales. To jointly account for these facts, I propose a novel theory of firm dynamics grounded in directed search in the product market. Search frictions microfound the customer accumulation process and the presence of heterogeneous markups. The rise in returns to scale explains 62-70% of the decline in business dynamism; 29% of the increase in markups; and 14-45% of the growth in expenditures devoted to customer acquisition. Additionally, the model rationalizes further facts: the aging of US firms, the reallocation of sales toward high markup firms, and firms' declining responsiveness to productivity shocks.

This second chapter, co-authored with Sampreet Goraya, studies the macroeconomic implications of the rise of intangible capital in firm-level production processes. Intangible capital has risen dramatically in the last decades, accounting for more than 30% of aggregate investment by 2015. However, we still know little about its importance in the production process and its associated properties. We estimate the firm-level production function, finding that intangible capital is an important factor for production: its share increased from 0.03 to 0.12 at the expense of labor between 1980 and 2015. We label this phenomenon intangible capital biased technological change (IBTC). Further, we provide novel empirical evidence showing that the investment process of intangible capital is associated

with higher sunk costs, meaning that it entails higher investment adjustment costs relative to tangible capital. Finally, using a model of firms and investment dynamics, we show that IBTC can explain many of the trends witnessed in the US economy since the 1980s. Specifically, it quantitatively explains the rise in the average firm size and concentration, the changes in aggregate factor shares, the increase in the profit rate, the decline in the tangible capital investment rate, and the decrease in allocative efficiency. Our findings suggest that a significant fraction of these transformations can be an outcome of the efficient response of the economy to changes in firm-level production technology.

The third chapter, co-authored with Marta Morazzoni and Danila Smirnov, studies the cyclicality of firm-level markups and their aggregate implications for the business cycle. Firms' markup cyclicality is at the heart of monetary policy transmission in the New Keynesian model. Using US Compustat data and employing local projection techniques, we uncover a novel empirical fact: dominant firms have a more countercyclical markup response after an unexpected contractionary monetary policy shock. Using a heterogeneous firms New Keynesian model with demand accumulation and endogenous markups that evolve over the life-cycle of producers, we show that this is due to the different demand elasticities faced by the firms. Dominant firms face a more inelastic demand, which implies a lower pass-through rate from costs to prices. Therefore, after a contractionary monetary policy shock, dominant firms pass less the reduction in marginal costs to prices compared to competitors, and increase their markups by more, as documented empirically. After calibrating the model to US micro-level data, we find that considering firms' heterogeneous demand elasticities has important implications for monetary policy amplification.

Sumari

Índex de figures

Índex de taules

Capítol 1

THE MACROECONOMICS OF RISING RETURNS TO SCALE: CUSTOMERS ACQUISITION, MARKUPS, AND DYNAMISM

1.1 Introduction

Over the last decades, firm-level production processes have undergone spectacular transformations. The introduction of new technologies, such as information and communication technologies (ICT), and extensive data availability have changed the way firms organize their production. The potential of these technological advancements to expand firms' economies of scale—the cost advantages that firms obtain due to their scale of operation—has captured the attention of academic researchers.[1] Meanwhile, US policymakers' concerns about the effect of these changes on

[1] Bloom et al. (2014) show the link between better information technologies and a wider firm-level span of control.

firms' pricing strategies and competition for customers have gained momentum.[2] This is because, rising scale economies, manifesting through lower costs for the largest firms, may have enabled these same firms to become highly effective in pricing, attracting customers, and exerting market power. Simultaneously, firms have devoted increasing resources to customers acquisition throughout activities such as advertisement and trademarks, suggesting their active involvement in building and exploiting scales.[3]

These technological transformations in firm-level production processes may explain why the US economy has experienced noteworthy trends over the same period of time. In particular, business dynamism—the entry rate of new firms and the reallocation rate of labor across firms—has declined steadily while markups have risen.[4] This has led some observers to speculate that the engine of US productivity may have slowed down, and that its economy may have moved from a competitive to a rent-based one.[5]

However, to date, few studies have systematically analyzed the evolution of firm-level scale economies. What are the consequences of this technological transformation for the above US trends? This project aims to provide an explanation that links these phenomena and makes two contributions. First, I use firm-level data from Compustat to investigate the evolution of firm-level returns to scale in production in the US between 1980 and 2014. Second, I propose a novel theoretical framework to study

[2]Khan (2016), now chair of the Federal Trade Commission, argued extensively about her worries regarding the pricing strategies adopted by Amazon and how this might be the outcome of the firm's scale economies.

[3]Kost et al. (2019) document the rise in trademark activities in the US and how this is associated with market power, whereas De Loecker et al. (2020) show that firms spending more on selling general and administrative are associated with higher markups.

[4]Decker et al. (2014) document the decline in business dynamism, that is, the slowdown in the entry rate of new firms and the reallocation rate of labor across firms. De Loecker et al. (2020) show the rise in markups.

[5]Decker et al. (2016) explain how declining business dynamism may impair the reallocation process across firms, and hence, lower US productivity. Philippon (2019) and Eeckhout (2021) discuss some of the potential reasons why the US economy has become less competitive.

the implications of changes in returns to scale through their impact on customer accumulation.

To study the evolution of returns to scale in the US economy, I estimate the firm-level production function. Here, I follow two state-of-the-art techniques. The first is the control function approach, as in Ackerberg et al. (2015), widely used by the empirical Industrial Organization literature. Second, I use the cost shares approach adopted by Syverson (2004) and Foster et al. (2008). Estimating production technologies in two-digit sectors and over time, as in De Loecker et al. (2020), I find a 5% increase in the average returns to scale, going from 1 in 1980 to 1.05 in 2014. Additionally, this rise shows an acceleration around 1990, consistent with the ICT acceleration ongoing in the same period.[6] Estimating production technologies at the sector level makes it possible to go beyond the analysis to the evolution of the average returns to scale—which could neglect distributional changes across sectors—and to study alternative reasons for this rise, exploiting cross-sectional variation. In particular, there are two potential reasons why the average returns to scale may have risen. First, returns to scale may have increased *within* all sectors. Second, there could have been a reallocation of economic activity *between* sectors toward sectors with ex-ante higher returns to scale. To study these two possibilities, I exploit a statistical decomposition at the sector level, which shows that the rise in the average returns to scale is a within-sector phenomenon.

Although other works have noticed the rise in returns to scale, this paper is the first to highlight the within-sector nature of this phenomenon. This novel fact is consistent with the view that US firms have undergone a technological transformation that has enabled them to increase their scale of operations.[7] I interpret the estimated increase in returns to scale as an

[6]For instance, the World Wide Web entered everyday life in the first period of the 1990s.

[7]Haskel and Westlake (2018) argue in their book that the rise of intangible capital—which is highly related to the digital revolution—has increased the ability of firms to scale their production. Newman (2014), Agrawal et al. (2018), Begenau et al. (2018), Goldfarb and Trefler (2018), Carriere-Swallow and Haksar (2019), and Jones and Tonetti (2020) all emphasize the potential role of data, particularly gathering information from the customer base, for the rise of returns to scale and the presence of increasing returns.

exogenous technological change, seeking to understand its consequences for the US economy and the recent trends mentioned above.

To understand the consequences of this technological change, I propose a novel model of customer accumulation. The framework builds on Gourio and Rudanko (2014) and Roldan-Blanco and Gilbukh (2020) and brings additional tools from the labor-search literature to model customer switching across firms, which in the data is between 10-25% a year.[8] Accounting for customer switching imposes discipline on market power dynamics, as firms internalize the effect of their pricing decisions on their customer base endogenous attrition. To do so, I introduce directed search in the product market, which implies that firms use prices and markups to compete for customers. Further, search frictions imply that firms devote resources to contact new customers. In the model, the presence of fixed operating costs introduces the endogenous entry and exit of firms as standard in most firm dynamics frameworks à la Hopenhayn (1992). Therefore, while remaining tractable for computational analysis, the framework can manage a rich set of firm-level facts and aggregate trends.

The model is grounded in search frictions in the product market. Search frictions microfound firm-level investments in the customer base and firms' strategic use of prices and markups to attract and retain customers, which are an established feature of the firms' activities.[9] Perhaps most importantly, they align the model with the literature pioneered by Foster et al. (2008), which shows that firms mostly grow by accumulating demand. In this vein, recent empirical works by Afrouzi et al. (2020) and Einav et al. (2020) show that customer accumulation accounts for 70% of firms' overall life-cycle growth.

Lashkari et al. (2021) document, using French data, that the adoption of ICT inputs has allowed firms to improve their organization, helping them to improve their scale economies and giving rise to higher returns to scale.

[8] See, for example, the value surveyed by Gourio and Rudanko (2014) from industry estimates.

[9] Dubé et al. (2010) and Bronnenberg et al. (2012) document the prevalence of long-term customer relations. Ruhl and Willis (2008) and Eaton et al. (2009) show that the buildup of market shares is a slow process. Paciello et al. (2019) show that customers are sensitive to prices and that firms consider this while setting them.

I calibrate the model to the 1980s period using identifying moments of the firms' life-cycle, business dynamism statistics from that period, and moments related to firm-level markups. First, as a validation exercise, I show that the model is consistent with a range of cross-sectional and firm-level facts. Second, I demonstrate that the introduction of customer accumulation through search frictions improves the general fit of the model on a series of important but often neglected firms' life-cycle facts. In particular, the model captures the upward sloping life-cycle path of markups and the downward sloping life-cycle path of selling-expenditures, relative to production costs, as observed in the microdata.

In the model, a rise in returns to scale reduces the marginal cost of production and, due to the properties of increasing returns to scale in production, reduces it by more for the biggest firms. This implies that the biggest firms in the economy become very effective in pricing, attracting customers, and charging markups. Therefore, although all firms are subject to the same change, its outcome is highly unequal, as it favors the biggest firms in the economy. This decline in marginal costs has three direct implications: (i) it increases the willingness of firms to scale up, and hence, their expenditures devoted to customer acquisition; (ii) it raises the firm-level markups due to the presence of incomplete pass-through; and (iii) it weakens the selection process in the model, implying a lower entry and reallocation rate. It is noteworthy that the first prediction—that is, the *endogenous* rise in selling-related expenditures relative to production costs after a rise in returns to scale—is a unique feature of this model, where firms invest in their demand through selling-related expenditures.[10] I test and confirm all the predictions in the cross-section of sectors of the Compustat data: I find that higher returns to scale in a sector are positively associated with higher average markups and higher average selling-related expenditures, relative to production costs, and negatively associated with

[10]Models in which market power comes from horizontal differentiation (Dixit and Stiglitz (1977), Kimball (1995), and Atkeson and Burstein (2008)) or search frictions, with only strategic pricing (Paciello et al. (2019) and Roldan-Blanco and Gilbukh (2020)), would not be able to produce the aforementioned facts as, normally, the only non-production costs they feature are fixed costs.

entry and reallocation rates.

I use the calibrated model to study the macroeconomic consequences of the observed rise in returns to scale. This technological change explains 62-70% of the decline in business dynamism; 29% of the increase in markups; and 14-45% of the growth in expenditures devoted to customer acquisition. Additionally, I show that this technological change is consistent with the phenomenon of the aging of firms, as documented in the data by Hopenhayn et al. (2018). It reproduces the reallocation of economic activity toward high markup firms, which gives rise to the fattening of the right tail of the markup distribution, as documented by Autor et al. (2020), De Loecker et al. (2020), and Kehrig and Vincent (2021). It explains the decline in firm-level responsiveness to productivity shocks, which Decker et al. (2020) document as a central component of the decline in business dynamism. Although the rise in returns to scale does not fully account for the markup increase, my investigation suggests that they are an important factor.

Literature Review. This paper contributes to several strands of the literature. It first relates to the search and matching literature on both the labor and the product market. Labor market papers that first introduced some of the techniques used in this paper are Moen (1997), Menzio and Shi (2010), and Menzio and Shi (2011). I build on the methodology developed by Schaal (2017), which, however, focuses on the labor market. Closer to my focus are Gourio and Rudanko (2014), Paciello et al. (2019), and Roldan-Blanco and Gilbukh (2020), which all develop heterogeneous firms models with search frictions in the product market.[11] Relative to Gourio and Rudanko (2014) and Roldan-Blanco and Gilbukh (2020), I allow incumbent customers to search, which is a feature of reality and an important factor for firms' pricing decisions. Moreover, compared to

[11] Burdett and Coles (1997) study the role of firm size for pricing when firms use the price to attract new customers. Dinlersoz and Yorukoglu (2012) provide a theoretical model of industry dynamics in the presence of information frictions. Burdett and Judd (1983), Menzio and Trachter (2015), Burdett and Menzio (2018), and Menzio and Trachter (2018) study equilibrium price dispersion without relying on firm heterogeneity.

Gourio and Rudanko (2014), I allow for commitment on the firm side, which enables firms to charge different prices, even to their incumbent customers. In the absence of commitment, all firms would ask the same price to the incumbent customers, equal to their marginal evaluation, which would make the model quantitatively unsuited to study dispersion in markups coming from different pricing strategies. Differently from Paciello et al. (2019) and Roldan-Blanco and Gilbukh (2020), I allow for increasing returns production technology and firm-level expenditures for customer accumulation, which are all fundamental features for the objective of this paper.

This paper also contributes to the empirical literature that has analyzed technological changes in the firm-level production process. Chiavari and Goraya (2021) show that firms' production technology has become more intangible intensive, at the expense of labor, and that this has had significant implications for the changes in the US factor shares. More closely related to this paper is the work by Lashkari et al. (2021), using French data to show that firms employed ICT investment to increase their firm-level returns to scale; however, they do not analyze its implications for markups. Relative to them, I focus on the US, documenting the within-sector increase in firm-level returns to scale, showing that this has had sizeable consequences for the rise in markups. Despite the focus on the evolution of markups, De Loecker et al. (2020) also document a rise in returns to scale. Yet, they do not focus on sector-level patterns, which I claim are essential in understanding the source of this increase.

Furthermore, this paper complements the growing literature that studies the potential explanations behind the rise in markups and the decline in business dynamism. A strand of this literature emphasizes demographic changes as a relevant factor behind these trends. Papers of this kind are Karahan et al. (2019), Hopenhayn et al. (2018), Peters and Walsh (2019), and Bornstein (2018). Alternatively, Liu et al. (2020) hypothesize that lower interest rates can explain certain recent trends. Relative to this strand of the literature, this project emphasizes technological factors as a potential force driving these trends.

Another strand of the literature, closer to this project, emphasizes the

technological factors behind the rise in markups and the decline in business dynamism. Papers in this vein are Akcigit and Ates (2021), De Ridder (2019), Weiss (2019), and De Loecker et al. (2021).[12] Akcigit and Ates (2021) argue that a decline in productivity spillovers from leaders to laggards is a driver of some recent trends. De Ridder (2019) emphasizes that the rise of firms that are better at using intangibles (as intangibles make other factors more productive) is important for the rise in markups, the decline in business dynamism, and productivity growth. Weiss (2019) shows how intangibles can explain the rise in markups and concentration. De Loecker et al. (2021) document that the rise in fixed costs and the decline in the number of potential entrants can jointly explain the rise in markups and the decline in business dynamism. I contribute to this literature by studying a different technological change—the rise of returns to scale in production—grounded outside the model in a detailed micro-level analysis. Leveraging Industrial Organization techniques to estimate the firm-level production function allows me to infer the strength of the technological change occurring in the US, bringing extra discipline outside the model to the quantitative analysis. The analysis of an alternative technological transformation also provides a new perspective to the ongoing debate regarding the causes of these US trends. Moreover, using a novel quantitative framework, I study additional implications compared to the previous literature. In particular, the model explains the rise in firm-level expenditures devoted to customer accumulation as firms desire to increase their scale of operation to take full advantage of the rise in scale economies.

Outline. Section 1.2 presents the empirical methodology and empirical findings of the paper. Section 1.3 introduces the theoretical model.

[12] Korinek et al. (2018) and Martinez (2018) relate automation to the rise in concentration and to the labor share decline. Crouzet and Eberly (2019) and Zhang (2019b) relate the rise in intangibles with the rise in concentration. Hsieh and Rossi-Hansberg (2019) suggest that the shift toward more productive technologies with higher fixed costs can explain the divergence behind local and aggregate concentration. Aghion et al. (2019) and Olmstead-Rumsey (2019) link the rise in concentration to the decline in productivity growth.

Section 1.4 calibrates the model and evaluates the performance of the model using firm-level and cross-sectional facts. Section 1.5 analyzes and discusses the impact of rising returns to scale before quantifying implications for the aggregate trends objective of this paper. Section 1.6 concludes.

1.2 Empirical Evidence

In this section, I present the empirical analysis of this project: (i) I introduce the main dataset used throughout the analysis; (ii) then, I introduce the main empirical methodology used to estimate firm-level returns to scale; (iii) finally, I document a rise in returns to scale in production within the last three decades.

1.2.1 Data

In this paper, I use two main data sources: Compustat and BDS data. The former is used to obtain information on US firms, while the latter is used to obtain representative measures for the US economy.

Compustat. The main data source is Compustat, a firm-level database with all US publicly traded firms between 1977 to 2014.[13] In this section, I discuss the strengths and limitations of this dataset. I provide more details on the data-cleaning process in Appendix 1.7.1.

The choice of data is driven solely by the ability of these data to cover the period of interest and the largest number of sectors. These characteristics make these data an excellent source of firm-level information to study technological changes in production undertaken by US firms.

Even though publicly traded firms are few relative to the total number of firms (as they tend to be the largest firms in the economy) they account for roughly 30% of US employment (see, Davis et al. (2006)). The

[13] This is also the frame for which the BDS data are available.

Compustat data contain information on firm-level financial statements, including measures of sales, input expenditures, capital stock information, and a detailed industry activity classification.

However, despite its many virtues, these data present two main limitations: (i) the fact that it is impossible to distinguish quantity and prices, which makes measurement of the production function elasticities significantly more challenging as extensively explained in the next section;[14] and (ii) the possible selection issues arising from using only publicly traded firms. To address the first concern, I follow the methodologies explained in Section 2.9.2. Moreover, whenever possible, I compare my results with additional data sources to isolate the potential bias of using only publicly traded firms.

BDS data. To obtain representative aggregate US measures of the firms' size distribution and business dynamism, I use the publicly available dataset from the Business Dynamics Statistics (BDS) program of the Census Bureau.

1.2.2 Production Function Estimation

To estimate firm-level returns to scale, I follow De Loecker et al. (2020) and use two main approaches: (i) the control function approach and (ii) an "augmentedçost shares approach. Both of these approaches are popular methods used to estimate firm-level production functions. I review here the two methodologies, emphasizing their virtues and their limitations.

Control Function Approach

The control function approach was pioneered by Olley and Pakes (1996), and developed further by Levinsohn and Petrin (2003) and Ackerberg et al. (2015). The main insight from this literature is that firm-level unobservable productivity can be proxied by some variable expenditure.

[14] This challenge is present in most of the production data.

To overcome some of the criticism emphasized in Gandhi et al. (2020), I work with a structural value-added specification, as in Ackerberg et al. (2015) and De Loecker and Scott (2016), given by:

$$\mathcal{Q}_{it} = \min\left\{K_{it}^{\beta^k} L_{it}^{\beta^\ell} \exp(\omega_{it} + \varepsilon_{it}),\ \beta^m M_{it}\right\}, \tag{1.1}$$

where $\mathcal{Q}_{it}$ is output, K_{it} is capital, L_{it} is labor, ω_{it} is log-productivity, ε_{it} is the error term, and M_{it} is the materials. This structural value-added production function yields the following first-order condition:

$$\mathcal{Q}_{it} = K_{it}^{\beta^k} L_{it}^{\beta^\ell} \exp(\omega_{it} + \varepsilon_{it}), \tag{1.2}$$

justifying the regression of $\mathcal{Q}_{it}$ on capital and labor while ignoring materials. One caveat is that, in theory, equation 2.22 may not be satisfied in certain situations. If capital and labor are quasi-fixed, and the materials are a flexible input, then when output prices are sufficiently low relative to the price of materials, it will be better to set $M_{it} = 0$ and not produce at all. However, given that my data only include actively producing firms, I assume that equation 2.22 always holds.[15] Therefore, under the specification in equation 2.21, the estimation of the firm-level production function reduces to:

$$q_{it} = \beta^k k_{it} + \beta^\ell \ell_{it} + \omega_{it} + \varepsilon_{it}, \tag{1.3}$$

where $q_{it} = \log(\mathcal{Q}_{it})$, $k_{it} = \log(K_{it})$, and $\ell_{it} = \log(L_{it})$. As usual, the main identification challenge to the production function estimation is the simultaneity bias induced by the unobserved time-varying firm-level productivity, ω_{it}. I follow the control function literature, and in particular Ackerberg et al. (2015) and De Loecker et al. (2020), to estimate the production function in 2.23 using a two-step approach based on the use of a control function for the productivity process. The identification relies on the observation that the firm's labor demand is given by a policy function of the form $\ell_{it} = \ell(k_{it}, \omega_{it})$. Then, providing that the policy function is

[15] For a more detailed discussion on this issue, see Ackerberg et al. (2015).

invertible, the productivity process can be proxied by a control function given by $\omega_{it} = \omega(k_{it}, \ell_{it})$, where $\omega(\cdot) = \ell^{-1}(\cdot)$.[16]

Therefore, in the first stage of this estimation procedure, I clean the firm-level output value from the measurement errors and unanticipated productivity shocks, regressing output on a polinomial of capital, labor, and potential demand shifters, given by:

$$q_{it} = \mathcal{P}(k_{it}, \ell_{it}, \mathbf{d}_{it}) + \varepsilon_{it}. \tag{1.4}$$

Then, in the second stage, using the estimate $\widehat{\mathcal{P}}$ from the previous stage, I can construct a measure of productivity that does not depend on the measurement error ε_{it}, given by:

$$\omega_{it}(\beta^k, \beta^\ell) = \widehat{\mathcal{P}}(k_{it}, \ell_{it}, \mathbf{d}_{it}) - \beta^k k_{it} - \beta^\ell \ell_{it}. \tag{1.5}$$

Finally, taking advantage of the assumption that productivity follows an AR(1) process, it is possible to construct a measure of productivity innovations given by:

$$\xi(\beta^k, \beta^\ell, \rho) = \omega_{it}(\beta^k, \beta^\ell) - \rho\omega_{it-1}(\beta^k, \beta^\ell). \tag{1.6}$$

Therefore, using the productivity innovations, I construct a set of moment conditions to estimated the parameters of the production function, given by:

$$E(\xi(\beta^k, \beta^\ell, \rho) \times \mathbf{z}_{it}) = \mathbf{0}_{Z\times 1}, \tag{1.7}$$

where $Z \geq 3$ and, under the assumption that firms react to unanticipated productivity shocks contemporaneously and that capital is predetermined, the set of admissible instruments is $\mathbf{z}_{it} \in \{k_{it}, \ell_{it-1}, k_{it-1}, \dots\}$. Once the output elasticities are obtained, it is straightforward to recover the returns to scale as:

[16]The assumptions needed to ensure the invertibility of the policy functions associated with a wide class of production functions have been discussed extensively by Pakes (1994), Olley and Pakes (1996), Levinsohn and Petrin (2003), and Ackerberg et al. (2015).

$$\alpha = \beta^k + \beta^\ell. \tag{1.8}$$

Units. It is well known that most of the time, standard production data, such as Compustat, record revenues and expenditures rather than the physical production and input used. In the presence of product differentiation (be it through physical attributes or location), an additional source of endogeneity presents itself through unobserved output and input prices.[17] This implies that, when bringing the model to the data, the structural value-added production function takes the following form:

$$q_{it} + p_{it} = \beta^k(k_{it} + p_t^k) + \beta^\ell(\ell_{it} + p_{it}^\ell) + \omega_{it} + \varepsilon_{it}, \tag{1.9}$$

where p_{it} is the output price, p_t^k is the common user cost of capital, and p_{it}^ℓ is the price of labor. This empirical specification produces the following structural error term:

$$\omega_{it} + p_{it} - \beta^k p_t^k - \beta^\ell p_{it}^\ell. \tag{1.10}$$

I follow De Loecker et al. (2016) and let the wedge between the output and input price (scaled by the output elasticity) be a function of the demand shifters and productivity difference.[18] Including demand shifters $\mathbf{d}_{it}$ in the control function, constructed using the measures of market shares, as in De Loecker et al. (2020), should therefore capture the relevant output and input market forces that generate differences in the output and input price.[19] As discussed in De Loecker et al. (2016), this is an exact control when output prices, conditional on productivity, reflect input price variation, and when the demand is of the (nested) logit form.

[17] See De Loecker et al. (2016) for a recent treatment of these issues.

[18] De Loecker et al. (2020) note that not observing output prices perhaps has the unexpected benefit that output price variation absorbs input price variation, thus eliminating part of the variation in the error term.

[19] I also use industry dummies to capture persistent variation in the demand across sectors.

This is clearly a second-best solution to address the above challenge in estimating the production function; however, it is impossible to go beyond this second-best solution to the problem without more detailed data on the output quantities.

Cost Shares

The cost shares approach has been prominently adopted in Foster et al. (2008), and it exploits the first-order conditions of the firm. To make fruitful use of the firm's first-order conditions, two assumptions are needed: (i) there are constant returns to scale in production and (ii) all inputs are variable. With these assumptions, we can calculate output elasticities from the cost shares. The cost shares of both inputs are defined as:

$$\theta^{\ell} = median\left\{ \frac{w_{it}\ell_{it}}{w_{it}\ell_{it} + r_t k_{it}} \right\} \quad and \quad \theta^{k} = 1 - \theta^{\ell}, \tag{1.11}$$

where $w_{it}\ell_{it}$ is the wage bill, and $r_t k_{it}$ is the rental cost of capital. Therefore, an extra requirement in this method involves the possibility of calculating the return on the physical capital, r_t.

The assumptions required to apply this methodology seem to be incompatible with the objective of this project, that is, the estimation of returns to scale in production. However, I explain how these assumptions have been relaxed by the literature, rendering this methodology flexible for a wide scope of applications.

First, following Foster et al. (2008), one can use moving averages of the cost shares to accommodate for slow adjustments of the inputs due, for example, to adjustment costs. Second, following Syverson (2004), returns to scale can be calculated, even when using a cost shares approach. In particular, he assumes the following functional form for the technology based on cost shares but without constant returns:

$$q_{it} = \alpha\Big[\theta^{k} k_{it} + \theta^{\ell}\ell_{it}\Big] + X_{it}'\delta + \omega_{it} \tag{1.12}$$

with all variables in logs, θ^k and θ^ℓ are given by 2.30, and X_{it} is a vector of potential controls. Therefore, while each cost share determines the output elasticity, the technology does not need to be constant returns, and the curvature is captured by α, which can be estimated with a simple OLS.

1.2.3 The Rise in Returns to Scale

Here, I document the rise in returns to scale under both specifications. Then, I look into the sectoral distribution of returns to scale, finding that this rise is due to an increase across all sectors.

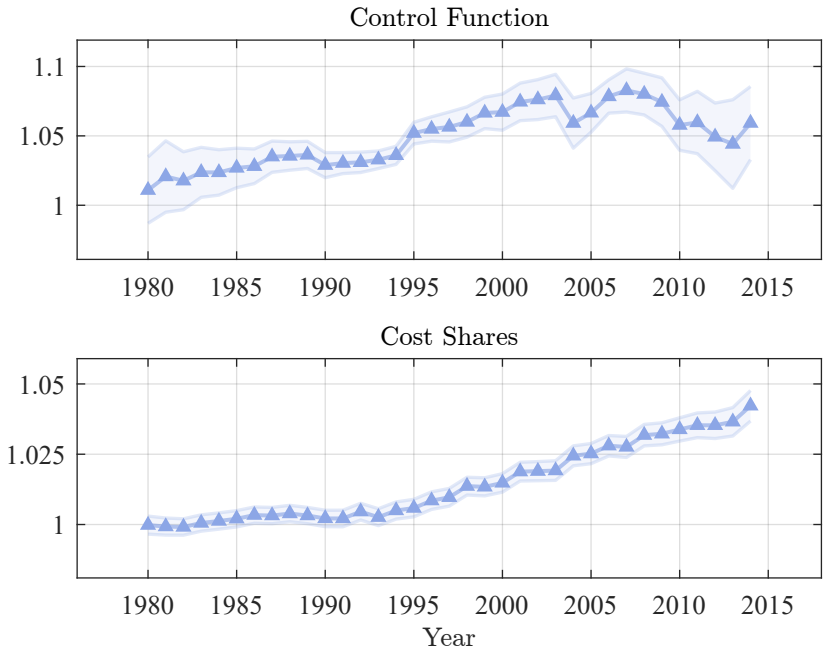
Average Returns to Scale in Production

To estimate the returns to scale for the US economy over a period spanning three decades, I need to assume the particular level at which the production technology is shared across firms. I begin by estimating the returns to scale under the assumption that all firms in the economy share the same production technology. I relax this seemingly unrealistic assumption later on in the analysis. Moreover, to allow for time variation in the elasticities, I estimate equation 2.23 using a ten-year rolling window around the year of interest.[20] Finally, for the choice of variable input in the production, I refer the interested reader to Appendix 1.7.1.

Figure 1.1 shows the evolution of returns to scale for both the control function approach and cost shares approach. The dashed dark blue lines show the point estimates of the returns to scale, whereas the solid light blue lines show the 90% confidence interval. Despite some qualitative differences between the two approaches, the overall quantitative message is similar. In 1980, returns to scale were 1, that is, there were constant returns to scale that rose approximately by 5% by 2014. Therefore, both estimation techniques suggest that, in recent years, US firms' production technology exhibits increasing returns to scale.

[20]Because of data scarcity, I choose a relatively long rolling window. However, the results do not depend on this assumption and are robust to different rolling window schemes.

Figura 1.1: Returns to Scale with Common Technology



Note. The figure on the top shows the evolution of the returns to scale computed with the control function approach. The figure on the bottom shows the evolution of the returns to scale computed with the cost shares approach. The dashed dark blue line shows the point estimates, whereas the solid light blue line shows the 90% confidence interval. Output elasticities are time-varying and calculated from 1980 to 2014.

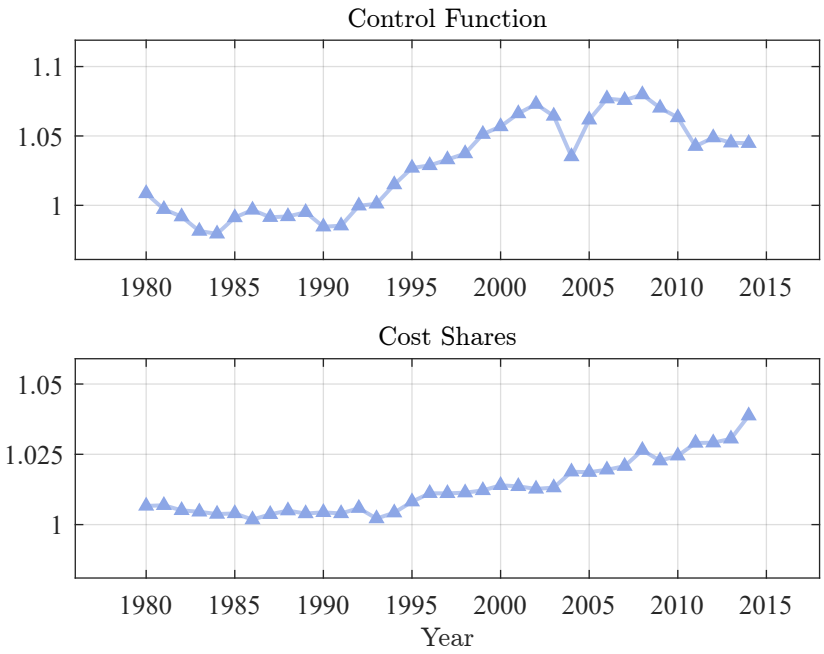
Now I relax the previous assumption of common technology across sectors. To do so, I re-estimate the production technology from equation 2.23 for each two-digit NAICS industry, again using a ten-year rolling window around the year in which I estimate the technology.[21] Therefore, as I estimate a different production technology for each two-digit NAICS industry and year, I define the average returns to scale in the US economy as:

[21]The assumption that firms within a two-digit NAICS industry share the same technology makes the results comparable with those in De Loecker et al. (2020).

$$\alpha_t = \sum_s m_{st} \cdot \alpha_{st}, \tag{1.13}$$

where m_{st} is the weight of each sector, and α_{st} is the sectoral returns to scale. In the main specification, I use sales shares as weights.

Figura 1.2: Returns to Scale with Sector-Level Technology



Note. The figure on the top shows the evolution of the returns to scale computed with the control function approach. The figure on the bottom shows the evolution of the returns to scale computed with the cost shares approach. Output elasticities are time varying and sector specific (two-digit). The average is sales-weighted. The figure illustrates the evolution of the average returns to scale in production from 1980 to 2014.

The graph on the top in Figure 1.2 reports the evolution of the baseline measure—obtained with the control function approach—of average returns to scale across the economy over time. At the beginning of the sample, returns to scale are equal to 1 and remain constant until the end

of the $1980s$; then, they start to rise steeply and by the end of the sample, are around 1.05.[22] In 2014, the average returns to scale is 5% higher compared to the one in 1980.

To validate the robustness of the result from the benchmark measure, the graph on the bottom in Figure 1.2 shows the evolution of the average returns to scale calculated with the cost shares approach. The salient characteristics of this measure closely resemble the patterns of the benchmark measure. From the beginning of the sample to the end of the $1980s$, returns to scale are flat and close to 1; then from the $1990s$ onward, they start to rise, reaching approximately 1.04 in 2014. Therefore, under the cost shares approach, the average returns to scale is roughly 4% higher relative to 1980.

Overall, the rise in returns to scale does not seem to be driven by the specific methodology applied and follows very close patterns across the different specifications. Appendix 1.7.2 reports further robustness exercises using an additional form of capital (such as intangible capital) and an alternative specification of the functional form of the production function (for example, the translog production function). The bottom line is that the finding for the benchmark measure of average returns to scale is robust.

Sectoral Analysis of Rising Returns to Scale

Although the average returns to scale is a useful statistics, it does not fully capture the underlying distributional changes in returns to scale. The advantage of estimating sector-specific production functions is that I obtain a distribution of returns to scale. This allows me to study whether the documented rise in returns to scale is due to a reallocation of economic activity across sectors or whether it is due to a rise in all sectors.

To do so, I decompose the rise in the average returns to scale into the component that is attributable to the rise in returns to scale at the

[22] My estimates are consistent with those reported by Gao and Kehrig (2017) using census data; they find that production technology in the US between 1982 and 1987 had constant returns to scale.

sector level and the component that is attributable to the reallocation of economic activity toward high-returns to scale sectors. Formally, the rise in the average returns to scale can be decomposed as:

$$\Delta\alpha_t = \sum_s m_{st-1}\Delta\alpha_{st} + \sum_s \Delta m_{st}\alpha_{st-1} + \sum_s \Delta m_{st}\Delta\alpha_{st}. \quad (1.14)$$

Therefore, the change in average returns to scale can be exactly decomposed into three components: (i) a *within* component, which captures the portion of the change in the average returns to scale at the industry level; (ii) a *between* component, which captures the portion of the change in the average returns to scale due to the reallocation of economic activity toward high-returns to scale industries; and (iii) finally, a *cross-term* component, which captures the portion of the change in the average returns to scale due to the joint change in returns to scale and in reallocation.

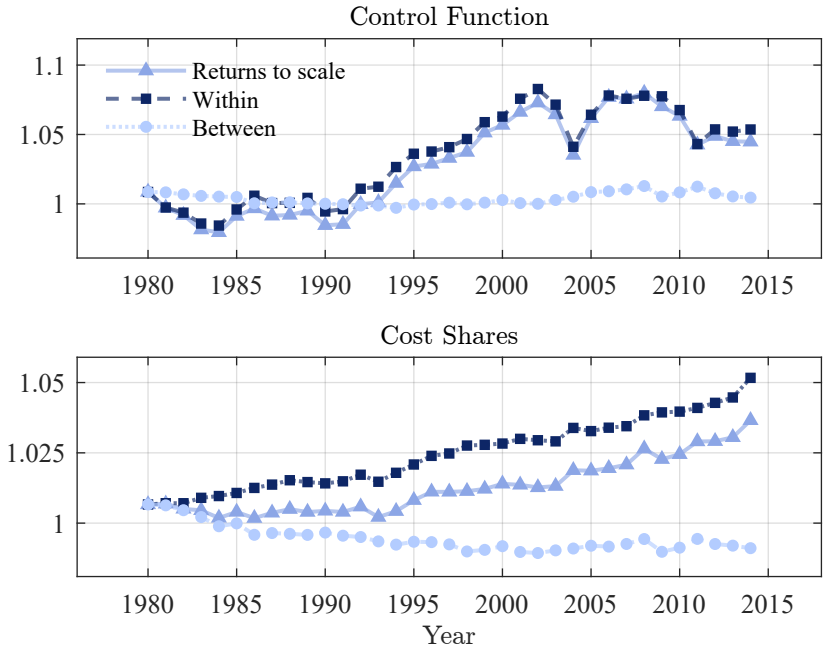
I perform this decomposition across sectors in the entire economy. To best present this decomposition, Figure 1.3 plots the average returns to scale, calculated with both methodologies, as well as two counterfactual experiments, the within and between experiments, based on the decomposition starting in 1980. I do not plot the cross-term experiment, as it is of little economic interest and substantially zero across the entire period. Finally, I set the initial level to 1980 and then cumulatively add the changes of each component from equation 1.14.

The first experiment (dashed dark blue line with squares) shows the counterfactual evolution of the average returns to scale as if there were only the Δwithin component, and all the other components were zero. This experiment shows that the within component tightly follows the average returns to scale in the case of the control function approach and exceeds the average returns to scale in the cost shares approach.[23] The second experiment (dotted light blue line with circles) shows the path of the counterfactual returns to scale if the only change had been due to

[23] With the cost shares approach, the within component exceeds the average returns to scale. Thus, in the absence of reallocation of economic activity across sectors, the rise in returns to scale with this methodology would have been even higher.

Δreallocation. This shows a flat profile over the period for the control function approach and a decreasing profile for the cost shares approach. From these two experiments, it is apparent that the rise in the average returns to scale is indeed a within-sector phenomenon and, if anything, the cross-sectoral reallocation of economic activity has slightly dampened its rise.

Figura 1.3: Decomposition of Returns to Scale Growth at Sector Level



Note. The figure plots the counterfactual evolution implied by the decomposition from equation 1.14 for the control function approach (upper figure) and the cost shares approach (lower figure). The solid blue line with triangles shows the (benchmark) average returns to scale. The dashed dark blue line with squares shows the evolution of the average returns to scale only if the Δwithin component is at play. The dotted light blue line with circles shows the evolution of the average returns to scale only if the Δbetween component is at play.

Taking stock, returns to scale have risen substantially in the US eco-

nomy, and this rise is occurring across all sectors. This transformation in the firms' production technology could stem from many things. For instance, since the $1980s$, and with an acceleration from the beginning of the $1990s$, a digital revolution took place in the US. New technologies such as the internet, mobile phones, computers, and software were developed. These new technologies brought forth an incredible transformation in the way production and business models could be organized. All of a sudden, firms could share internal information at a higher pace and could reach customers at a speed and on a scale previously not possible. The ability of these new technologies to increase the scale at which firms can operate has been the object of interest among researchers since the beginning of the aforementioned digital revolution.[24] I acknowledge that drawing a clear causal link between the digital revolution in information technology and the rise of returns scale requires better data than what I have. However, in this project, I will nonetheless interpret the rise of returns to scale as a pervasive technological transformation that US firms are experiencing across all sectors.

1.3 Model

To study the implications of the technological change outlined above for firms' investment in their customer base, business dynamism, and markups, I build a firm dynamics model with search frictions in the product market. Search frictions are a natural choice to microfound (i) the presence of heterogeneous endogenous markups in equilibrium; (ii) firms' expenditures to attract new customers; and (iii) the empirical observation that firms grow over their life span mostly by accumulating new custo-

[24] A particularly relevant paper is Lashkari et al. (2021), which documents, via rich firm-level data from France that investment in ICT allowed French firms to increase their returns to scale in production in recent years. Newman (2014), Agrawal et al. (2018), Begenau et al. (2018), Goldfarb and Trefler (2018), Carriere-Swallow and Haksar (2019), and Jones and Tonetti (2020) emphasize the potential role of data, particularly gathering information from the customer base, as a source of increasing returns to scale.

mers.[25] I refer the interested reader to Appendix 1.8.1 for a discussion of the technical features of the model.

1.3.1 Population and Technology

Time is discrete. The economy is populated by a representative household, comprising a continuum of measure one of potential buyers and by a large number of workers, and by an endogenous measure of firms with free entry.[26] The representative household discounts the future at a rate β. The instantaneous utility of the household is:

$$uC - v(L), \tag{1.15}$$

where uC is the utility from the consumption of the frictional good, and $v(L)$ is the disutility of labor.[27] The representative household aggregates consumption C is a bundle of the consumption of each active buyer via the following CES aggregator:

$$C = \int_{i \in \mathcal{I}} c_i di, \tag{1.17}$$

where c_i is buyers' consumption of the frictional good, and $\mathcal{I} \subseteq 1$ is the set of active buyers. Equation 1.17 assumes that the goods of the different firms are perfect substitutes, so we can interpret the continuum of firms as effectively selling the same product. Moreover, I assume that buyers wish to buy exactly one unit of the firm's good, and hence, their shopping

[25] Afrouzi et al. (2020) and Einav et al. (2020) show that 70% of firm growth comes from accumulating new customers over their life cycle.

[26] In the text, I refer to buyers and customers interchangeably.

[27] As a consequence, the labor supply of the household will be given by:

$$\lambda^{BC} w = v'(L), \tag{1.16}$$

where λ^{BC} is the Lagrange multiplier associated with the household budget constraint, w is the wage, and $v'(L)$ is the marginal disutility of labor. For convenience, I normalize λ^{BC} to 1 without loss of generality.

value will be equal to the marginal utility of the household's consumption, $u \geq 0$.[28]

Firms differ in their idiosyncratic productivity z, independent across firms, that lies in the finite set $\mathcal{Z}$ and follows a Markov process $\pi(z'|z)$. A firm with a measure ℓ of workers operates with the production technology:

$$y = e^z F(\ell), \tag{1.18}$$

where F is a strictly increasing production function with $F(0) = 0$. Upon entry, firms must pay a sunk entry cost κ. Following Hopenhayn (1992), I assume that firms must pay a fixed operating cost $f \geq 0$ every period to use the production technology. This operating cost is crucial in generating endogenous exit in the model. Finally, I also assume that firms exit exogenously with probability $\delta \in (0, 1)$.

1.3.2 Frictional Product Market

The product market is frictional, and the search is directed on buyers' and firms' sides. Firms announce contracts to attract buyers. Because utility is transferable between buyers and firms, a sufficient statistic for each contract is the utility x that it delivers to the buyer upon matching. Firms offering identical contracts compete in the same market segment; therefore, I describe the product market as a continuum of submarkets indexed by the utility $x \in [\underline{x}, \overline{x}]$ that firms promise to buyers. Firms must pay a cost c for each *ad* they post.[29] Moreover, firms that change their customer base are subject to a convex cost $\mathcal{K}(n_i; n)$, where n_i is the number of new customers that the firm wants to acquire.[30] Buyers can

[28] The fact that buyers wish to purchase exactly one unit implies that only the extensive margin of demand matters in the model, that is, to how many buyers I should sell. This assumption implies that $c_i = 1, \ \forall i \in \mathcal{I}$.

[29] The term *ad* in the model is a stand-in for a broader notion of marketing and selling effort, and will be interpreted as such later on.

[30] The convex cost slows down the adjustment of firms' customer base and is pivotal in generating a realistic endogenous firm life cycle. Moreover, this convex cost is the key friction, together with the exogenous exit shock, preventing the model from settling on a degenerate distribution of firms.

direct their search and choose in which submarket to look.

A standard matching function with constant returns to scale governs match creation in each market segment. I denote by $\theta(x)$ the ads-buyers ratio or tightness of submarket x. In a submarket with tightness θ, buyers find a firm with probability $m(\theta)$, while firms find potential customers with probability $q(\theta) = m(\theta)/\theta$. As standard in the search literature, I assume that m is increasing, while q is decreasing, and that $m(0) = 0$, $q(0) = 1$. Buyers and firms must solve a trade-off between the level of utility of a given contract and the corresponding probability of being matched. The search process takes time, and I assume that firms and buyers can only visit one submarket at a time.

Buyers are allowed to search while already being attached to a firm. The equilibrium market tightness can be written as $\theta(x) = a/\mu$, where a stands for the number of ads posted in submarket x, and μ stands for the corresponding efficiency-weighted number of searching buyers.[31] The number of ads a that a firm posts is not required to be discrete and should be interpreted as a mass. As a result, the law of large numbers applies, and firms do not face uncertainty about the number of buyers they recruit. In particular, a firm that posts a ads exactly meets a measure $aq(\theta) = n_i$ of buyers.

1.3.3 Contractual Environment and Timing

Contracts specify various elements relevant to the firm and its customers. I assume that contracts are state-contingent, and that there is full commitment from the firm side. A contract specifies $\{p_{t+j}, \tau_{t+j}, d_{t+j}\}_{j=0}^{\infty}$, where p is the price, τ is a separation probability, and d is an exit dummy. Each element at time $t + j$ is contingent on the entire history of shocks (z^{t+j}). A more detailed exposition of the contractual environment and its implications for the model is in Appendix 1.8.1.

The contracts offered by firms are large objects but can be written in their recursive form. Contracts are rewritten every period after matc-

[31] In particular, $\mu = \mu_u + \mu_a$, where μ_u is the number of unattached buyers and, μ_a is the corresponding number of attached customers searching on the market.

hing occurs and when production takes place. At this stage, the firm starts with some utility $\mathcal{C}$, promised in the past to its incumbent customers or new ones. A recursive contract $\omega = \{p, \tau, d, \mathcal{C}'\}$ for the current period specifies the current price p and the next period's quantities $\{\tau(z'; w), d(z'; w), \mathcal{C}'(z'; w)\}$, contingent on the next period's state, where $\mathcal{C}'(z'; w)$ is some future promised utility. Because of commitment on the firm side, contract ω is required to deliver at least the promised utility $\mathcal{C}$ to the customers.

The timing of the model is the following. At the beginning of period t, firms decide whether to enter or not. Immediately afterward, incumbent and entering firms learn their idiosyncratic productivity z and their exogenous exit shock δ. Then, conditional on surviving, they decide whether to exit ($d = 1$) or stay. In the following stage, separation occurs with probability τ. Search and matching follow with new and incumbent firms on one side and unattached/attached customers on the other side. Production takes place in the final stage of the period, and the markets clear.

1.3.4 Customer's Problem

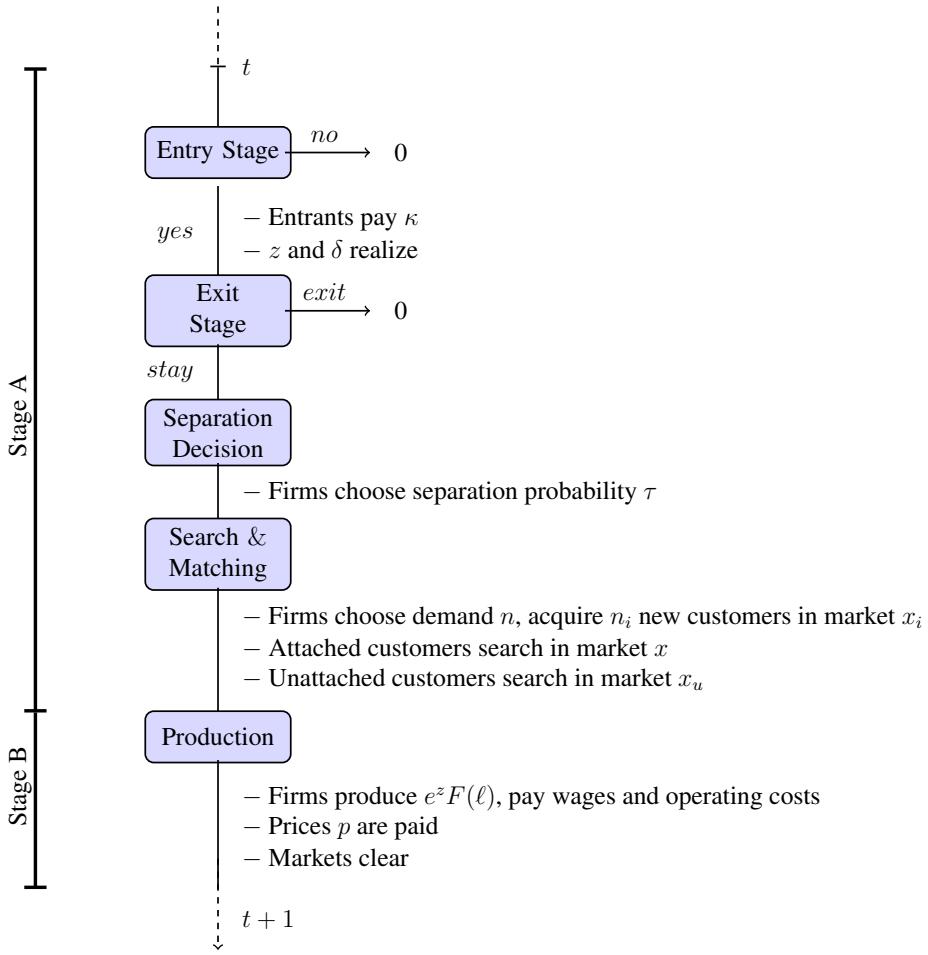
As conventional in the search literature, the value functions below are expressed at stage B of the period when production takes place. I write the value of an unattached buyer as follows:

$$\boldsymbol{\mathcal{U}} = \max_{x_u} \beta[m(\theta(x_u))x_u + (1 - m(\theta(x_u)))\boldsymbol{\mathcal{U}}']. \tag{1.19}$$

If a buyer is not attached to a firm, she does not enjoy any utility in that period. In the following period, she chooses a market segment, x_u, where to search. In doing so, she must solve a trade-off between the offered utility, x_u, and the likelihood of getting a job, $m(\theta(x_u))$. When successful, she enjoys the promised utility x_u, but she remains unattached otherwise.

In the case of a customer attached to a firm with productivity z under the contingent contract $\omega = \{p, \tau(z'; w), d(z'; w), \mathcal{C}'(z'; w)\}$, the value can be written as:

Figura 1.4: Timing of the Model



$$\boldsymbol{\mathcal{C}}(z,\omega;w) = u - p + \beta\mathbb{E}\{(\delta + (1-\delta)d + (1-\delta)(1-d)\tau)\boldsymbol{\mathcal{U}}' + (1-\delta)(1-d)(1-\tau)\max_{x'}[m(\theta(x'))x' + (1-m(\theta(x')))\mathcal{C}'(z';w)]\}. \tag{1.20}$$

An attached customer buys one unit of the firm's output at a price p and values it at the marginal utility of the representative household, $u \geq 0$. The following period may then lead to three different outcomes, which correspond to the three terms in brackets: (i) in the case of exit, that is, exogenously with $\delta \in (0,1)$ or endogenously if $d = 1$, or in the case of destructing the relation, $\tau \in (0,1)$, the customer goes back to the potential buyers' pool with value $\boldsymbol{\mathcal{U}}'$; (ii) she moves to a different firm under a contract with value x' with probability $m(\theta(x'))$; or (iii) she stays in the current firm and receives a promised utility $\mathcal{C}'(z';w)$ in the following period. Notice that customers entering the pool of potential buyers in the given period cannot search in the same period.

1.3.5 Firm's Problem

Consider the problem of a firm at the production stage with a measure n of customers. Customers within the same firm may differ in their level of promised utility. Each customer is identified by an index $j \in [0,n]$ and a corresponding level of promised utility $\mathcal{C}(j)$.

The problem of a firm consists of choosing a list of contracts for its customers:

$$\omega(j) = \{p(j), \tau(z';w,j), d(z';w), \mathcal{C}'(z';w,j)\}, \quad \forall j \in [0,n]. \tag{1.21}$$

In addition, the firm must decide on a submarket $x_i(z';w)$ in which to search for new potential customers, and it must choose the number of new customers that it wants to acquire $n_i(z';w)$. I describe the problem faced by firms as follows:

$$
\begin{aligned}
&\boldsymbol{\mathcal{V}}(z,n,\{\mathcal{C}(j)\}_{j\in[0,n]};w)\\
&\quad=\max_{n_i'(z';w),x_i'(z';w),\{\omega(j)\}_{j\in[0,n]}}\int_0^n p(j)\mathrm{d}j-w\ell-wf\\
&\qquad+(1-\delta)\beta\mathbb{E}\Bigg\{-n_i'\frac{wc}{q(\theta(x_i'))}\\
&\qquad-w\mathcal{K}(n_i';n)+\boldsymbol{\mathcal{V}}(z',n',\{\widehat{\mathcal{C}}(z';w,j')\}_{j'\in[0,n']};w)\Bigg\}^+,
\end{aligned}
\tag{1.22}
$$

subject to:

$$
n'(z';w)=\int_0^n(1-\tau(z';w,j))(1-m(\theta(x'(z';w,j))))\mathrm{d}j+n_i'(z';w),
\tag{1.23}
$$

$$
\widehat{\mathcal{C}}(z';w,j')=\begin{cases}\mathcal{C}(z';w,j) & \text{for} \quad j'\in[0,n'(z';w)-n_i'(z';w)] \quad \text{and} \quad j'=\Phi(z';w,j),\\ x_i(z';w) & \text{for} \quad j'\in[n'(z';w)-n_i'(z';w),n'(z';w)],\end{cases}
\tag{1.24}
$$

$$
y=e^zF(\ell),
\tag{1.25}
$$

$$
y=n,
\tag{1.26}
$$

where $\Phi(z';w,j)=\int_0^j(1-\tau(z';w,k))(1-m(\theta(x'(z';w,k))))\mathrm{d}k$.

In the current period, the firm earns revenue, $\int_0^n p(j)\mathrm{d}j$, minus the cost of labor, $w\ell$, and minus the fixed operating cost, wf. In the following period, the firm survives with probability $(1-\delta)$ and then it chooses whether to exit or not. The $\{.\}^+$ notation, standing for $\max(.,0)$, captures this decision, which I summarize in the dummy $d(z';w)\in\{0,1\}$ ($d=1$ for exit). Following this decision, the firm then chooses a number of new customers to acquire $n_i'(z';w)$ and the submarket $x_i'(z';w)$ in which to direct its selling effort. Because each ad has a probability $q(\theta(x_i'))$ of being successful, the total selling cost incurred for these new customers is $n_i'wc/q(\theta(x_i'))$. Additionally, to slow down the adjustment pace of firms' customer base, I introduce a convex cost, that is, $w\mathcal{K}(n_i';n)$, which each

firm must pay to change its customer base. This is one of the two fundamental assumptions that allows the model to produce a realistic life cycle.[32] Moreover, the constraint that this convex cost imposes on the firm's ability to expand its customer base is the key friction, together with the exogenous exit shock, that prevents the economy from settling on a degenerate distribution of firms.

Constraint (1.23) is the law of motion of total customers. Customers n' in the next period are the sum of the new customers $n'_i(z';w)$ with the remaining customers after the departure of those separated with probability $\tau(z';w,j)$ and of those moving to other jobs with probability $m(\theta(x'(z';w,j)))$. Constraint (1.24) keeps track of the promised utilities across customers. Because the measure of customers evolves over time, I use the mapping Φ to re-index the customers that stay and make sure that a customer with an original index $j \in [0, n'(z';w) - n'_i(z';w)]$ is assigned a new index $\Phi(z';w,j) \in [0, n'(z';w) - n'_i(z';w)]$ in the next period. Newly recruited customers with promised utility, $x'_i(z';w)$, are assigned an index in $[n'(z';w) - n'_i(z';w), n'(z';w)]$. Constraint (1.25) defines the technology with which the firm operates; therefore, this determines the amount of labor ℓ that a firm will hire in each period. Finally, constraint (1.26) states that the output must be equal to the number of available customers n in the given period.

In addition to these constraints, and due to commitment on the firm side, the firm is subject to the following *promise-keeping* constraint:

$$\forall j \in [0,n], \quad \mathcal{C}(j) \leq \boldsymbol{\mathcal{C}}(z, \omega(j); w). \tag{1.27}$$

Constraint (1.27) ensures that the contract $\omega(j)$, assigned to customer j, delivers at least the promised lifetime utility $\mathcal{C}(j)$. Note that there is no non-negativity constraint on the firm's profits, implying that firms have deep pockets and no limited liability.

[32]The second fundamental assumption, as explained later, is related to the fact that each firm enters with a predetermined measure of initial customers. In the quantitative section of the paper, I will calibrate this to be lower than the average mass of customers attached to incumbent firms.

1.3.6 Firm's Pricing

Until now, I have allowed firms to charge different prices to their customers, conditional on their past histories. In this section, I present the optimal prices charged by the firms to their different customers.

Because firms have commitment but customers do not, when a firm designs a contract, it must take into consideration two constraints. First, the contract must take into account a *participation constraint*, given by:

$$m(\theta(x'))x' + (1 - m(\theta(x')))\mathcal{C}(z') \geq \boldsymbol{\mathcal{U}}, \tag{1.28}$$

which states that the continuation value for a customer, conditional on remaining matched, given by equation (1.20), must be higher than the value of being unmatched, given by equation (1.19). This ensures that the customer does not prefer to be unmatched. Second, the contract must take into account the following *incentive constraint*:

$$x' = arg\max_{\tilde{x}} m(\theta(\tilde{x}))\tilde{x} + (1 - m(\theta(\tilde{x})))\mathcal{C}'(z'; w), \tag{1.29}$$

which states that the submarket in which the customer will search is the one that maximizes the continuation value, conditional on remaining matched, given by equation 1.20. This verifies that the submarket in which the customers search is the optimal submarket in which they would like to search. A contract satisfying constraints (1.28) and (1.29) is said to be an incentive-compatible contract. It is now easy to derive prices from the promise-keeping constraint (1.27). The price for a customer j is given by:

$$p(j) = \boldsymbol{\mathcal{C}}(z, \{p = 0, \tau, d, \mathcal{C}'\}; w) - k(j), \tag{1.30}$$

where $k(j) \in \{\mathcal{C}(j), x(j), x_u\}$, depending on the customer's past history.

Notice that the price charged to each customer for the good is the difference between the present value of being attached to a firm evaluated at today's price equal to zero, that is, $\boldsymbol{\mathcal{C}}(z, \{p = 0, \tau, d, \mathcal{C}'\}; w)$, minus the history-dependent promised utility $k(j)$. Therefore, the higher the value

customers get from the match, the higher the price charged by the firm. Conversely, the higher the utility a firm promises, the lower the prices charged to its customers.

Equation (1.30) captures one of the main trade-offs for the firms in the model. In particular, firms are always subject to two opposite tensions. On the one hand, firms that want to grow need to attract customers; to do so, they must give a high promised utility, meaning low prices. On the other hand, firms want to extract value from their matches, meaning that they want to charge high prices to their customers. Therefore, the evolution of prices, and hence of markups, strictly follow the life cycle of firms: young firms, being small, must invest in their customer base, and hence, charge low prices and markups. On the contrary, old firms—which are on average bigger—want to harvest their customer base, and hence, charge high prices and markups.

1.3.7 Free Entry and Equilibrium Definition

To close the model, I am left to specify the process of entry. Every period, before the idiosyncratic shock z is realized, the potential entrants decide whether or not to enter. Upon entry, firms must pay an entry cost κ, after which they draw their idiosyncratic productivity from a distribution g_z. Depending on the outcome, firms may decide to exit or stay, in which case they can start searching for customers and producing as any normal firm.

I define the problem faced by an entering firm of type z as follows:

$$\boldsymbol{\mathcal{V}}^e(z;w) = (1-\delta)\max_{x_e}\left\{-n_e\frac{wc}{q(\theta(x_e))} + \boldsymbol{\mathcal{V}}(z,n_e,\{\mathcal{C}(j)\}_{j\in[0,n_e]};w)\right\}^+. \tag{1.31}$$

Having drawn the idiosyncratic productivity z and surviving the exit shock $\delta \in (0,1)$, the potential entrant first decides whether or not to exit, a decision captured by the notation $\{.\}^+$ and summarized in the dummy $d_e(z;w)$. If it stays, the firm searches $n_e \in \mathbb{R}^+$ new customers, and chooses a submarket, x_e, to maximize its expected value of operating, minus

the total ad cost $n_e wc/q(\theta(x_e))$. I do not allow the entering firms to choose n_e optimally. This is the second necessary ingredient, together with the convex adjustment cost that firms must pay to change their customer base, to obtain a well-defined notion of life cycle within the model.[33]

Due to the presence of free entry, firms enter as long as expected profits exceed the entry cost κ, driving these expected profits down to κ. Therefore, the expected surplus from entering must be equal to κ in equilibrium:

$$w\kappa = \int \boldsymbol{\mathcal{V}}^e(z; w) g_z(dz). \tag{1.32}$$

1.3.8 Firm Distribution Dynamics and Recursive Equilibrium

Using the optimal decision of firms, we may now describe the evolution of customers over time. Let $g(z, n; w)$ be the distribution of customers across firms in stage B of the current period when production takes place. The dynamics of the distribution of customers across firms can be described by:

$$\begin{aligned} g(z', n'; w) = &\sum_{z,n} 1\{n'(z'; w, n) = n'\}(1 - d(z'; w, n))(1 - \delta)\pi(z'|z)g(z, n; w) \\ &+ m_e 1\{n_e(z'; w) = n'\}(1 - d_e(z'; w))(1 - \delta)g_z(z'), \end{aligned} \tag{1.33}$$

where $1\{\cdot\}$ denotes an indicator function. Equation (1.33) defines the mass of firms with an individual state (z', n') in the next period as the sum of surviving incumbent and entering firms that end up in this state. The term m_e is the endogenous measure of new entrants, defined as the number of entering firms required to reach the equilibrium market tightness on every market segment.

[33] I let entering firms enter with an n_e lower than the average size. Together with the convex adjustment cost described earlier, this implies that new firms start small and grow slowly to reach the average size in the economy.

Finally, I define the stationary recursive equilibrium in this economy. A *stationary recursive competitive equilibrium* consists of value functions $\{\boldsymbol{\mathcal{U}}, \boldsymbol{\mathcal{C}}, \boldsymbol{\mathcal{V}}, \boldsymbol{\mathcal{V}}^e\}$, policy functions $\{x_u, x, p, \tau, d, \mathcal{C}', n_i, x_i, d_e, x_e\}$, a wage $\{w\}$, an invariant measure of incumbents g, and a measure of entrant firms m_e, such that: (i) $\boldsymbol{\mathcal{U}}$ and x_u solve the unattached buyers' problem (1.19); (ii) $\boldsymbol{\mathcal{C}}$ and x solve the attached buyers' problem (1.20); (iii) $\boldsymbol{\mathcal{V}}$, τ, d, n_i, and x_i solve the incumbent firms' problem (1.22); (iv) $\boldsymbol{\mathcal{V}}^e$, d_e, and x_e solve the entrant firms' problem (1.31); (v) p and $\mathcal{C}'$ solve (1.29) and (1.30); (vi) the labor market clears; and (vii) the invariant measure of incumbents g and the measure of entrants firms m_e satisfy the dynamics of the distribution of customers across firms, given by (1.33) and the free-entry condition (1.32).

1.4 Model Parametrization and Validation

In this section, I bring the model presented in Section 1.3 to the data. Particularly, the model is estimated to replicate certain salient moments from the cross-section of firms around 1980. First, I present the functional forms and the stochastic processes used in the quantitative analysis. Second, the aforementioned salient moments are used to discipline some deep parameters that are not directly observable to the researcher. Third and finally, I validate the model on non-targeted moments of both the cross-section and the life cycle of firms.

1.4.1 Functional Forms and Stochastic Processes

The household disutility of labor is given by:

$$v(L) = \vartheta \frac{L^{1+\frac{1}{\psi}}}{1+\frac{1}{\psi}}, \tag{1.34}$$

where ϑ is a parameter governing the cost of supplying labor for the household, and ψ is the Frisch elasticity.

The firm-level production function is given by:

$$F(\ell) = \ell^{\alpha}, \tag{1.35}$$

where α governs the firm-level returns to scale of production. Given that time is discrete, I choose a functional form for the probability that a customer finds a firm bounded between 0 and 1, which rules out the Cobb-Douglas matching functions. In particular, I pick the following functional forms:

$$m(\theta) = \theta/(1+\theta)^{-1}, \quad \text{and} \quad q(\theta) = (1+\theta)^{-1}. \tag{1.36}$$

The convex cost of relaxing the customer base is given by:

$$\mathcal{K}(n_i; n) = \chi_1 \left(\frac{n_i}{n}\right)^2 n^{\chi_2}, \tag{1.37}$$

with $\chi_1, \chi_2 \geq 0$.The idiosyncratic productivity shock follows an AR(1) process, given by:

$$z_t = \rho z_{t-1} + \sigma \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(0,1), \tag{1.38}$$

where z_t is the time-varying idiosyncratic productivity, $\rho \in (0,1)$ is the parameter governing the persistence of the process, and σ is the standard deviation of the innovation to the process.

1.4.2 Parametrization

The model is parametrized in two steps. First, I fix a set of parameters to match the standard targets in the steady state. Second, given the values of those parameters, I choose the remaining parameters to match identifying moments from the data. A model period is one year, and the calibration targets moments from the 1980s.

I set the discount rate β equal to 0.97 so that the annual interest rate is about 3%, a value standard in the literature. The degree of firm-level returns to scale α is set equal to 1. This implies constant returns to scale, a value consistent with the empirical estimates presented in Section 1.2.3. I set the persistence of the productivity shock ρ equal to 0.8, the value

found by Foster et al. (2008).[34] The standard deviation of the innovations to the productivity process σ is set to 0.2, a value close to Foster et al. (2008) and common in the firm dynamics literature. The marginal utility from consumption u is set equal to 1, implying a unitary evaluation of each extra unit of consumption. Finally, the Frisch elasticity ψ is set equal to 2.84, corresponding to the average aggregate Frisch elasticity of hours reported by Chetty et al. (2011).

The parameters left to be internally calibrated are $\{c, \chi_1, \chi_2, n_e, f, \kappa, \delta, \vartheta\}$. All these parameters are disciplined through cross-sectional and life-cycle moments. The linear cost c, paid by firms to search for an extra customer, is disciplined by the average markup in 1980. This is identified because this is a sunk cost that firms must recover—in the long run. Hence the higher this cost is, the higher the markup that a firm must charge to operate. The convex cost of increasing the customer base χ_1 is deeply tight with respect to the life cycle of the firms. Particularly, it influences the speed at which firms increase their size. Hence, I use the average size of firms that are five years old in 1980 to identify this value. The initial mass of customers that each entering firm has, n_e, together with the aforementioned convex cost, completely informs the endogenous life cycle in the model. Specifically, given χ_1, the mass of customers upon entry informs us about the size of the entrant firms, which is indeed used as the identifying moment for this parameter. The operating cost f is used to match the average firm size in the period. This is so because if this cost increases, only relatively more productive firms can operate, meaning that the average firm in the market becomes bigger. The entry cost κ is identified with the entry rate in 1980, as it is standard in the literature. The exit shock probability δ is identified with the aggregate excess reallocation rate, as the higher the exit probability is, the higher the reallocation of labor in the model. The convex cost parameter χ_2 is disciplined with the share of firms that are greater or equal to eleven years. This is because the higher χ_2 is, the

[34] Foster et al. (2008) is an important reference, as they disentangle from firm-level sales the contribution of prices from the contribution of true productivity. This is particularly important in our setting, given that the model differentiates firm-level prices and firm-level productivity.

more costly it is to grow for larger firms. Hence, the more likely they will exit at a younger ages. Finally, the labor supply shifter ϑ is set such that the equilibrium wage in 1980 is equal to one.

The parameters are estimated using the following routine. For arbitrary values of the vector of parameters, $\mathcal{P} = (c, \chi_1, \chi_2, n_e, f, \kappa, \delta, \vartheta)$, the dynamic programming problem is solved, and the policy functions are generated. Using these policy functions, the decision rules are simulated until the distribution of firms over $\{n, z\}$ is converged. I draw from this stationary distribution, simulating the economy for many periods, and construct a panel of firms. I compute the aforementioned moments of interest, which I denote as $\mathcal{M}(\mathcal{P})$, whereas the empirical moments are denoted as $\widehat{\mathcal{M}}$. I estimate the fitted parameters $\widehat{\mathcal{P}}$ using a minimum distance criterion, given by:

$$\mathcal{L}(\mathcal{P}) = \min_{\mathcal{P}} \Big(\widehat{\mathcal{M}} - \mathcal{M}(\mathcal{P})\Big)' \boldsymbol{\mathcal{W}} \Big(\widehat{\mathcal{M}} - \mathcal{M}(\mathcal{P})\Big). \tag{1.39}$$

Following Asker et al. (2014), I set the weighting matrix $\boldsymbol{\mathcal{W}} = \boldsymbol{\mathcal{I}}$ and use a grid search algorithm to find the vector $\widehat{\mathcal{P}}$ that minimizes the objective function (1.39).

Table 1.1 summarizes the parameter values resulting from the calibration, along with the fit of the model. The fit is, overall, quite satisfactory. In the calibration, I focus on the average cost-weighted markup. However, the model-implied average sales-weights markup is 1.28, very close to the 1.25 value from the data. Finally, the model implies a slope of selling-related activities on sales of 0.15, close to the value of 0.49 documented by Afrouzi et al. (2020).[35] The next section validates the calibration in deeper detail.

[35] To obtain the slope of selling-related activities on sales, I follow the recent paper by Afrouzi et al. (2020), and I run the following regression specification:

$$s_j = \beta_1 \int_0^{n_j} p_\kappa \mathrm{d}\kappa + \beta_2 w\ell_j + \varepsilon_j,$$

where, in the model, s, the selling-related expenditure, is computed as $wcn_i/q(\theta) + w\chi_1(n_i/n)^2 n^{\chi_2} + wf$, total sales are $\int_0^n p_\kappa \mathrm{d}\kappa$, and $w\ell$ is the labor cost. Hence, the coefficient of interest is given by β_1.

Taula 1.1: Estimated Parameters and Moments

Fixed	Value	Description			
β	0.97	Annual interest rate			
α	1	Returns to scale			
ρ	0.8	Autocorrelation idiosyncratic productivity			
σ	0.2	Standard deviation idiosyncratic productivity			
u	1	Marginal utility			
ψ	2.84	Frisch elasticity			
Fitted	Value	Description	Moments	Model	Data
c	0.45·1e-3	Linear cost of searching	Avg. markup	1.20	1.17
χ_1	0.46	Convex cost of searching 1	Avg. size age 5	12.32	10.16
χ_2	1.91	Convex cost of searching 2	Share of old firms	0.32	0.32
n_e	6.79	Customers' entrant firms	Avg. entrant size	5.98	5.97
f	0.78	Fixed operating cost	Avg. firm size	20.24	20.25
κ	6.92	Entry cost	Entry rate	0.14	0.13
δ	0.98	Exit shock probability	Reallocation rate	0.29	0.31
ϑ	0.985	Labor supply shifter	Wage	1	–

Note. The table reports the values of the parameters and model-implied moments. All the moments have been calculated from 1977 to 1985. I do this because BDS reports data only from 1977; by 1980, not all moments of interest can be computed accurately. Firms size is measured by the total labor ℓ employed in a given period—which is consistent with the measure reported by BDS. The average markup is calculated with cost weights, as in the data.

1.4.3 Validation

To validate the model, I test the overall calibration against two different dimensions of interest. First, I document the model's performance on the cross-section and the life cycle of firms. Second, I test the cross-sectional and life-cycle implications produced by the model for the markups and for the selling ratio—the ratio of non-production to production costs. The reader interested only in the main results can go directly to Section 1.5. Additional steady-state implications of the model are presented in Appendix 1.8.3.

Cross-Sectional and Life-Cycle Implications

The model is designed to capture some relevant aspects of the cross-sectional differences in the micro-data. Part of this cross-sectional heterogeneity is inherently linked with the life cycle of firms. In particular, firms enter small and, conditional on surviving, slowly expand their size when accumulating new customers. This implies that firms of different cohorts have different sizes, with younger firms exhibiting fewer employees—our measure of size, consistent with the BDS data. Moreover, only a few firms survive and keep operating, making the mass of firms belonging to the old cohorts a decreasing share of the total.

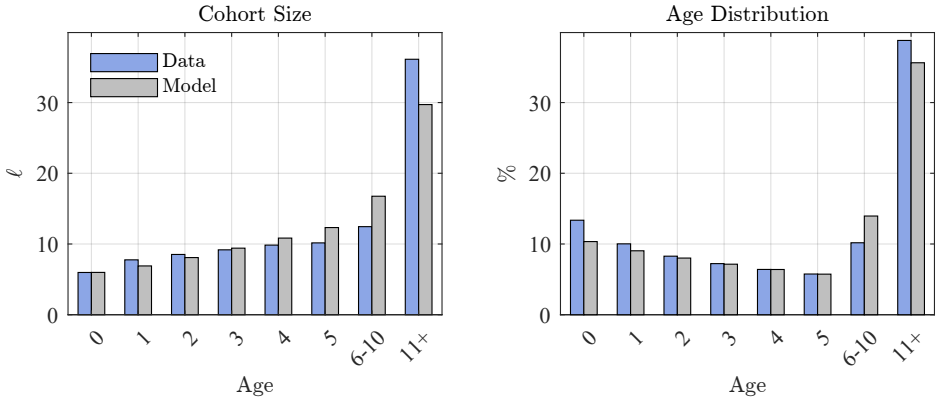
Figure 1.5 presents the aforementioned facts, both for the model and data. The figure on the left shows the size of each cohort, measured as the average number of employees within each firm of a given age, in the model and the data. It can be seen that the model and data track each other well; this is not surprising, given that average employment for firms of age 0 and age 5 used as a target in the calibration. Nonetheless, the model slightly understates the size of the oldest (11+) firms. Instead, the figure on the right documents the distribution of firms across cohorts in the data and the model. The model manages to track satisfactory data.

Empirical works on firm-level data have established many regularities about the life cycle of firms. Since the seminal work by Dunne et al. (1989), we know that in the US manufacturing sector, establishment growth is unconditionally negatively correlated with age.[36] Moreover, Cabral and Mata (2003), using a comprehensive data set of Portuguese manufacturing firms, show that the employment distribution shifts to the right and becomes less right-skewed as cohorts age. Figure 1.6 shows the aforementioned life-cycle facts in the model.

The model aptly captures the life-cycle facts. In the model, firms enter small, with few customers, and grow only slowly, accumulating new customers. Moreover, the accumulation of customers is less costly for young firms; hence, they experience higher growth relative to older firms.

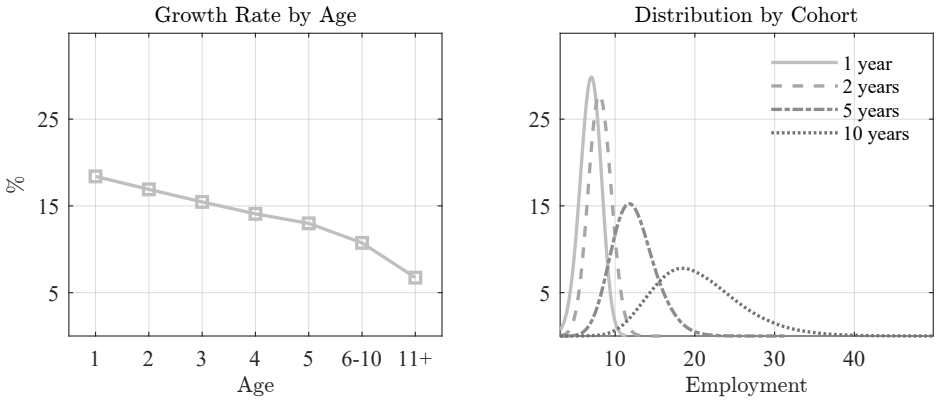
[36]This finding was confirmed for a variety of sectors and countries. See Coad (2009) for a recent survey of the literature.

Figura 1.5: Model Cross-Section



Note. The figure on the left shows the size of each cohort, measured as the number of employees within firms. The figure on the right shows the distribution of firms across cohorts. The light blue bars represent BDS data; the light grey bars show the model predictions.Data reported are between 1977-1985.

Figura 1.6: Model Life Cycle



Note. The figure on the left shows the employment growth rate by age, that is, $g_{it}^{\ell} \equiv (\ell_{it} - \ell_{it-1})/\frac{1}{2}(\ell_{it} + \ell_{it-1})$. The figure on the right shows the employment distribution across cohorts. Both y-axes are in percentage points.

The same mechanism explains the results presented in the right figure. In particular, while firms age, they expand their size, pushing the distribution of their cohort to the right. Overall, the model fits well with many non-targeted moments of the cross-section and the life cycle of firms.

Implications for Markups and Selling-Related Activities

The model produces clear predictions about the evolution of markups and selling-related activities over the life cycle of the firms. In particular, young firms charge lower markups and spend more on selling-related activities (relative to production costs) to grow faster. Therefore, in the data, we should expect to observe a growing profile for markups and a declining profile for selling-related activities over production costs as firms age.

To map the model's expenditures to an empirically meaningful empirical counterpart, I define the *selling ratio* in the model as:

$$\varrho = \frac{f + n_i c/q(\theta) + \chi_1 (n_i/n)^2 n^{\chi_2}}{\ell}, \tag{1.40}$$

where the numerator is composed of the total non-production costs (which, through the lens of the model, I interpret as selling-related activities), whereas the denominator is composed of the total production costs.[37]

Moreover, to test the aforementioned predictions of the model in the data, I exploit the following regression specification, given by:

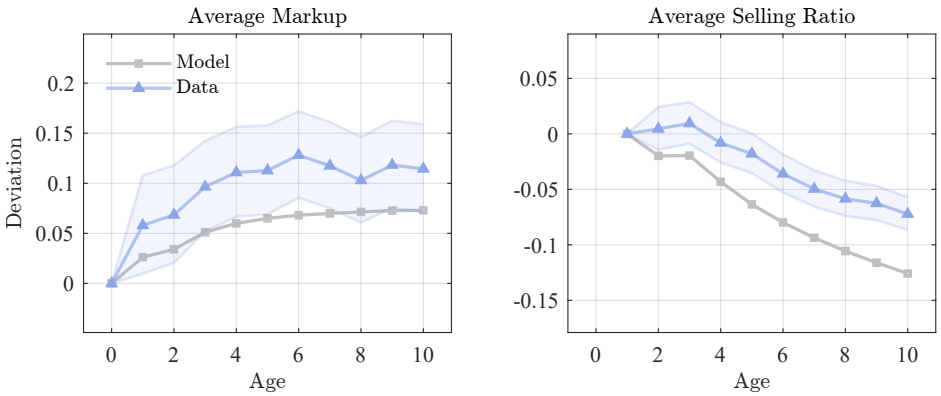
$$\log y_{it} = \alpha + \sum_{a=1}^{10} \gamma_a 1\{\text{age}_{it} = a\} + \phi_{st} + \varepsilon_{it}, \tag{1.41}$$

where $y_{it} \in \{\mu_{it}, \varrho_{it}\}$, the firm-level markup μ_{it} is defined in Appendix 1.7.1, the selling-ratio ϱ_{it} is the ratio of selling-related expenditure to the cost of goods sold, where selling-related expenditure is defined in Appendix 1.7.1, age_{it} is the firm's age, and ϕ_{st} are sector-year fixed effects.

[37] Notice that both the numerator and the denominator should be multiplied by w, the wage, which however is not reported, as it is canceled out.

The coefficients γ_a are the parameters of interest that measure the average $\log y_{it}$ for each age group using within sector-year variation.

Figura 1.7: Life Cycle of Markups and Selling Ratio—Model and Data



Note. The figure on the left shows the average markup across firms of different ages, both in the model (light grey line with squares) and in the data (light blue line with triangles); the figure on the right shows the average selling ratio across firms of different ages, both in the model (light grey line with squares) and in the data (light blue line with triangles). The light blue areas are the 90% confidence interval. All variables are reported relative to the initial year, which is normalized to zero.

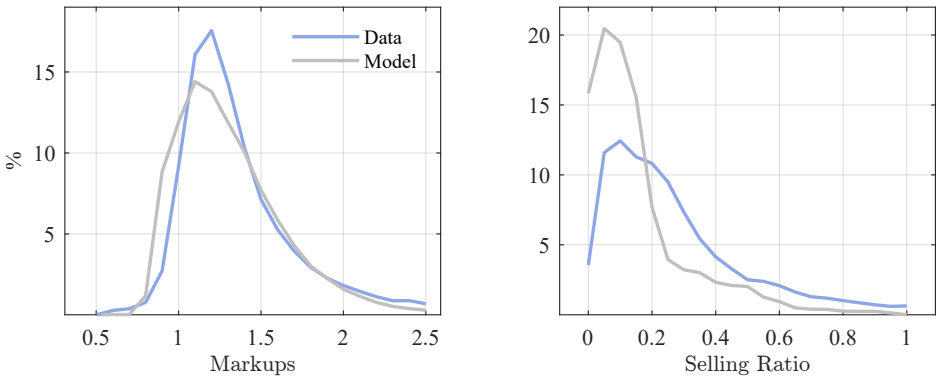
Figure 1.7 shows the evolution of the average markup and average selling ratio for firms of different ages.[38] The model-implied markups over the life cycle satisfactorily follow the one in the data; if anything, the model-implied one has a slightly flatter profile over the life cycle.[39] The model-implied selling ratio declines over the life cycle of the firm, as we also see in the data. However, in this case, the model performs quanti-

[38]I plot the results for the initial part of firms' life cycle; however, the patterns remain similar when the age is more than ten. The selling ratio is plotted only when the age is greater than zero because in the model, entrant firms face a cost composition that is different, as they do not pay the convex cost.

[39]Similar empirical findings have also been documented by Alati (2021) in Compustat and by Peters (2020) in Indonesian data. They both find that markups increase over the firms' life cycle.

tatively less well than in the markups case. In the data, the selling ratio declines less compared to the model.

Figura 1.8: Distributions of Markups and Selling Ratio—Model and Data



Note. The figure on the left shows the markup distribution in the data (light blue) and in the model (light grey). The figure on the right shows the selling ratio distribution in the data (light blue) and in the model (light grey). The distributions in the data are calculated within the period 1977-1985. The distributions of markups are showed within the [0.5, 2.5] range, whereas, the distributions of the selling ratio are shown within the [0, 1] range.

As a final validation exercise, I compare the model-implied distribution of the markups and the selling ratio with their empirical counterparts. Figure 1.8 shows the comparison. The figure on the right shows the model-implied distribution of markups (light grey) and its empirical counterpart (light blue); the figure on the left shows the model-implied distribution of the selling ratio (light grey) and its empirical counterpart (light blue). Overall, the qualitative fit is satisfactory.

The distribution of markups implied by the model is very close to the empirical counterpart. This is a successful outcome of the model, as the only targeted moment of that distribution is its cost-weighted average. Moreover, the model aptly captures the right skewness of the empirical distribution of the selling ratio. However, as none of the moments of this distribution has been used to calibrate the model, there are some quanti-

tative differences: (i) the data show a higher mass near zero; and (ii) the empirical distribution of the selling ratio is less dispersed compared to the one implied by the model. Without further data, is impossible to say where these differences come from; however, in Appendix 1.8.3, I show that by using an alternative measure of selling-related expenditure, the overall qualitative features of the empirical distribution of the selling ratio remain unchanged.

1.5 Rising Returns to Scale and the Macroeconomics

Having calibrated and validated the model, in this section, I move forward to study the macroeconomic implications of a rise in the returns to scale, as documented in Section 1.2.3. To this end, I will analyze, within the model, the effect of rising returns to scale from 1 to 1.05, keeping all the other parameters fixed. First, to shed light on the main mechanism, I discuss the qualitative implications of such a rise in returns to scale in the model. Second, I present suggestive evidence for the mechanism inbuilt in the model. Third, I use the model to study the quantitative implications of this 5% rise in returns to scale, as documented in section 1.2.3.

1.5.1 Inspecting the Mechanism

In this section, I explore the qualitative implications of a rise in returns to scale from 1 to 1.05, keeping all the other parameters fixed to the 1980 calibration. First, I link the effect that rising returns to scale have on the marginal costs of production at the firm level. Second, I explain how changes in the marginal cost of production affect markups and business dynamism.[40]

Given the production structure of the model, as specified in Section 1.3, the firm-level marginal cost of production is given by:

[40] Appendix 1.8.2 extends the intuitions provided in this section to a more general case in which firms produce also using capital.

$$\mathcal{MC}(z, n; w) \equiv \frac{1}{\alpha}\left(\frac{n}{e^z}\right)^{\frac{1-\alpha}{\alpha}} \frac{w}{e^z}, \tag{1.42}$$

where α is the firm-level returns to scale, n is the mass of customers, that is, the firm-level size, e^z is the idiosyncratic productivity, and w is the wage. Notice that, when $\alpha = 1$—in the presence of constant returns to scale—the marginal cost of production reduces to the more familiar w/e^z; hence, it is just the ratio of the wage to the idiosyncratic productivity.

However, when $\alpha > 1$, the marginal cost of production not only depends on the firm's size, but also decreases in it—this is under the quantitative-relevant scenario in which $n/e^z \geq 1$.[41] Therefore, this model, once calibrated to the empirical findings presented in Section 1.2.3, implies that the bigger a firm is, the better it becomes to produce, and hence, the lower its marginal cost of production is. This link can be interpreted as the model microfoundation of a technological change biased toward larger firms. In particular, the negative dependence of firm-level marginal costs of production with size stems from the notions that bigger firms (with bigger economic activities) manage to gather more information about their production processes (and potentially about their customers as well) and use it, owing to new information and communication technologies (ICT), to improve production. This mechanism creates a virtuous circle where bigger firms are better at producing; hence, become even bigger and better at producing, and so on.[42]

[41] In the model, I do not restrict this ratio to be greater than one, but when I calibrate it to match the firm size distribution, as explained in Section 1.4.2, I indeed find that this is the case. In this sense, this should not be seen as an assumption but as a quantitative result.

[42] Lashkari et al. (2021) document (using rich firm-level data from France) that investment in ICT has allowed French firms to increase their returns to scale in production in recent years.

Newman (2014), Agrawal et al. (2018), Begenau et al. (2018), Goldfarb and Trefler (2018), and Carriere-Swallow and Haksar (2019) provide additional microfoundations for the same concept. All of them emphasize the potential role of data, particularly gathering information from the customer base, which can give rise to increasing returns to scale.

Figure 1.9 on the left shows, for firms with a different number of customers n, that is, for firms of different size, how the firm-level marginal cost of production changes as the returns to scale change—the analysis is performed under the already stated quantitative-relevant scenario in which n/e^z is big enough, that is, is weakly greater than 1.

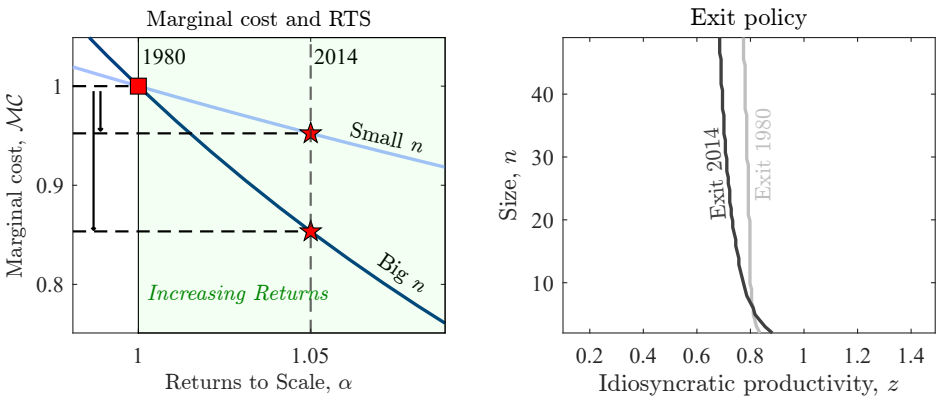
In the model, the firm-level marginal cost of production declines monotonically from the 1980 steady state to 2014 steady state, meaning that an increase in the returns to scale lowers the marginal cost of production for all firms in the latter economy. Moreover, as shown by the graph, the marginal cost of production, after an increase in the returns to scale above one (the green area), declines much more for bigger firms. This is a well-known feature of increasing returns to scale in the production function (which, as shown in Section 1.2.3, is the empirically relevant case), where an increase in the input allows firms to produce more than proportionally, effectively lowering the quantity of input needed to achieve a given level of outputs. Therefore, the increase in returns to scale has a differential effect across firms, favoring bigger firms in the economy.

The decline in the marginal cost of production has three direct implications: (i) it increases the willingness of firms to scale up, and hence, their expenditures devoted to customers acquisition; (ii) it raises the firm-level markups; and (iii) it weakens the selection process in the model.

First, with lower marginal costs of production, firms want to achieve a bigger size; as a consequence, they devote more resources to activities related to accumulating new customers. This implies that, in the 2014 steady state, firms will devote relatively more resources to non-production costs compared to production costs. As a consequence, there will be a shift away from production costs toward non-production costs, as observed in the Compustat data by De Loecker et al. (2020).

Second, the decline in the marginal cost of production increases the surplus generated by the customer-firm relation, as firms are effectively better at producing. However, because there is an incomplete pass-through of costs in the model, only a fraction of this increase in the surplus will be passed on customers in the form of lower prices. Firms will retain the remaining fraction in the form of higher markups. Therefore, due to

Figura 1.9: Returns to Scale, Marginal Costs, and Selection



Note. The figure on the left shows the relation between the firm-level marginal cost of production and the returns to scale, α, for different levels of customers, that is, size. The dark blue line represents the marginal cost of a Big firm (high customer firm), whereas the light blue line represents the marginal cost for a Small firm (low customer firm). The figure on the right shows the exit threshold in the 1980 and 2014 steady state over the firms' state space. The 2014 steady state has the same calibration as the 1980 one but with higher returns to scale, that is, $\alpha = 1.05$, The dark light grey line is the 1980 threshold, whereas, the dark grey line is the 2014 threshold.

the decline in the marginal cost of production, firms will experience an increase in markups in the 2014 steady state.

Third, the decline in the marginal cost of production weakens, on average, the firms' selection process. This can be seen in Figure 1.9 on the right. The figure plots the exit threshold over the firms' state space in the 1980 and 2014 steady state. It can be seen that, in the latter steady state, the exit threshold moves, on average, to the left, implying that less productive firms will be able to operate in the economy. This is because, in the 2014 economy, firms are better at producing, which increases their resilience to adverse productivity shocks.

This decline in selection has two direct implications: (i) it lowers the entry rate of firms in the economy; and (ii) it decreases the churning of firms, which has as a consequence a decline in the reallocation of labor.

First, when the selection declines, the exit rate declines as well. In a stationary equilibrium, where the exit rate must equal the entry rate, this translates into a one-to-one decline in the entry rate. Second, the decline in the entry and exit rate translates into a firms' lower attrition rate. This implies that the reallocation of labor between entrant and exciters declines, and hence, the overall labor reallocation declines. Thus, the aforementioned decline in the selection translates into a decline in business dynamism.

As a final remark, I emphasize that, although selection weakens on average, it increases for the smallest firms in the economy, that is, firms with few customers. This is because small firms have to attract new customers. However, this is more costly in the 2014 steady state because these small firms must compete with the biggest firms that can now exploit their scale economies to compete for customers through very low prices.[43] Therefore, only marginally more productive small firms can do so, and consequently, this increases the selection process for small firms. Moreover, given that small firms are mostly new entrant firms, this acts as an entry barrier, which ulteriorly exacerbates the decline in the entry rate in the new economy.

1.5.2 Mechanism Validation

In this section, I test in the data the main qualitative predictions outlined above. The model predicts that a rise in firm-level returns to scale should increase markups and selling-related expenditures and decrease business dynamism. To this extent, I first show in the data that there has been a rise over time in the firm-level selling-related expenditures relative to production costs.[44] Then, exploiting only cross-sectoral variation in the data, I

[43]This kind of behavior has recently received a great deal of attention in the antitrust debate; see Khan (2016). The model rationalizes this behavior as the outcome of the rise of scale economies that big firms, such as Amazon, can take advantage of to set prices lower those of their smaller competitors.

[44]I focus my attention only on the rise over time in firm-level selling-related expenditures relative to production costs because it is relatively less known. The rise in markups and the decline in business dynamism have been extensively documented; see De Lo-

document that, in sectors where returns to scale are higher, selling expenditures and markups are higher, whereas business dynamism is lower.

The Rise in Selling-Related Expenditures

The model predicts that a rise in returns to scale increases firm-level expenditure in selling-related activities at the expense of production costs. Therefore, it is natural to look at the evolution of this ratio over time as a first test of the implications outlined above. To this extent, I look at the evolution over time of the average selling ratio, as defined in Appendix 1.7.1.[45]

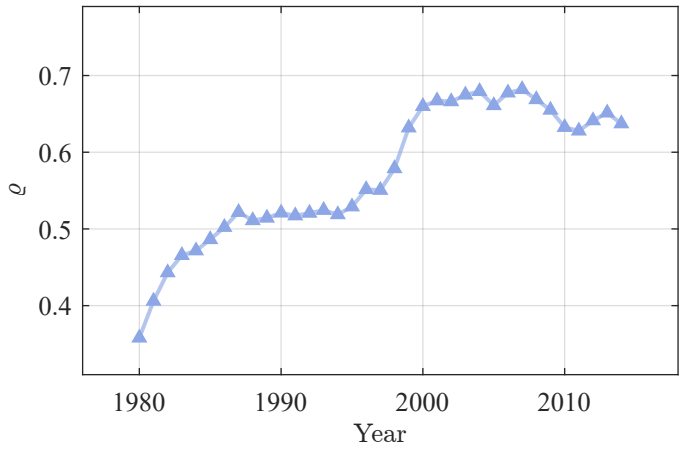
Figure 1.10 shows the evolution of the average selling ratio. At the beginning of the sample, the average selling ratio was approximately 0.4, then rose almost up to 0.7 around 2000, and then went back roughly to 0.65 by the end of the period. Hence, the measure has experienced an increase slightly above 62% over the period of analysis. Therefore, I conclude that, with higher returns to scale over time, we should expect to observe higher firm-level expenditures in selling-related activities, relative to production costs, from the data. In Appendix 1.7.3, I show that using an alternative measure of selling-related activities produces similar results.

I finish emphasizing that this is an aspect peculiar to the theory outlined in this paper. Only a model in which the market power is a long-term investment would produce such an empirical pattern in selling-related activities relative to production costs. Models in which the market power is derived from the love for variety (see, for example, Dixit and Stiglitz

ecker et al. (2020) and Decker et al. (2014).

[45]The rise in non-production costs over time relative to total costs has already been documented by De Loecker et al. (2020). However, I focus on a different measure, that is, the ratio of selling-related costs (these are similar to the non-production costs analyzed in De Loecker et al. (2020); see Appendix 1.7.1 for a more detailed explanation) to production costs. This has two advantages: (i) it avoids the challenges of computing the costs of holding capital, which requires additional assumptions; and (ii) it focuses directly on the shifts in those particular costs emphasized by the theory in this paper. However, regarding the results, both measures show a clear rise over time.

Figura 1.10: Average Selling Ratio



Note. The figure plots the evolution of the average selling ratio between 1980 and 2014. The measure is constructed using a simple average.

(1977), Kimball (1995), and Atkeson and Burstein (2008)) or from search frictions with only pricing strategies and no expenses devoted to the acquisition of new customers (see, for example, Paciello et al. (2019) and Roldan-Blanco and Gilbukh (2020)), would not be able to produce an endogenous increase in selling-related activities relative to production costs, as in the one documented above.

The Cross-Sectoral Implications of Higher Returns to Scale

Here, I test in the cross-section of sectors the qualitative predictions of the model outlined above. The model would predict that, in sectors where returns to scale are higher, we should expect to observe lower business dynamism (lower entry and reallocation rates), higher markups, and higher selling-expenditures relative to production costs (selling ratio). To do so, I regress all these variables against sector-level returns to scale, as estimated in Sections 1.2.3. The sector-level entry and reallocation rates are from the BDS data; the sector-level cost-weighted markups are computed

with the method proposed by Hall (1988) and De Loecker and Warzynski (2012); and the sector-level selling ratio is computed, as described in Appendix 1.7.1.

Taula 1.2: Returns to Scale and Cross-Sectoral Correlations

	Business Dynamism			
	(1)	(2)	(3)	(4)
	Entry rate	Reallocation rate	Markups (log)	Selling ratio
Returns to scale	−0.047***	−0.145***	0.354*	0.839***
	(0.010)	(0.020)	(0.213)	(0.113)
Observations	518	518	722	722
R-squared	0.602	0.764	0.144	0.687
Sector-Time FE	v	v	v	v

Notes. Fixed effects are at the sector-time level, where the sector is at the 1-digit level. Robust standard errors are in parenthesis. *** p-value ¡ 0.01, ** p-value ¡ 0.05, * p-value ¡ 0.1.

Table 1.2 shows the results.[46] The coefficients are estimated using only within sector-time variation; this is important because most of these variables have time trends, which could give rise to spurious correlations. The regressions clearly show that in sectors where firms produce with higher returns to scale, business dynamism is lower; that is, entry and reallocation rates are lower.[47] All coefficients are significant. In Appendix 1.7.3, I show that using alternative measures of selling-related activities produces similar results.

[46]It is worth noticing that the coefficient related to business dynamism is estimated over a smaller sample. This is because the BDS data merge some sectors; for example, manufacturing, which normally is classified by NAICS codes 31-32-33, in BDS is a unique sector.

[47]In related work, Gao and Kehrig (2017) use Census data to show that, where firms produce with higher returns to scale, the average firm size and concentration are higher. This reinforces the correlations documented above, as it confirms in a different dataset similar patterns compared to the analysis emphasized in this section.

Although these correlations seem to propose a rise in returns to scale as an underlying factor behind some recent firm-level trends, I should caution the reader from any causal interpretation of these relations. However, the presence of these correlations indeed supports the economic forces outlined by the model.

1.5.3 Quantitative Implications

This section explores the main quantitative implications of the rise in returns to scale. First, it analyzes the effect that rising returns to scale have on business dynamism, markups, and other aggregate trends. Second, it studies the implication of this technological change on the distribution of markups. Third, it examines the consequences of the rise in returns to scale for firm-level responsiveness of employment growth to productivity shocks. Appendix 1.8.4 shows additional quantitative results.

Rising Returns to Scale and Aggregate Trends

Here, I study the quantitative implication of a 5% rise in returns to scale (from 1 to 1.05, as documented in Section 1.2.3) for the decline in business dynamism, the rise in markups, and the evolution of other trends, such as the rise in concentration and the rise in firm-level selling-related activities. To this end, I compare two steady states, the 1980 one, calibrated as documented in Section 1.4.2, and the 2014 one, where I only let the returns to scale α rise from 1 to 1.05, keeping all the other parameters fixed.

Table 1.3 shows the quantitative implications of the rise in returns to scale for the aggregate trends. The model can explain an important share of the decline in business dynamism, as it explains 62% of the decline in the entry rate and 70% of the decline in the reallocation rate. Moreover, because the rise in returns to scale inherently favors the bigger and oldest firms in the economy, the model can explain 90% of the rise in the share of the old firms (firms with 11+ years) and 96% of the decline in the employment share of young firms (firms with less than 5 years).

Taula 1.3: Effect of Rising Returns to Scale

			Change			
	1980 S.S.	2014 S.S.	Model	BDS	Compustat	Model/Data
Business Dynamism						
Entry rate	0.139	0.104	-25%	-40%	–	62%
Reallocation rate	0.294	0.237	-19%	-27%	–	70%
Share of old firms	0.322	0.467	$+45\%$	$+50\%$	–	90%
Employment share of young firms	0.204	0.094	-69%	-56%	–	96%
Markups						
Avg. markup (cost-weighted)	1.202	1.229	$+2\%$	–	$+7\%$	29%
Others						
Avg. selling ratio	0.4	0.65	$+9\%$	–	$+62\%$	14%
Concentration (HHI)	7.003e-06	7.440e-06	$+6\%$	–	$+33\%$	18%

Notes. All variables are calculated coherently with their definitions, as used in the data. The average markup is calculated using cost weights, whereas the average selling ratio is calculated using a simple average across firms. Concentration is calculated as described in Grullon et al. (2019). The data sources are BDS and Compustat. To calculate the empirical moments from the 1980s I use the time window 1977-1985, whereas for the empirical moments from the 2014, I use simple the values in that year. The last column shows the fraction of the overall empirical variation explained by the model.

Moreover, the model is able to explain 29% of the rise in the average cost-weighted markup.[48] I focus on the evolution of the cost-weighted measure, which is the welfare-relevant aggregate measure, as documented by Grassi (2017) and Edmond et al. (2018). However, in the next session, I look at the evolution over time of the markup distribution to analyze the features of the rise in the sales-weighted measures that are related to the reallocation of economic activity toward bigger firms. Although the model explains a non-negligible fraction of the rise in the aggregate

[48] I document a 7% increase in the cost-weighted markup; De Loecker et al. (2020) report a rise of approximately 10%. This difference is mainly due to the way I clean Compustat. In particular, I drop all firms that are not incorporated in the US and all utilities and financial firms. Notice that these choices do not change the qualitative behavior of markups compared to the paper above. However, they lower their rise. Therefore, readers should keep this caveat in mind when interpreting the numbers.

markups, it cannot explain most of it. This shows that the rise in returns to scale does not seem to be the only force behind the rise in the data, suggesting that there may be additional mechanisms at work in the US economy that can account for the unexplained rise.

Finally, the model is also consistent with the rise in selling-related expenditures and product market concentrations, as observed in the data. In particular, the model can explain 14% of the rise in the benchmark measure of selling-related expenditures and 45% of the increase in the alternative advertisement-based measure, as documented in Appendix 1.7.3. Although the model explains only a fraction of this rise, we can still define this as a success, given that this endogenous rise is a distinctive feature of this model, where firms actively invest in their market power (see Section 1.5.2 for a more detailed explanation of this point). The model also explains 18% of the rise in concentration, which shows that the model captures the reallocation of economic activity toward bigger firms that have been documented empirically by Kehrig and Vincent (2021), Autor et al. (2020), and De Loecker et al. (2020).

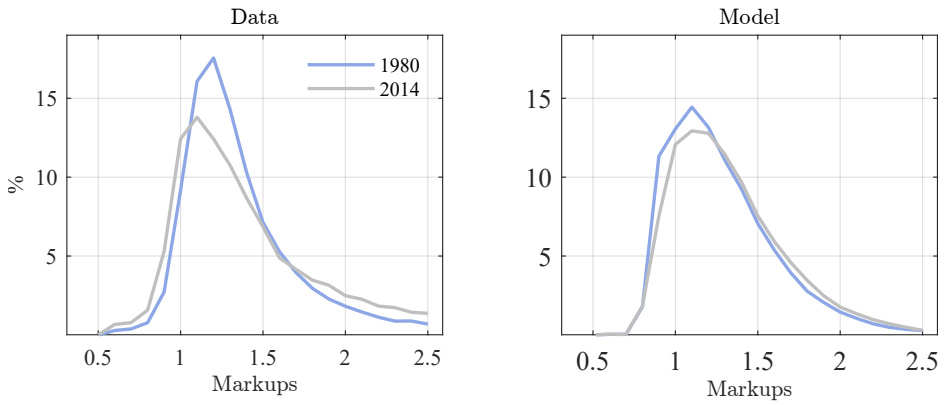
Evolution of the Markup Distribution

Analyzing the rise in markups, De Loecker et al. (2020) show that there has been a substantial change in their distribution overall. In particular, they notice that much of the rise in the average markup is due to reallocation of the economic activity toward the right tail of the distribution—in their words, there has been a fattening of the right tail of the markup distribution.

In this section, I look at this prediction in the model. Hence, I compare the model-implied distribution of markups in both steady states, that is, in the 1980s and 2014, with the one in the data as documented by De Loecker et al. (2020).

Figure 1.11 presents the results. The figure on the left shows the empirical distribution of markups in the 1980s, light blue, and in 2014, light grey. The figure on the right shows the model-implied distribution of markups in the 1980s (light blue) and in 2014 (light grey). It can be seen

Figura 1.11: Distributions of Markups—Model and Data



Note. The figure on the left shows the empirical markup distribution in 1977-1985 (light blue) and in 2005-2014 (light grey). The figure on the right shows the model-implied markup distribution in the 1980 calibration (light blue) and in the 2014 calibration (light grey). Both distributions are shown within the [0.5, 2.5] range.

that the model qualitatively captures the overall change in the distribution of markups. Specifically, in the 2014 steady state, the model exhibits a considerable fattening of the right tail, compared to the 1980s steady state, as the one portrayed in the data and emphasized in De Loecker et al. (2020).

This, in the model, happens because the rise in returns to scale reallocates the economic activity toward bigger firms, which are also the ones the higher markups. This reallocation toward bigger firms translates into a fatter right tail of the markup distribution. Therefore, the model produces the rise in the average markup jointly with the distributional changes emphasized by the empirical works of Kehrig and Vincent (2021), Autor et al. (2020), and De Loecker et al. (2020).[49]

[49]Edmond et al. (2018) demonstrated with an exact decomposition that this rise in the variance of the markups distribution is the main reason why the sales-weighted average markup rose by more compared to the cost-weighted one. The model captures this qualitatively, as it produces a rise in sales-weighted markup of 3%, which is approximately

Declining Responsiveness

In this section, I look into additional facts highlighted by the empirical literature related to the decline in business in the US. In particular, Decker et al. (2020) show that an important component of the decline in business dynamism is the fact that firms in recent decades have responded less to productivity shocks, that is, conditional to a productivity shock, they expand (or contract) less.

To analyze this feature in the model and in the Compustat data, I proceed in two steps: (i) I replicate the spirit of their empirical investigation, both in the model and in the data; and (ii) I propose an exact decomposition to shed light on the forces behind this decline in responsiveness.

Therefore, in the data, I implement the following regression:

$$g^{\ell}_{it+1} = \alpha + \beta\, a_{it} \otimes \mathcal{F}(t) + \boldsymbol{X}'_{it}\boldsymbol{\gamma} + \phi_{st} + \varepsilon_{it}, \tag{1.43}$$

where $g^{\ell}_{it+1} \equiv 2 \times (\ell_{it} - \ell_{it-1})/(\ell_{it} + \ell_{it-1})$ is the growth rate of employment, a_{it} is the empirical measure of total factor productivity revenue (TFPR), that is, the residual from the production function in Section 1.2.3, $\mathcal{F}(t)$ is a flexible function of time, $\boldsymbol{X}_{it}$ is a vector of controls, and ϕ_{st} are sector-time fixed effects. The symbol $\otimes$ represents the full interaction between the two variables. Therefore, the coefficient of interest will be the β associated with the interaction between a_{it} and $\mathcal{F}(t)$, which captures the evolution over time of the marginal effect of changes in productivity.[50]

Results are presented in Table 1.4. The first three columns show the

17% of what I observe in my calculations, compared to a 2% rise in the cost-weighted markup.

[50] In the model, as I only have a simulated panel for the two distinct steady states, I have to run a different regression. In particular, I run the following regression in both steady states:

$$g^{\ell}_{it+1} = \alpha + \beta\, a_{it} + \boldsymbol{X}'_{it}\boldsymbol{\gamma} + \phi_{st} + \varepsilon_{it}, \tag{1.44}$$

where I do not allow for time-dependent functions. However, the regression follows the same spirit and allows for a very similar interpretation. Therefore, to analyze the decline in responsiveness, within the model, I look at the difference of the estimated coefficients in the two steady states, that is, $\beta^{2014} - \beta^{1980}$.

Taula 1.4: Declining Firm-Level Responsiveness

	(1) Compustat	(2) Compustat	(3) Compustat	(4) Model
$\widehat{\beta}^{2014} - \widehat{\beta}^{1980}$				-0.027
$a_{it} \times Year$	-0.001^{***}			
	(0.000)			
$a_{it} \times \mathcal{I}_{t\geq 2000}$		-0.006^{**}		
		(0.002)		
$a_{it} \times \mathcal{I}_{t\in[1990,2000)}$			-0.007^{**}	
			(0.003)	
$a_{it} \times \mathcal{I}_{t\in[2000,2010)}$			-0.009^{***}	
			(0.003)	
$a_{it} \times \mathcal{I}_{t\in[2010,2015)}$			-0.011^{***}	
			(0.004)	
Controls	v	v	v	v
Sector-Time FE	v	v	v	
Observations	$143,771$	$143,771$	$143,771$	
R-squared	0.037	0.038	0.038	

Note. The table reports the change in firm-level responsiveness to productivity shocks. The controls are size, the interaction of employment with the time function, and past productivity. In column (1), I allow for a simple linear trend. In columns (2) and (3), I instead allow for a more flexible set of dummies, where $\mathcal{I}_{t\in T}$ equals 1 when $t \in T$.

decline in firm-level responsiveness in the data with three different specifications: (i) a linear trend; (ii) a dummy capturing responsiveness after 2000; and (iii) a set of dummies that captures the responsiveness in each decade having as a benchmark the first decade. Both the first (parametric) and last two (semi-parametric) regressions show a statistically significant decline in firm-level responsiveness over time. In particular, the first specification shows a decline in responsiveness, between 1980 and 2014, of 0.035, whereas the last specification shows a decline of 0.011.

The last column shows the evolution of responsiveness in the model.

In the model, firm-level responsiveness also declines between the two steady states. In particular, we can see that this decline of 0.027 lies in the empirical range reported above.

Finally, to understand which forces lie behind the above decline in the model, I define firm-level responsiveness as:

$$\frac{\Delta \log \ell_{it}}{\Delta z_{it}} = \frac{1}{\alpha} \times \left[\frac{\Delta \log y_{it}}{\Delta z_{it}} - 1 \right], \tag{1.45}$$

where α is the returns to scale, and $\Delta \log y_{it}/\Delta z_{it}$ is the output growth associated with productivity growth.[51] Equation (1.45) shows that the rise in returns to scale translates directly into a decline in firm-level responsiveness.[52] Moreover, the rise in returns to scale can also affect responsiveness indirectly through its effect on the output growth associated with productivity growth. Taking stocks, in the model, the direct effect of rising returns to scale dominates, and hence, firm-level responsiveness declines after the aforementioned technological change, making the model consistent with the findings documented by Decker et al. (2020).

1.6 Conclusion

In this paper, I documented empirically that US firms have undergone a technological change biased toward higher returns to scale. In particular, leveraging the Compustat data and state-of-the-art production function estimators, I document that firm-level returns to scale experienced a 5% increase, going from 1 in 1980 to 1.05 in 2014. Moreover, I find

[51] This definition of firm-level responsiveness is slightly different from the one implied by the regressions above. However, notice that controlling in the regressions for past productivity allows for a similar interpretation of firm-level responsiveness: the growth in employment associated with productivity growth. In light of this and consistent with Decker et al. (2020), I stick with the above regression analysis as the benchmark measure of firm-level responsiveness. However, equation (1.45) is still useful in understanding which mechanism is behind the decline in the model.

[52] The difference in the brackets is always positive. In particular, it can be shown that $\Delta \log y_{it}/\Delta z_{it} = 1 + \alpha \Delta \log \ell_{it}/\Delta z_{it}$, where $\Delta \log \ell_{it}/\Delta z_{it} > 0$.

that this rise is happening within all sectors—suggesting a technological interpretation—and is not the outcome of a reallocation of economic activity toward high returns to scale sectors.

To understand the implications of this technological change for some of the main trends in the US economy, I propose a novel heterogeneous firms model grounded in search frictions in the product market. Search frictions make the model consistent with several features of the microdata: (i) they microfound endogenous heterogeneous markups; (ii) entail firms' active expenditures to attract customers, and (iii) imply that firms grow through the accumulation of new customers, which empirically accounts for 70% of their life-cycle growth. In the model, because of the central role of prices for attracting and retaining customers, changes in returns to scale, affecting firm-level marginal costs, influence the firm-level ability to price, grow, and charge markups.

I calibrate the model with firm-level data and use it to quantify the effect of the 5% rise in returns to scale. In the model, such a technological change can explain between 62-70% of the decline in business dynamism, 29% of the increase in the average cost-weighted markup, and between 14-45% of the rise in expenditures devoted to customer acquisition. The model captures all these, while being consistent with additional microfacts, such as the aging of US firms, the reallocation of economic activity toward high-markup firms, and the decline in firm-level responsiveness to productivity shocks.

Several potential directions are left unexplored. It would be interesting to study the implications of the increase in returns to scale for the increase in merger and acquisition activities, as witnessed in recent decades. Moreover, it would be valuable to introduce in the model horizontal product differentiation as an additional source of market power. I leave these questions to future research.

1.7 Empirical Analysis Appendix

1.7.1 Data

This section presents the construction of the main sample and main variables, providing summary statistics for the final sample. Then, it shows how to construct the user cost of capital and explains which variable is used in the production function estimation as labor (variable) input. Finally, it shows the construction of the measures of selling-related activities and markups.

Main Sample, Variables, and Summary Statistics

I use Compustat from 1977 to 2014. I drop all firms whose Foreign Incorporation Code (FIC) is not equal to USA. Then, I linearly interpolate when there is one missing between two available data points SALE, COGS, XSGA, EMP, PPEGT, PPENT, XRD, XLR, XPR, XRENT, RECD, DP for data quality. I exclude utilities (SIC codes between 4900-4999) because they are heavily price regulated, and I also exclude financial firms (SIC codes between 6000-6999) because their balance sheets are dramatically different from other firms.

To construct the firm-level total stock of capital, I use the perpetual inventory method (PIM). In particular, with PIM, capital is defined as:

$$k_{it} = (1 - \delta)k_{it-1} + x_{it}, \tag{1.46}$$

where $x_{it} - \delta k_{it-1} = \texttt{PPENT}_{i,t} - \texttt{PPENT}_{it-1}$ is the net investment, and the initial capital stock, k_{i0}, is initialized using the first available entry of PPEGT.[53]

For data quality, I interpret as mistakes zero or negative in SALE, k, EMP, or XSGA, and I drop those observations; moreover, if SALE, k, EMP are missing, I drop these observations too; however, if XSGA is missing, I set it to zero. Finally, if XRD, XLR, XPR, XRENT, RECD, or DP are negative

[53] Given that a measure of real capital is needed for the analysis, I deflate the measure of net investment with the appropriate deflator.

or missing, I treat them as zeros. To obtain a real measure of the main variables, I deflate them with the GDP deflator; I deflate investment and capital stock by the investment good deflator.[54] The table below presents a few basic summary statistics for a few leading variables used in the analysis.

Taula 1.5: Summary Statistics (1977-2014)

	Sales	Cost of Goods Sold	Employment	Capital Stock (*Book Value*)	Capital Stock (*PIM*)	Age
Mean	1,873,553	1,296,868	7,056	1,005,617	728,260	13
25^{th} Percentile	22,553	13,896	115	5,756	3,552	5
Median	139,060	84,909	638	36,079	24,323	11
75^{th} Percentile	751,619	483,007	3,500	241,352	169,204	19
No. Obs.	168,496	168,496	168,496	167,884	168,496	168,496

Note. Summary statistics of cleaned Compustat dataset between 1977 and 2014. All variables but Age are in thousands US$. Sales and Costs of Goods Sold are deflated with the GDP deflator using the base year 2012, whereas both capital stocks are deflated using the investment deflator with the base year 2012.

User Cost of Capital

As mentioned in the main body of the paper, one of the challenges of using the cost shares approach is that it requires a measure of the user cost of capital. To this end, I define the user cost of capital as:

$$r_t = i_t - E_t\pi_{t+1} + \delta, \tag{1.47}$$

where i_t is equal to the nominal interest rate, $E_t\pi_{t+1}$ is expected inflation at time t, and δ is the depreciation rate of capital. I take the annual Moody's Seasoned Aaa Corporate Bond Yield as an empirical proxy for the nominal interest rate, the annual growth rate of the Investment Nonresi-

[54] Deflators are taken from the NIPA tables.

dencial Price Deflator to calculate the expected inflation, and the depreciation rate is calibrated to $\delta = 0.1$, as in the rest of the paper.[55,56,57]

Variable Input in Production

Recent work based on Compustat, particularly since De Loecker et al. (2020), has used the item Cost of Goods Sold (COGS) as the preferred measure of variable input in production. This choice was motivated by the need for a bundle of variable input expenditures to calculate firm-level markups. However, despite being an unavoidable choice, using it as a measure of variable input imposes an additional assumption in the estimation, as it assumes that labor and materials are perfectly substitutable.

However, as the primary goal of this paper is to estimate the returns to scale, and hence output elasticities, and not the markups, I favor a direct measure of the firm-level variable input. In particular, I use as a benchmark measure the variable EMP, which represents the number of employees in a given firm, and show robustness exercises using COGS. Therefore, to be consistent with this approach, when I calculate cost shares, I need to construct a measure of labor cost, $w_{it}\ell_{it}$. To do so, I use the labor cost expenditure (XLR) reported by a subsample of firms. For the firms that report it, I calculate the labor cost per worker defined as $w_{it} \equiv \texttt{XLR}/\texttt{EMP}$, and then I calculate its within-sector median and use it to impute the labor cost for the firms that do not report it as $w_{it}\ell_{it} = \widehat{w}_{st} \cdot \texttt{EMP}_{it}$.

Selling-Related Expenditure

In this section, I present the two main approaches used to compute firm-level selling-related activities. Unfortunately, in Compustat, there is no

[55]Moody's Seasoned Aaa Corporate Bond Yield: https://fred.stlouisfed.org/series/AAA

[56]Investment Price Deflator: https://fred.stlouisfed.org/series/A008RD3Q086SBEA

[57]I estimate an AR(1) process on the annual growth rate of the Investment Nonresidential Price deflator and define the contemporaneous expected inflation as $E_t\pi_{t+1} = \mu + \rho\pi_t$.

perfect way to compute firm-level selling-related activities; therefore, while presenting the two approaches, I will emphasize their virtues and their weaknesses.

Benchmark measure. To measure firm-level selling-related expenditures, I use Selling General and Administrative (XSGA). This item in Compustat has been the focus of many recent studies such as: Gourio and Rudanko (2014), Ptok et al. (2018), Afrouzi et al. (2020), and Morlacco and Zeke (2021).[58] However, despite the acknowledged ability of Selling General and Administrative to capture firm-level selling-related expenditure, it is well known that this item reports many expenditures that are not directly related to selling efforts, such as bad debt expenses, expenditure in pensions and retirement, rents, and expenditure in research and development.[59] Therefore, to partially overcome the aforementioned limitations, my adjusted measure of *selling-related expenditure* is defined as:

$$\mathtt{S}_{it} = \mathtt{XSGA}_{it} - \mathtt{XRENT}_{it} - \mathtt{XPR}_{it} - \mathtt{RECD}_{it} - \mathtt{XRD}_{it}, \tag{1.48}$$

where XSGA is an expenditure in Selling General and Administrative, XRENT is an expenditure in Rents, XPR is an expenditure in Pensions and Retirement, RECD is an expenditure due to Bad Debts, and XRD is an expenditure in Research and Development.

Alternative measure. As an alternative measure to the above measure, I use the Compustat variable XAD, which reports the firm-level expenditure in advertisements. This is the only available item in Compustat that

[58] In particular, Ptok et al. (2018) document that Selling General and Administrative is particularly good at capturing firm-level sales force spending.

[59] For a more exhaustive discussion on how research and development are accounted for in Compustat, see Peters and Taylor (2017). For an extensive list of items reported in Selling General and Administrative, see Afrouzi et al. (2020). In my list, I reported to the best of my knowledge only the items reported in Compustat that are accounted for in Selling General and Administrative.

measures only (and somehow cleanly) selling-related costs; however, this measure suffers from two main drawbacks: (i) it reports the cost of advertising media (radio, television, newspapers, periodicals) and promotional expenses but excludes selling and marketing expenses, and (ii) half of the observations are missing.

Firm-Level Markups

Throughout the paper, markups are constructed following Hall (1988) and De Loecker and Warzynski (2012); hence, the firm-level markup is given by:

$$\mu_{it} = \widehat{\beta}_{st}^{cogs} \cdot \frac{\texttt{SALE}_{it}}{\texttt{COGS}_{it}}, \tag{1.49}$$

where the $\widehat{\beta}_{st}^{cogs}$ is the output elasticity to COGS. To ease the comparability between this paper and the seminal work by De Loecker et al. (2020), I use their measure of this elasticity. However, the results are robust to using the alternative measure of $\widehat{\beta}_{st}^{cogs}$ presented in Appendix 1.7.2.

1.7.2 Additional Robustness Production Function

Here, I document the robustness of the results in Section 1.2.3. To this end, first, I present the alternative specification that I will use. Second, I present the results from these specifications, both for the average returns to scale and for the within-between sectors decomposition.

Alternative Control Function: Investment. Here, I document the robustness of the rise in returns to scale to alternative control functions such as investment. This particular control function has been pioneered by Olley and Pakes (1996) and discussed extensively by Ackerberg et al. (2015).[60] To apply the methodology presented in Section 2.9.2 to the case

[60] A known drawback of using this alternative measure as a control function for the estimation is the presence of many zeros in investment (see, Levinsohn and Petrin (2003)). However, in Compustat, this is a minor issue, as the number of observations that are zero

in which investment is used as a control function, equation (2.24) has to be modified as:

$$q_{it} = \mathcal{P}(k_{it}, \ell_{it}, x_{it}, \mathbf{d}_{it}) + \varepsilon_{it}, \tag{1.50}$$

where x_{it} is now the firm's investment. Given this new augmented equation, the rest of the procedure is the same as the one outlined in Section 2.9.2.

Alternative Variable Input: Cost of Goods Sold. Here, I adopt an alternative specification of the production function, as used in the recent paper by De Loecker et al. (2020). To this end, I use `COGS` instead of `EMP` as the variable input. This is a necessary shortcut to estimate firm-level markups in Compustat. However, it imposes an alternative set of assumptions as the true estimated production function is:

$$q_{it} = \beta^k k_{it} + \beta^{cogs}(\ell_{it} + m_{it}) + \omega_{it} + \varepsilon_{it}, \tag{1.51}$$

where m_{it} is the firm's materials. Equation (1.51) implicitly entails two additional assumptions: (i) first, now the production function is defined as the gross output, and hence, is partially subject to the identification criticisms laid out in Gandhi et al. (2020); and (ii) second and last, given that `COGS` is the sum of all production costs (particularly labor and materials), its adoption as an input in production implicitly assumes that labor and material are perfect substitutes within the production process.

Additional Dynamic Input: Intangible Capital. Recently, there has been a particular emphasis on the role played by the rise of intangible capital at the firm level.[61] Therefore, this could generate some concerns as that the rise in returns to scale could potentially be partially driven by the rise in unmeasured intangible capital as input in production. To

is particularly small relative to most of the dataset—this is due to the fact that Compustat is a firm-level dataset containing mostly big firms.

[61] In particular, Chiavari and Goraya (2021) show that intangible capital is rising dramatically as an input in production.

address this concern, I estimate a new production function, augmented by intangible capital, given by:

$$q_{it} = \beta^k k_{it} + \beta^\imath \imath_{it} + \beta^\ell \ell_{it} + \omega_{it} + \varepsilon_{it}, \tag{1.52}$$

where $\imath_{it}$ is the intangible capital in production. This new specification entails an additional challenge: namely, the firm-level measurement of intangible capital. To this end, I take advantage of the balance sheet intangible capital and capitalize firm-level knowledge capital, as done in Chiavari and Goraya (2021). The balance sheet intangible capital is given by:

$$\imath_{it}^{balance\ sheet} = \texttt{INTAN}_{it} + \texttt{AM}_{it} - \texttt{GDWL}_{it}, \tag{1.53}$$

where `INTAN` is the net balance sheet intangible capital, `AM` is the ammortization of the balance sheet intangible capital, and `GDWL` is goodwill. Knowledge capital is given by:

$$\imath_{it}^{knowledge} = (1 - 0.30)\imath_{it-1}^{knowledge} + \texttt{XRD}_{it}, \tag{1.54}$$

where the depreciation rate is set to 30%, close to the empirical estimates by Ewens et al. (2019), `XRD` is the firm-level expenditure in research and development, and $\imath_{i0}^{knowledge}$ is set equal to zero. Finally, the total firm-level intangible capital is given by:

$$\imath_{it} = \imath_{it}^{balance\ sheet} + \imath_{it}^{knowledge}. \tag{1.55}$$

Alternative Production Function: Translog. Finally, I explore the robustness of the rise in returns to scale to an alternative production function specification. In particular, in this section, I adopt the following translog specification given by:

$$q_{it} = \theta_1^k k_{it} + \theta_1^\ell \ell_{it} + \theta_2^k k_{it}^2 + \theta_2^\ell \ell_{it}^2 + \theta_3^{k\ell} k_{it}\ell_{it} + \omega_{it} + \varepsilon_{it}. \tag{1.56}$$

To estimate the translog production function, I follow the methodology outlined in Section 2.9.2. However, the output elasticities are now given by:

$$\beta^k = \text{median}\Big\{\theta_1^k + 2\theta_2^k k_{it} + \theta_3^{k\ell}\ell_{it}\Big\}, \tag{1.57}$$

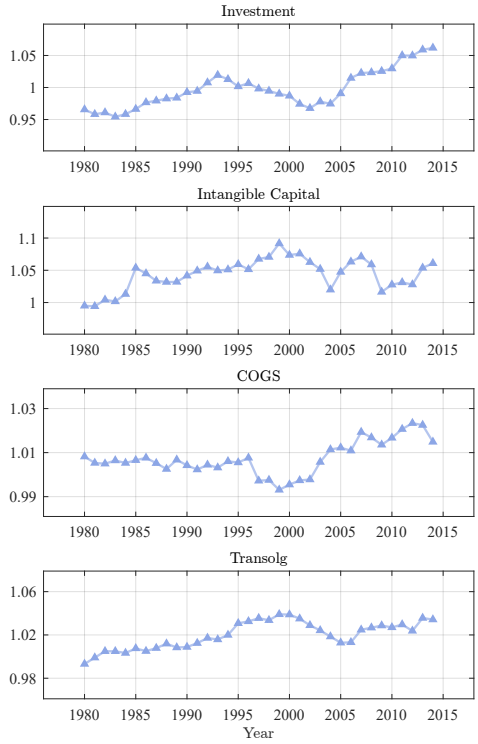
$$\beta^\ell = \text{median}\Big\{\theta_1^\ell + 2\theta_2^\ell \ell_{it} + \theta_3^{k\ell} k_{it}\Big\}. \tag{1.58}$$

Therefore, the returns to scale implied by the production technology from equation (1.56) is given by $\alpha = \beta^k + \beta^\ell$.

Results from Alternative Specifications. The results from the above specifications are presented in Figures 1.12 and 1.13. Figure 1.12 shows the evolution of the sales-weighted average returns to scale in production from 1980 to 2014 for the different alternative specifications. The first graph shows the robustness exercise when we use investment as a proxy variable. The second graph shows the robustness exercise when we augment the production function with intangible capital as an additional dynamic input. The third graph shows the robustness exercise when we use the cost of goods sold (COGS) as the variable input. Finally, the fourth graph shows the robustness exercise when we adopt a translog specification for the production function.

Figure 1.13 plots the counterfactual evolution of the within and between components implied by the decomposition from equation (1.14); that is, it shows the evolution of the average returns to scale only if the Δwithin component is at play and the evolution of the average returns to scale only if the Δbetween component is at play. The first graph shows the robustness exercise when we use investment as a proxy variable. The second graph shows the robustness exercise when we augment the production function with intangible capital as an additional dynamic input. The third graph shows the robustness exercise when we use the cost of goods sold (COGS) as the variable input. The fourth graph shows the robustness exercise when we adopt a translog specification for the production function.

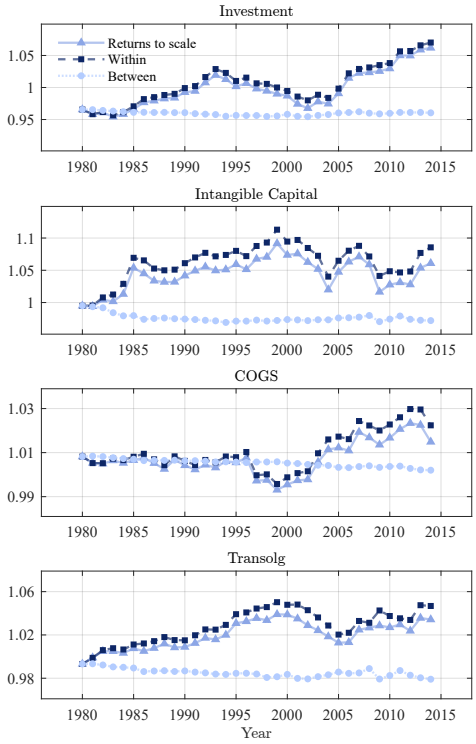
Figura 1.12: Alternative Specifications – Robustness 1



Note. The figures above show the evolutions of the average returns to scale for all four robustness specifications. The first figure shows the evolution of the average returns to scale when we use investment as a proxy variable. The second figure shows the evolution of the average returns to scale when we augment the production function with intangible capital as an additional dynamic input. The third figure shows the evolution of the average returns when we use the cost of goods sold (COGS) as the variable input. The fourth figure shows the evolution of the average returns to scale when we adopt a translog specification for the production function.

Overall, these robustness exercises show qualitative patterns that are similar to the benchmark specification presented in Section 1.2.3. In the 1980s, the average returns to scale are very close to 1 in all specifications, and by 2014, reaches a value between 1.02-1.06. This implies a rise in

Figura 1.13: Alternative Specifications – Robustness 2



Note. The figures above show the results of the decomposition (1.14) for all four robustness specifications. The first figure shows the evolution of the average returns to scale, the within component, and the between component when we use investment as a proxy variable. The second figure shows the evolution of the average returns to scale, the within component, and the between component when we augment the production function with intangible capital as an additional dynamic input. The third figure shows the evolution of the average returns, the within component, and the between component when we use the cost of goods sold (COGS) as the variable input. The fourth figure shows the evolution of the average returns to scale, the within component, and the between component when we adopt a translog specification for the production function.

line with the benchmark specification. Therefore, regardless of the preferred specification, returns to scale in recent years exhibit an increasing trend. Moreover, when we look at the outcome of the decomposition for

all the alternative specifications, we can see that, in all cases, the total increase in the average returns to scale is due to the within component. This reinforces the view that returns to scale are increasing within all sectors of the US economy, regardless of the specification at hand. Taking stocks, we can see from these additional exercises that the main results are a solid feature of the data, suggesting a technological change that is shaping firms' production processes in all sectors of the US economy.

1.7.3 Selling-Related Activity Robustness

In this section, I explore the extent of the robustness of the results concerning the selling-related expenditure measure. In particular, I check whether using an alternative measure based on the firm-level advertisement expenditure, as reported in Appendix 1.7.1, has any effect on the results and the conclusions from the main text. To do so, first, I look at the evolution of this alternative measure over time. Second, I look at the cross-sectoral correlation between this measure and the sector-level measure of returns to scale.

Trend

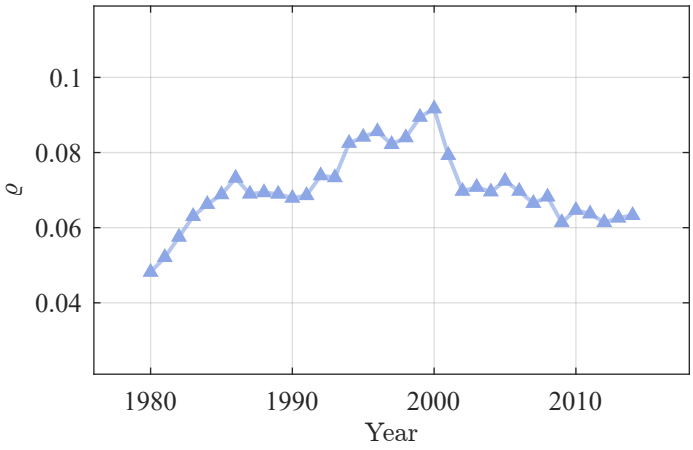
One main prediction of the theory is that the rise in returns to scale implies that the firms spend more on selling-related activities relative to production costs. Therefore, even using the alternative measure of selling-related expenditure, we should observe a rise over time— although we should expect to observe different levels, as explained in Appendix 1.7.1. With this alternative specification, the selling ratio becomes:

$$\varrho_{i,t} = \frac{\texttt{XAD}_{i,t}}{\texttt{COGS}_{i,t}}. \tag{1.59}$$

Figure 1.14 shows the evolution of the selling ratio, as defined in equation (1.59), between 1980 and 2014. This alternative measure of the selling ratio shows a qualitative pattern that is reasonably similar to the benchmark specification. In particular, it increases since 2000 and then

declines slightly until the end of the sample, but, overall, it shows an increase over time, as predicted by the theory.

Figura 1.14: Average Selling Ratio



Note. The figure shows the evolution of the unweighted average selling ratio, as defined in equation (1.59), between 1980 and 2014.

However, the quantitative behavior is very different compared to the benchmark measure. This is not surprising, as this alternative measure (as explained in Appendix 1.7.1) reports only the costs related to advertising media (radio, television, newspapers, periodicals) and promotional expenses, but it excludes selling and marketing expenses. Therefore, despite being highly related to the firm's selling activities, it underrepresents the true costs incurred by the firm to attract and retain customers. Overall, the main takeaway, regardless of the preferred measure to calculate the selling ratio, is that in the US, over the last thirty years, there has been a sizeable increase in selling-related activities relative to production costs.

Cross-Sectoral Correlation

Here, I show that using this alternative measure of firm-level selling-related expenditure yields a similar sign in the correlation between returns

to scale and the measure itself.

Taula 1.6: Effect of Rising Returns to Scale

	Selling ratio − alternative
Returns to scale	0.086***
	(0.025)
Observations	722
R-squared	0.361
Sector-Time FE	v

Notes. Fixed effects are at the sector-time level, where the sector is at the 1-digit level. Robust standard errors are in parenthesis. *** p-value < 0.01, ** p-value < 0.05, * p-value < 0.1.

Table 1.6 shows the cross-sectional correlation between sector-level returns to scale and the selling ratio. The presence of sector-time-level fixed effects is necessary to ensure that the variation that informs the coefficient estimate does not come from common time trends. The table shows a clear positive correlation between the two variables. Therefore, I can conclude that, regardless of the preferred measure to calculate the selling ratio, the sectors in which firms operate with higher returns to scale are correlated with a higher average selling ratio, as predicted by the theory.

1.8 Model Appendix

1.8.1 Model Details

In this section, I go through additional details of the model, emphasizing important concepts related to its solution method. Most of the discussion follows the logic developed in Schaal (2017). First, I present a less general contractual environment relative to the one presented in Section

1.3.3. In this environment, I can solve the model without taking care of the distribution of promised utilities—which is an infinite-dimensional object. This allows the characterization of the real allocation in the economy with standard recursive methods. Second, I comment on how the prices from Section 1.3.6 implement the same allocation as the one characterized under the less realistic contractual environment.

Alternative Contractual Environment

Here, I assume that contracts are complete, state-contingent, and that there is full commitment on both the customer and firm side. Relative to Section 1.3.3, the contracts are complete, and customers also have commitment; this is a very convenient formulation of the contractual environment, despite its lack of realism.

Therefore, in this case, the contract specifies $\{p_{t+j}, \tau_{t+j}, x_{t+j}, d_{t+j}\}_{j=0}^{\infty}$, where p is the price, x is the submarket where the customer searches while being matched, τ is a separation probability, and d is an exit dummy. Each element at time $t+j$ is contingent on the entire history of shocks (z^{t+j}). The fact that the contract specifies x (the submarket in which a firm's customer must search) is a feature of completeness.

Joint Surplus

The additional assumptions embedded in the alternative contractual environment allow the simplification of the problem of the firm. The completeness of contracts, the commitment assumption, and the transferability of utility guarantee that the optimal policies always maximize the joint surplus of a firm and its customers. The model can thus be solved in two stages: a first stage in which I maximize the surplus, and a second stage in which I design the contracts that implement the allocation. The following Bellman equation gives the joint surplus maximization problem for a firm and its current customers:

$$\begin{aligned}\boldsymbol{\mathcal{S}}(z,n) = \max_{\ell,d,n_i',x_i',\tau,x'} \; & nu - w\ell - wf \\ & + \beta\mathbb{E}\Bigg\{(\delta + (1-\delta)d)n\boldsymbol{\mathcal{U}}' + (1-\delta)(1-d)\Bigg[\tau n\boldsymbol{\mathcal{U}}' \\ & + (1-\tau)m(\theta(x'))nx' - \left(\frac{wc}{q(\theta(x_i'))} + x_i'\right)n_i' - w\mathcal{K}(n_i';n) + \boldsymbol{\mathcal{S}}(z',n')\Bigg]\Bigg\},\end{aligned} \tag{1.60}$$

subject to:

$$n' = (1-\tau)(1-m(\theta(x')))n + n_i', \tag{1.61}$$

$$y = e^z F(\ell), \tag{1.62}$$

$$y = n. \tag{1.63}$$

The surplus maximization problem characterizes the optimal allocation of physical resources within a firm: the optimal amount of separations, firm-to-firm transitions, the number of new customers, and the decision of whether to exit or not. Because the utility is transferable, transfers between the firms and their customers leave the surplus unchanged. Elements of the contracts describing the way profits are split, such as prices and continuation utilities, disappear in the surplus maximization problem. In particular, the distribution of promised utilities, $\{\mathcal{C}(j)\}_{j\in[0,n]}$, is not part of the state space, and only the size of the customer base at the production stage n matters.

The first element in the surplus maximization problem is the total utility of the customers followed by the wages and operating cost wf paid by the firm. In the next period, conditional on surviving the exit shock δ, the firm chooses whether to exit or not, a decision captured by the exit dummy d. If a firm chooses to exit, all the customers become unmatched while the firm's value is set to zero, yielding a total utility of $n\boldsymbol{\mathcal{U}}'$. If it chooses not to exit, the firm may then proceed with its separations. The total mass of separations is τn, which provides a total expected utility of $\tau n\boldsymbol{\mathcal{U}}'$ to the customer-firm group. After searching, some customers move to other firms with value x' and contribute the amount $(1-\tau)m(\theta(x'))nx'$

to the total surplus. Simultaneously, the firm proceeds with its customer acquisitions. For each new customer acquisition in the product market segment x_i', the firm incurs a cost of $wc/q(\theta(x_i'))$ and must offer on average a lifetime utility-price x_i' to its new customer, which appears as a cost to the current customer-firm group, and pays, to adjust its customer base, the convex cost $w\mathcal{K}(n_i'; n)$.

Free Entry

Under this different contractual environment, the free entry condition stated in (1.31) can be restated in terms of joint surplus maximization. I redefine the problem faced by an entering firm of type z as follows:

$$\boldsymbol{\mathcal{V}}_e(z) = (1-\delta)\max_{x_e}\left[\boldsymbol{\mathcal{S}}(z, n_e) - n_e\left(x_e + \frac{wc}{q(\theta(x_e))}\right)\right]^+. \tag{1.64}$$

Having drawn the idiosyncratic productivity z, the potential entrant first decides whether to exit, a decision captured by the notation $\{\cdot\}^+$ and summarized in the dummy d_e. If it stays, the firms acquire a measure of customers, $n_e \in \mathbf{R}^+$, and choose a market x_e in which to search, to maximize the joint surplus minus the total advertisement cost $n_e wc/q(\theta(x_e))$ and the total utility $n_e x_e$ that the firm must deliver to its new customers.

An important feature of this economy is that the submarket in which customers are acquired, x_e, solely appears through the term $wc/q(\theta(x_e))+x_e$, which is an acquisition cost per customer common to both entering and incumbent firms. The first term, $wc/q(\theta(x_e))$, captures the total advertisement cost of acquiring exactly one customer. The second term, x_e, is the utility price that firms offer to their new customers. Firms choose submarkets that minimize the advertisement cost per customer. Define the minimal advertisement cost as:

$$\text{cost} = \min_x \left[x + \frac{wc}{q(\theta(x))}\right]. \tag{1.65}$$

The optimal entry further requires that only the submarkets that minimize this advertisement cost be open in equilibrium, which I summarize

in the following complementarity slackness condition:

$$\forall x, \quad \theta(x)\left[x + \frac{wc}{q(\theta(x))} - \text{cost}\right] = 0. \tag{1.66}$$

This condition means that submarkets wether minimize the advertisement cost, cost $= x + c/q(\theta(x))$, or remain unvisited, $\theta(x) = 0$. In equilibrium, active submarkets will have the same hiring cost, and firms will be indifferent between them. Therefore, the equilibrium market tightness on every active market is:

$$\theta(x) = q^{-1}\left(\frac{wc}{\text{cost} - x}\right). \tag{1.67}$$

Notice that because q is a decreasing function, the equilibrium market tightness decreases with the level of utility promised to the customers, as these offers succeed in attracting more customers, while firms refrain from posting such expensive contracts. The probability of finding a firm for customers thus declines with the attractiveness of the offer.

Prices and the Main Model

Once the real allocation of the economy is solved under the contractual environment specified in Section 1.8.1, building on the results in Schaal (2017), one can solve equations (1.29) and (**??**) to construct the prices (equation (1.30)) that implement the exact same allocation from (1.60).

1.8.2 Capital, Marginal Costs, and Labor Share

Here, I discuss a potentially useful extension that allows the meaningful disjoint analysis of both the labor share and markups in the model. To do so, I augment the model with physical capital. For the sake of exposition, I assume that firms do not own their own capital but borrow it in every period. The firm's problem would then be:

$$
\begin{aligned}
&\boldsymbol{\mathcal{V}}(z, n, \{\mathcal{C}(j)\}_{j\in[0,n]}; w) \\
&\quad = \max_{n_i'(z';w), x_i'(z';w), \{\omega(j)\}_{j\in[0,n]}} \int_0^n p(j)\mathrm{d}j - w\ell - (r+\delta^k)k - wf \\
&\quad + (1-\delta)\beta\mathbb{E}\Big\{ - n_i' \frac{c}{q(\theta(x_i'))} - w\mathcal{K}(n_i'; n) \\
&\quad + \boldsymbol{\mathcal{V}}(z', n', \{\widehat{\mathcal{C}}(z'; w, j)\}_{j\in[0,n']}; w) \Big\}^+,
\end{aligned} \tag{1.68}
$$

subject to:

$$
n'(z'; w) = \int_0^n (1-\tau(z'; w, j))(1 - m(\theta(x'(z'; w, j))))\mathrm{d}j + n_i'(z'; w), \tag{1.69}
$$

$$
\hat{\mathcal{C}}(z'; w, j) = \begin{cases} \mathcal{C}(z'; w, j) & \text{for} \quad j' \in [0, n'(z'; w) - n_i'(z'; w)] \quad \text{and} \quad j' = \Phi(z'; w, j), \\ x_i(z', w) & \text{for} \quad j' \in [n'(z'; w) - n_i'(z'; w), n'(z'; w)], \end{cases} \tag{1.70}
$$

$$
y = e^z k^{\alpha\omega} \ell^{(1-\alpha)\omega}, \tag{1.71}
$$

$$
y = n, \tag{1.72}
$$

where $\Phi(z'; w, j) = \int_0^j (1-\tau)(1 - m(\theta(x'(z'; w, k))))\mathrm{d}k$.

The firm now borrows the capital at a rental rate given by $r + \delta^k$, where r is the interest rate, and δ^k is the depreciation of physical capital. Therefore, the production function (1.71) takes both capital and labor as inputs. Moreover, now the production functions directly distinguish the output elasticity to labor from the returns to scale, which are given by ω. Therefore, the marginal product in this *augmented* economy is given by:

$$
\mathcal{MC} = \left(\frac{1}{\omega}\right)\left(\frac{1}{\alpha}\right)^{\alpha}\left(\frac{1}{e^z}\right)^{\frac{1}{\omega}} n^{\frac{1-\omega}{\omega}} (r+\delta^k)^{\alpha} w^{1-\alpha}. \tag{1.73}
$$

As can be seen from the equation, the marginal cost is still decreasing in the returns to scale; that is, it declines in ω. Therefore, the main

mechanism described in the main text is preserved in this case as well. Moreover, because the output elasticity to labor is now governed by a different parameter relative to the one that governs the returns to scale, one can accommodate both an increase in returns to scale and a decline in the labor share. This is indeed consistent with the empirical evidence presented in Section 1.2.3 and in Chiavari and Goraya (2021). Taking stocks, this model extension should be able to obtain both an increase in markups—as in the main text—and a quantitative-relevant decline in the labor share.

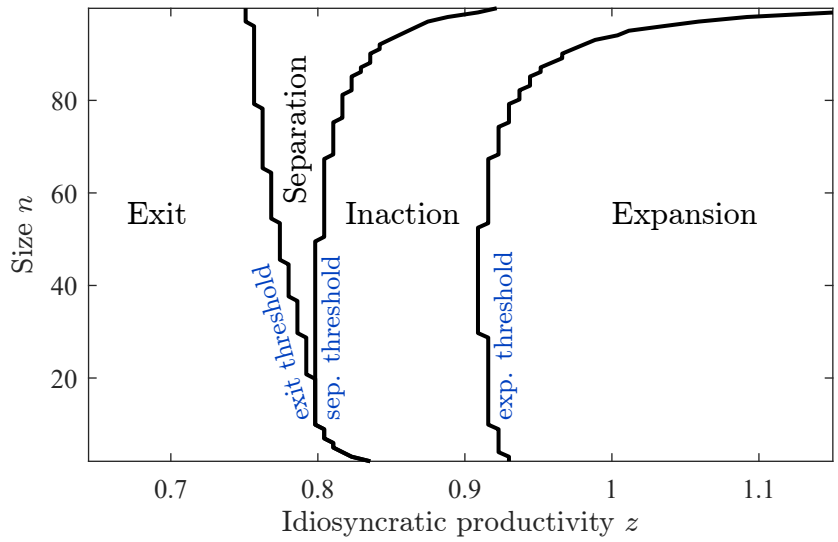
1.8.3 Additional Validation Exercises

Customer Base Policy at the Firm-Level

Firms, in the model, can use various margins—acquisitions, separations, or exit—to adjust employment. I examine here how the decision of firms to use these margins varies as a function of their individual characteristics (z, n) at the beginning of a period.

Figure 1.15 displays the optimal policy of firms as it appears in the baseline calibration. As expected, customer acquisitions take place in small productive firms, whose marginal value of adding customers is high, while separations occur in unproductive firms. Interestingly, because search frictions show up in the surplus (1.60) as a linear advertisement cost, cost $= wc/q(\theta(x_i)) + x_i$, a wedge appears in the adjustment cost faced by firms at $n' = n$. More specifically, separating from a customer earns a value of $\mathcal{U}$ to the customer-firm group, while acquiring new customers incurs the above cost, strictly greater than the value of being an unattached customer in equilibrium. Arising from this kink in adjustment costs, a band of inaction emerges between two thresholds: an expansion threshold, and a separation threshold. Whenever a firm falls in the expansion region, its optimal strategy consists of acquiring new customers until it slowly reaches the expansion threshold—a point at which the marginal value of adding a customer equals the overall cost of acquiring extra customers. Similarly, whenever a firm finds itself in the separation region, its optimal decision is to separate from its customers until it slowly reaches

Figura 1.15: Firms' Action Threshold in the Space (n, z)



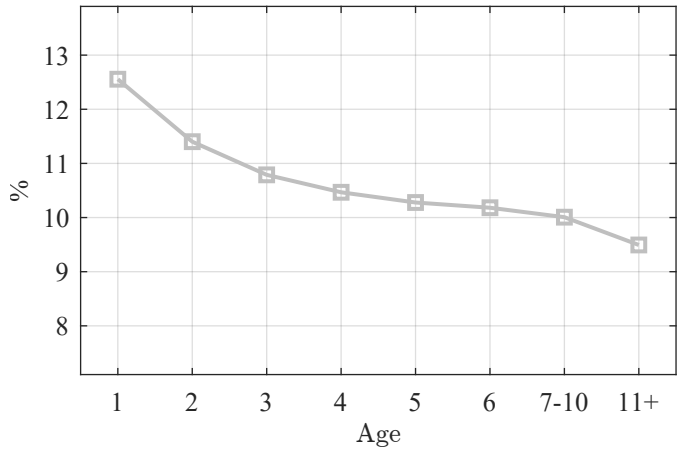
Note. The optimal policies depicted in this figure correspond to the baseline calibration. The areas corresponding to the different margins of adjustment are distinct and do not overlap. Notice that customer acquisitions and separations never occur at the same time because it is more costly for firms to acquire new customers than to retain the current ones.

the separation threshold. There, the marginal value of a customer equals the marginal value of separation. The presence of an inaction region implies the existence of a nonnegligible mass of firms that do not adjust their customer base within a period. Exit takes place in unproductive firms. Indeed, due to the presence of a fixed operating cost wf, the decision to exit mostly affects low productivity and low customer firms, as their current production and expected future surplus fall short of the total operating costs.

Additional Life Cycle and Cross-Section Implications

Another implication of the model is that firms with higher productivity and customers are less likely to exit the market; therefore, older firms are also less likely to exit. This feature of the model can be seen from Figure 1.15, which shows the exit threshold implied by the baseline calibration. Clearly, the exit region, conditional on having low productivity, is wider when firms have fewer customers than when they have many.

Figura 1.16: Exit Rate by Age



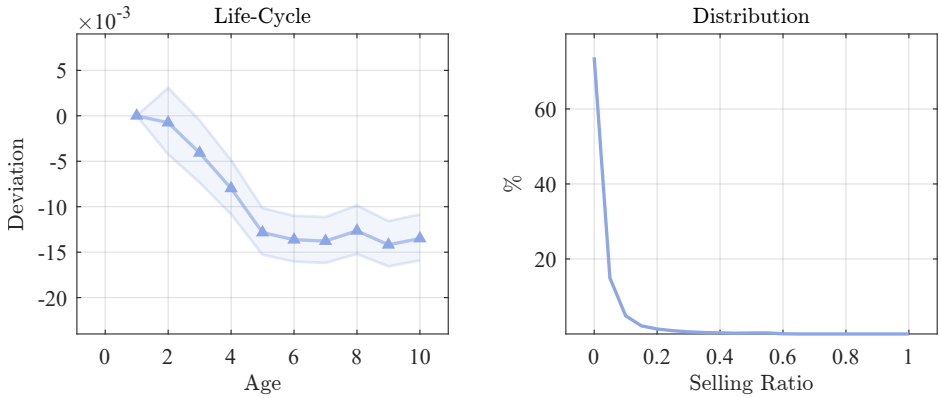
Note. The figure shows the exit rate by age group.

Figure 1.16 shows the exit rate for different age groups. As expected, the model produces a negative correlation between the exit rate and age, meaning that, on average, older firms are less likely to exit the market than younger ones. In the model, this happens because the bigger a firm is, the higher the demand it faces, and hence, the higher its ability to pay its fixed costs. This is an important prediction of the model, as this negative correlation is an empirical finding documented in many empirical papers, such as Haltiwanger et al. (2013).

Robustness on Selling Ratio Implications

In this section, I test the robustness of the patterns documented in Section 1.4.3 regarding the selling ratio. This is particularly important, as already explained in Appendix 1.7.1, because Compustat does not offer any ideal measure of selling-related expenditure at the firm level. Therefore, I review the life cycle and distributional patterns of the selling ratio using the alternative measure of selling expenditure defined in Section 1.7.1.

Figura 1.17: Selling Ratio Robustness



Note. The figure on the left shows the estimated age profile of the selling ratio from equation (1.41) together with the 90% confidence interval. The figure on the right shows the distribution of the selling ratio. The time frame is 1977-1985.

Figure 1.17 shows the results of this robustness exercise. Overall, the main patterns highlighted with the benchmark measure are robust to alternative definitions of selling expenditures. The life-cycle profile of the selling ratio is very similar, aside from the obvious level difference, to the one obtained with the benchmark measure. In the data, firms have a high selling ratio when they are young, which declines with their age.

Moreover, the selling ratio distribution with the alternative measure is very similar to the one obtained with the benchmark measure. In particular, both distributions are right-skewed with a long right tail. Both graphs

show that the model predictions regarding firm-level selling expenditures (relative to production costs) are a robust feature of the microdata, regardless of the measure adopted in the data for this ratio.

Prices and Customers Implications

Prices are one of the main tools that firms have to attract, or retain, customers. In the model, firms that want to grow will charge lower prices to attract and retain customers, and vice versa, firms that are already big will charge higher prices to extract value from their existing customers. Moreover, in the model, firms can discriminate across different customers, as explained in Section 1.3.6. Therefore, the model has two main sources of price dispersion: first, different firms charge different average prices, and second, within the same firm, customers are also charged different prices. Finally, it is worth emphasizing that the model has clear predictions on the customer side as well. Customers, as previously emphasized, will move from firms charging higher prices to firms charging lower prices. Therefore, the model produces an endogenous turnover over customers in equilibrium.

To look at the price dispersion generated by the model, I compare the standard deviation of the price distribution in the model with the one reported by Kaplan and Menzio (2015). This is particularly sensible, as they look at customer-level prices within a very narrow geography and product category, which maps very close to the model setup where output is identical and homogeneous. The model produces a standard deviation of 2.1e-4, which explains approximately 6% of what is observed in the data by Kaplan and Menzio (2015). This is, of course, only a partial success, but should not come as a surprise because it is well known from the work of Hornstein et al. (2011) that this class of models struggles in generating the empirically observed dispersion in prices.

Finally, the model produces an endogenous average customer turnover rate of around 11% a year. This is in the range of the estimates from the previous literature. In particular, Gourio and Rudanko (2014) find a

customer depreciation rate of 0.15.[62] Hence, the model is within the range found by the literature.

Size and Markups

In the data, firms that have bigger sales within a sector tend to have also higher markups; for instance, this has been documented in India by De Loecker et al. (2016). Here, I look at this prediction in the Compustat data and the model. To do so, I run the following regression specification:

$$\log \mu_{it} = \alpha + \beta_1 \log s_{it} + \beta_2 \log s_{it}^2 + \phi_{st} + \varepsilon_{it}, \tag{1.74}$$

where $\log \mu_{it}$ is the log-markup, $\log s_{it}$ is the log-sale, and ϕ_{st} is the sector-time fixed effect. I allow for a quadratic specification to permit a nonlinear relation between the two variables.

The regression estimates a positive relation, both in the model and in the data between the log-sale and log-markups. In particular, in the model, the regression estimates a $\beta_1 = 0.43$ and a $\beta_2 = -0.06$, whereas, in the data, the regression estimates a $\beta_1 = 0.30$ and a $\beta_2 = -0.01$. All coefficients are statistically significant, and the time frame is 1977-1985.

The model's estimates are close to the ones from the data. This is an important result as, in the model, these elasticities have not been a target in the calibration strategy. The model is hence able to replicate moments from the joint distribution of firms' size and markups. This is the case because, in the model, the biggest firms are the most productive, and hence, the ones that face a lower marginal cost of production. In turn, this implies that they are the ones that charge the lowest prices, and hence, are the ones that face a more inelastic demand, which allows them to charge the highest markups.

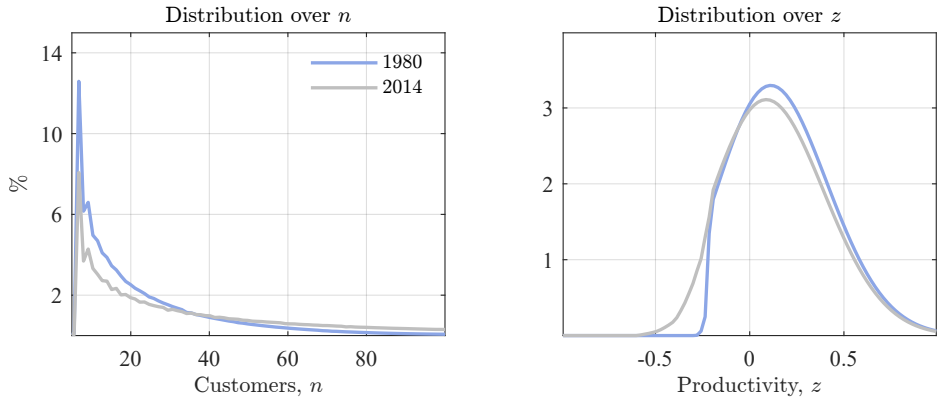
[62] Significant customer inertia has also been documented empirically by Dubé et al. (2010) and Bronnenberg et al. (2012).

1.8.4 Additional Quantitative Exercises

Evolution of the Firms' Distribution

The explored rise in returns to scale has direct implications for the distribution of firms across customers and productivity levels, as explained in Section 1.5.1. In particular, a rise in returns to scale (i) gives rise to some big firms that, exploiting their scale economies, can attract many customers and to a lot of small firms with instead few customers facing the competition of these big firms; (ii) lower the selection process in the economy, implying that more firms with lower productivity are indeed able to operate in the new equilibrium.

Figura 1.18: Firms' Distribution Across Customers and Productivities



Note. The figure on the left shows the distribution of firms across customers, n. The figure on the right shows the distribution of firms across the productivity levels.

Figure 1.18 shows the distribution of firms across customers and productivity levels in the 1980 and 2014 steady states. The figure on the right shows the distribution of firms across customers; it shows an increase in its right-skewness and a fattening of the right tail. This is the outcome of the presence—in equilibrium—of big firms that can exploit their scale economies to attract new customers. The fact that the distribution is more right-skewed speaks directly to the literature emphasizing the rise of

superstar firms (see, for example, Autor et al. (2020)).

The figure on the left shows the distribution of firms across productivity levels; it shows lower selection. In particular, in the new steady state, there is a fattening of the left tail, which is the outcome of the presence of new firms that, exploiting their scale economies, can operate even after adverse productivity shocks.

Aggregate Output and Welfare

In this final section of the Appendix, I look at the implications of the rise in returns to scale for aggregate output and welfare. To study the effect of this technological change in aggregate output and to highlight which factors are behind its changes, I present the following decomposition of its rise over time:

$$\Delta \log Y_t = \Delta \log \mathcal{Z}_t + \Delta \log L_t/m_t + \Delta \log m_t, \tag{1.75}$$

where $Y = \int_i y_i di$ is the aggregate output, $\mathcal{Z} = \int_i (y_i/\ell_i)(\ell_i/L)di$ is the aggregate productivity, L is the total labor, and m is the mass of firms. This decomposition helps us understand when the aggregate output changes because of a change in (i) the aggregate productivity, or (ii) the average firm size, or (iii) the mass of firms in the economy.

Taula 1.7: Evolution of Aggregate Output

	Productivity	Avg. Firm Size	Mass of Firms	Output
	$\log \mathcal{Z}$	$\log L/m$	$\log m$	$\log Y$
$100 \times \Delta_{2014-1980}$	-28.61	45.63	-15.23	1.79

Note. This table shows the percentage change in the aggregate output and its components, as highlighted in equation (1.75) between the 1980 and 2014 steady states. The first column reports the percentage change in the aggregate productivity, the second column reports the percentage change in the average firm size, the third column reports the change in the mass of firms, and the fourth column reports the percentage change in the aggregate output. Notice that columns one to three must sum up to column four by construction.

Table 1.7 reports the results from the decomposition highlighted in equation (1.75). In the model, after a 5% rise in returns to scale, the output increases by almost to 2% relative to the trend.[63] However, this moderate rise in the aggregate output masks sizeable changes in its different components: in particular, I observe a decline in aggregate productivity of approximately 28%, a rise in the average firm size of approximately 45%, and a decline in the mass of firms of approximately 15%.

The decline in labor productivity is the outcome of a rich set of forces. On the one hand, a rise in returns to scale, all else being equal, increases the firm-level average product of labor, y_i/ℓ_i; on the other hand, it weakens the selection process in the economy, allowing less productive firms to operate. Quantitatively this second force dominates and produces the decline in the aggregate productivity documented above.[64] The weakening of the selection process also produces the decline in the mass of firms, as shown in Table 1.7. As explained in Section 1.5.1, with lower selection, entry rates and reallocation rated decline, leading to a steady state with fewer firms.

The rise in the average firm size follows from similar forces as the one outlined above: the returns to scale rise increases the firm size and lowers the selection—that is, it allows smaller and less productive firms to operate. However, in this case, the first effect dominates. Overall, the rise in the average firm size dominates the other two factors, translating into an aggregate output rise.

[63] It is noteworthy to acknowledge that we cannot compare the two steady states, as we should view the model as a detrended version of an underlying framework with balanced growth. Therefore, this 2% output rise is an increase relative to a counterfactual experiment in which the output would have only increased due to balance growth, at a 3% rate, for example.

[64] Interpreting the model outlined in this paper as a detrended version of a model featuring balanced growth, we can think of the decline in the aggregate productivity as the model counterpart to the facts highlighted by Fernald (2015). This proves that the technological change documented in the paper can be consistent with the recent US productivity decline. Of course, one should exercise caution with such an interpretation. The model has not been designed to capture the growth phenomena fully, and thus does not allow for a straightforward mapping with the data in this aspect.

I now turn my attention to the evolution of aggregate welfare. In this economy, aggregate welfare is the representative household utility. Therefore, the change in welfare measured in consumption-equivalent terms is given by:

$$U(C_{1980}(1+\lambda), L_{1980}) = U(C_{2014}, L_{2014}), \tag{1.76}$$

where λ measures how much more (or less) the consumption in percentage terms makes the representative household indifferent between the 1980 and 2014 steady states. Given the specific functional form of the representative household preferences, λ is given by:

$$\lambda = \frac{C_{2014} - \vartheta\big(1+1/\psi\big)^{-1}\big(L_{2014}^{1+1/\psi} - L_{1980}^{1+1/\psi}\big)}{C_{1980}} - 1. \tag{1.77}$$

I find that welfare is approximately 37% below the trend. This decline is due to (i) the lower selection and (ii) higher firm-level selling-related expenditure. Lower selection translates into lower average productivity, which, together with the fact that the average firm becomes bigger, implies that the representative household must supply additional labor to sustain production. Higher firm-level selling-related expenditure, devoted to firm size expansion, is a deadweight loss that must be financed by the representative household with additional labor. These two forces together increase labor, and hence, labor disutility, which being convex, dominates the moderate linear increase in utility from consumption due to the rise in output.

I conclude with a few remarks related to the results on aggregate welfare. First, this does not imply that welfare is lower relative to the 1980s, but only that it is below the trend due to this technological change. To see this, we can compute the level of welfare in 2014, assuming that the economy has grown by 3% a year. In this case, welfare in 2014 would be about 27% higher than in 1980.[65] Second, in the model, consumer welfare, as measured by aggregate consumption and aggregate welfare, moves

[65]To see this, consider that absent any change, assuming a 3% growth, the aggregate welfare increased by 2014, which is given by $\log(C_{1980}(1.03)^{34}) - \log(C_{1980})$. Hen-

in the opposite direction. This fact illustrates the tension associated with the common practice of antitrust authorities of using aggregate consumer welfare as a shortcut for the overall welfare.[66]

ce, the counterfactual level of welfare after the rise in returns to scale is $(1 - 0.37) \times (\log(C_{1980}(1.03)^{34}) - \log(C_{1980}))$.

[66]The inability of the consumer welfare paradigm to fully capture stakeholders' interests has recently been a highly debated topic among antitrust scholars (Hovenkamp (2019, 2020a,b) and Marinescu and Hovenkamp (2019)).

Capítol 2

THE RISE OF INTANGIBLE CAPITAL AND THE MACROECONOMIC IMPLICATIONS

Joint with Sampreet Goraya

2.1 Introduction

In the last decades, technological improvements have reshaped the production process of US firms. Nowadays, investments in research and development, intellectual property products, and computerized information—commonly known as intangible capital—account for more than 30% of aggregate investment. This novel capital shows different characteristics compared to tangible capital, such as equipment and structures. Specifically, it is usually immaterial, specific to the firm that uses it, and often internally produced rather than acquired. The rise of intangible capital, with its unique characteristics, has ramifications for competition, for antitrust policy, for allocative efficiency, and hence for economic well-being

more broadly.

Despite the increasing importance of intangible capital, we know little about its properties and the implied consequences of its rise. In this paper, our goal is to document the properties of intangible capital and to shed light on the macroeconomic implications of its rise as an input in production for US firms. First, exploiting firm-level data, we document the changing nature of production technology. We show that the input share of intangible capital has seen a sizeable increase since the 1980s, going from approximately 0.03 to 0.12. This rise has come at the expense of the labor share in production. We label this phenomenon *intangible capital biased technological change* (IBTC).

Then, we provide novel empirical evidence on the behavior of investment in intangible capital. In the data, we see that the firm-level intangible capital investment process is lumpier compared to tangible capital investment, as it is characterized by long periods of inaction and by a high serial correlation. To rationalize these empirical findings, we use a general equilibrium model of firms and investment dynamics extended with intangible capital. We also allow for a flexible specification of adjustment costs associated with the investment process of both types of capital. The model attributes higher adjustment costs—particularly fixed adjustment costs—to intangible capital investment relative to tangible capital investment. These findings confirm the view that intangible capital investment is inherently different from tangible capital investment. For example, the presence of inherent indivisibilities in implementing the just-in-time (JIT) production process by the US manufacturing sector meant that large investment beforehand were required, and the necessity of retraining workers and restructuring production procedures translated into long setup times.[1]

[1]The JIT production process, pioneered by Japanese manufacturers, gained momentum among US firms in the 1980s. Nakamura et al. (1998) documented that the transfer of JIT requires a substantial effort on the part of U.S. manufacturers because so many different aspects of plant operation are involved, and Fullerton et al. (2003) argued that investment returns from JIT adoption are not immediately observable, due to the long-run nature of its implementation process."

Finally, we use the calibrated model to quantify the effects of IBTC on changes that the US economy experienced in average firm size, industry concentration, and allocative efficiency. A shift in production technology toward intangible capital—which entails high investment adjustment costs—makes it difficult for small firms to survive and operate. This creates a reallocation of economic activity toward larger firms, which eventually increases the size of average firms and industry concentration. Moreover, intangible capital tends to be more misallocated as its marginal product is more dispersed across firms relative to other inputs. Therefore, as intangible capital becomes more important as an input of production relative to physical capital and labor, the overall allocative efficiency (as measured by the dispersion in total factor productivity revenue, TFPR) in the economy increases as well. Finally, these findings suggest that a significant fraction of these transformations can be an outcome of the efficient response of the economy to changes in firm-level production technology.

In this paper, we first estimate an augmented firm-level production function with three inputs: tangible capital, intangible capital, and labor. Here, we follow two approaches. The first is the control function approach, as in Ackerberg et al. (2015) and as recently used in De Loecker et al. (2020). Second, we use the cost shares approach adopted by Foster et al. (2008). To do so, however, we need a firm-level measure of intangible capital, which is notoriously difficult to construct with commonly available firm-level datasets as the US Generally Accepted Accounting Principles (US GAAP hereafter) fails to fully account for intangible capital on firms' balance sheets.[2] To this end, we leverage the Compustat dataset, which encompasses all US publicly traded firms. Our baseline measure of intangible capital, between 1980 and 2015, is made up of two components: (i) internally generated intangible capital, through research and development expenditure; and (ii) identifiable intangible capital booked on the balance sheet. Finally, to validate this measure, we compare it with the one pro-

[2]See Lev and Gu (2016), and Ewens et al. (2019) for more discussion about measurement problems while using firm-level data. Corrado et al. (2009), Corrado and Hulten (2010), McGrattan and Prescott (2010b), McGrattan and Prescott (2014), and Koh et al. (2020) highlight the measurement issues while using aggregate data.

vided by Koh et al. (2020) and find similar trends and magnitudes among the two measures.

Using this measure, our production function estimations find that intangible capital is an important factor in production: its input share increased from 0.03 in 1980 to 0.12 in 2015. Moreover, most of this rise happened at the expense of the labor input in production. This finding is robust to different estimation techniques, production function specifications, and levels of disaggregation. We interpret this finding as a technological transformation that US firms are experiencing, where intangible capital is becoming a more prominent input in the production process at the cost of labor. We have labeled this phenomenon IBTC.

After assessing the role of intangible capital in production, we study the properties of the investment process of intangible capital. We focus on the technological frictions associated with the investment process, which has been emphasized in Haskel and Westlake (2018). To do so, we build a model of firms and investment dynamics in the spirit of Hopenhayn (1992) and Clementi and Palazzo (2016b). In the model, firms behave competitively and produce a single good using a Cobb-Douglas production function with tangible capital, intangible capital, and labor as inputs. Moreover, the model features entry and exit of firms and a flexible structure of investment adjustment costs for both types of capital. In particular, the adjustment costs associated with both types of capital have two components: (i) convex cost disciplining the intensive margin of investment, and (ii) a fixed cost disciplining the extensive margin of investment.

The predictions of the model related to the investment process of both tangible and intangible capital depend on the precise identification of the two sets of parameters that discipline the convex and fixed costs associated with the investment processes. Following the seminal papers of Cooper and Haltiwanger (2006) and Asker et al. (2014), we use inaction rates, defined as investment between $\pm 1\%$, to identify the fixed costs of adjusting each type of capital. This moment is informative because higher fixed costs of adjustment increase the inaction rate, as firms prefer not to invest instead of paying these costs. Then, we use the autocorrelation of the investment rate process to discipline the convex costs associated with

both types of capital. High convex costs make the investment process serially correlated, forcing firms to accumulate capital slowly.[3]

The calibrated model finds substantial differences in the investment processes of these two types of capital, with intangible capital having higher adjustment costs compared to tangible capital. In the model, the high adjustment costs associated with intangible capital make this input slower to adjust relative to tangible capital when productivity shocks occur. To validate this prediction, we estimate the elasticity of the average revenue product of both types of capital to productivity shocks. We find that this elasticity is higher for intangible capital, making the average revenue product of intangible capital, $ARPK_I$, more responsive relative to the average revenue product of tangible capital, $ARPK_T$. Moreover, we also document in the data that, consistent with the model predictions, the $ARPK_I$ is more volatile at the firm level and more dispersed in the cross section relative to the $ARPK_T$ in most of the sectors.

As a final validation exercise, we exploit cross-sector variation in intangible intensity to provide reduced-form support for the model mechanism. In particular, we look at the intangible investment share, the tangible investment share, the labor share, the profit rate, industry concentration, and allocative efficiency, which are the main focus of analysis when we study the consequences of IBTC. Overall, we find that the model captures the qualitative features of the data implying the direction of cross-sectoral correlations in line with the ones found in the data.

Finally, we use the calibrated model to quantify the macroeconomic effects of IBTC. In particular, we ask what are the effects of increasing the intangible capital share and decreasing the labor share in the firm-level production technology while keeping fixed the returns to scale? The changes in the returns to scale in production are of interest on their own, as studied by Lashkari et al. (2021) and Chiavari (2021); however, much less is known about the role of the changing composition of inputs in

[3]It is important to notice that these parameters are jointly calibrated. Moreover, the presence of counterbalancing forces—that is, high fixed costs decrease the autocorrelation of the investment rates, whereas, high convex costs increase it—makes it crucial for the correct identification of these two costs.

production, and this motivates our focus. Furthermore, many other technological changes have emerged over this period, and we abstract from them in the benchmark analysis. However, in our robustness analysis, we allow for some of these changes in our model, such as the decline in the relative price of intangible investments, and we disentangle the effects of IBTC.

The IBTC can quantitatively explain most of the increase in average firms' size and industry concentration, as measured by the Herfindahl-Hirschman Index and by the employment share of firms with more than 250 employees, as observed in the data. This happens because, while intangible capital becomes more important in production, firms rely more on an input that entails higher adjustment costs, and as a result, the value of entry decreases, pushing up the threshold productivity of the marginal entrant. This implies that a relatively small ? but more productive ? mass of firms operate in the economy. Thus, the average incumbent size increases. Furthermore, IBTC makes the growth of small firms costly, as they have to incur very high adjustment costs to build their stock of intangible capital, and it makes it easier for large firms to shrink, as the high depreciation rate of intangible capital favors its depletion. This mechanism, together with the above increase in selection, tilts a reallocation of sale shares toward the larger firms, leading to the rise in average firm size and industry concentration.

Further, through the lens of the model, we see that IBTC also accounts for most of the changes in the aggregate factor shares that have been emphasized in the literature. It generates the increase in the intangible capital share and the decline in both tangible capital and labor shares. This happens because micro-level technological change also affects the aggregate demand for each of the three inputs, favoring intangible capital in particular. Moreover, consistent with the findings in Koh et al. (2020), we find that the labor share would decline much less if intangibles were expensed instead of capitalized.[4] Finally, as the selection process increases and

[4]Barro (2019) and Atkeson (2020) also show that part of the decline in the corporate non-financial labor share is a result of the accounting procedures used by the Bureau of Economic Analysis of the US Department of Commerce.

only the more productive firms operate in the market, we see an increase in the firm-level profit rate of a magnitude consistent with that has been emphasized in De Loecker et al. (2020) and Barkai (2016).

Moreover, we find that IBTC can explain half of the decline in the tangible capital investment rate, as documented by Hall (2015), Gutiérrez and Philippon (2016), and Crouzet and Eberly (2019). This occurs because firms with this new intangible-intensive technology tilt a greater share of their expenditure toward intangible capital. As a consequence, investment in tangible capital declines in the new steady state. Hence, the model interprets half of the slack in the investment rate in tangible capital as the by-product of a technological change that makes tangible capital a less relevant input in production.

Finally, the quantitative model shows that IBTC can explain between 32% and 80% of the overall decline in the allocative efficiency of the US economy, as documented by Bils et al. (2020). This is driven by the fact that $TFPR$ in our framework is a weighted geometric mean of the average revenue product of inputs, where the weights are proportional to their output elasticities. The presence of adjustment costs means that dispersion in $TFPR$ is driven by dispersion in the average products of both types of capital. When the output elasticity of intangible capital increases, the dispersion in $ARPK_I$ becomes the primary driver of the dispersion in $TFPR$.[5] Therefore, dispersion in $TFPR$, which is our measure of allocative efficiency (where higher dispersion in $TFPR$ means lower allocative efficiency), increases. However, in our framework dispersion in $TFPR$—as already noted by Asker et al. (2014)—cannot be interpreted as misallocation, as in Hsieh and Klenow (2009), because the allocation still coincides with the planner's one.

Related Literature. This paper is related to the rising literature that measures intangible capital at the aggregate level, as in Atkeson and Kehoe (2005), Corrado et al. (2009), Corrado and Hulten (2010), McGrattan

[5] Adjustment costs associated with the investment process of an input do not allow its marginal product to equalize across firms, and hence generate dispersion in the average revenue product.

and Prescott (2010a), McGrattan and Prescott (2010b), McGrattan and Prescott (2014), Koh et al. (2020), and Atkeson (2020), and at the firm-level, as in Peters and Taylor (2017) and Ewens et al. (2019).[6] Relative to them, we structurally estimate a Cobb-Douglas firm-level production function augmented with intangible capital. We document that intangible capital is an important input in production and is rising over time at the expense of labor.

Furthermore, our paper is related to the extensive literature that examines lumpy investment dynamics, as pioneered by Abel and Eberly (1994), Abel and Eberly (1996), Doms and Dunne (1998), and Cooper and Haltiwanger (2006), highlighting the role of non-convex adjustment costs in the firm-level investment process. To the best of our knowledge, we are the first to highlight the presence of higher adjustment costs associated with the investment process of intangible capital relative to tangible capital.

This work is related to some recent papers by Brynjolfsson et al. (2021), De Loecker et al. (2021), and Kaplan and Zoch (2020). Our analysis on the rising importance of intangible capital and its associated measurement challenges is close to the view of Brynjolfsson et al. (2021). The rising importance of adjustment costs associated with intangible capital is in line with the rise in overhead costs, as documented in De Loecker et al. (2021). In line with our findings, Kaplan and Zoch (2020) highlights the rising expenditure in intangible investment.

Finally, the paper relates to Lashkari et al. (2021), Aghion et al. (2019), Hsieh and Rossi-Hansberg (2019), and Chiavari (2021) which present different mechanisms, all associated with technological factors, behind some of the macroeconomic trends emphasized in this paper. Moreover, De Ridder (2019), Zhang (2019a), and Caggese and Perez-Orive (2020) emphasize the role of intangible capital as a driving factor behind some recent trends. Relative to them, we use firm-level data to inform our mo-

[6]Moreover, our paper is also related to the literature on innovation and firm dynamics that documents the behavior of firms' R&D investment and related measurement issues (see, Cohen and Klepper (1992), Grilliches (1995), Klette and Johansen (2000), Klette and Kortum (2004), and Cohen (2010)).

del about the production process and the properties associated with this new capital. We find that intangible capital is a dynamic input in production whose importance is rising and that its investment process is highly distorted by technological frictions such as adjustment costs. Combining these novel insights with a quantitative model, we are the first, to the best of our knowledge, to jointly explain the rise in average firm size and concentration, the changes in aggregate factor shares, the decline in the tangible investment rate, and the decline in allocative efficiency in the US economy between 1980 and 2015.

Outline. Section 2.2 briefly discusses the data and shows the construction of our main variables. Section 2.3 documents the stylized facts. Section 2.4 presents our quantitative framework. Section 2.5 contains the calibration of the model and its external validation, and Section 2.6 oresents a discussion of the main mechanisms behind our results, Section 2.7 presents the main results and discusses the implications of IBTC for the US economy. Section 2.8 concludes.

2.2 Data and Measurement

In this section, we present the main dataset used throughout the analysis. We explain (i) the construction of the variables, unrelated to intangible capital, used for the empirical analysis and (ii) the measurement of firm-level intangible capital, emphasizing the main challenges, its virtues, and its drawbacks.

2.2.1 Main Measures

The main data source is Compustat, a firm-level database with all the US publicly traded firms between 1980 and 2015. In this section, we briefly discuss the strengths and limitations of this dataset. We provide more details on the data cleaning process and the construction of the sample of analysis in Appendix 2.9.1.

The choice of the data source is driven solely by its ability to cover the

period of interest and the largest number of sectors. These characteristics make these data an excellent source of firm-level information to study technological changes in production undertaken by US firms.

Although publicly traded firms are few relative to the total number of firms, as they tend to be the largest firms in the economy, they account for roughly 30% of US employment (see Davis et al. (2006)). The Compustat data contain information on firm-level financial statements including measures of sales, input expenditures, and capital stock information, as well as a detailed industry activity classification.

Despite the many virtues of these data, however, they present two main limitations: (i) the fact that it is impossible to distinguish quantity and prices, which makes the measurement of the production function elasticities significantly more challenging, as extensively explained in the next section;[7] (ii) the possible selection issues arising from using only publicly traded firms. To address the first concern, we follow the methodologies explained in Appendix 2.9.2. Moreover, whenever possible, we compare our results with additional data sources to isolate the potential bias of using only publicly traded firms.

As a measure of firm-level production, we use firms' sales (`SALE`); as a measure of variable inputs used in production, we use cost of goods sold (`COGS`); as a measure of firm-level employees, we use (`EMP`); and as a measure of tangible capital, we use gross capital (`PPEGT`). Summary statistics related to these variables are reported in Appendix 2.9.1.

2.2.2 Intangible Capital Measurement

The firm-level measurement of intangible capital is a challenging task as a substantial portion of it is internally generated rather than being externally acquired, and US GAAP does not allow its capitalization on the balance sheet (see Lev and Gu (2016), and Ewens et al. (2019)). As a consequence, following the accounting standards in force, nearly all of the internally generated intangible capital is recorded differently from tangible capital in the accounting books. In particular, all tangible investment is recorded

[7] This challenge is present in most of the production data.

on the balance sheet at its purchased price and then depreciated over its useful life; however, internally produced intangible investment, such as R&D, advertisement, or employee training, is fully expensed in the current period and hence appears in the firms' income statement but not on the balance sheet. Only externally acquired intangible capital is directly booked on the balance sheet. For a more in-depth discussion about accounting standards and related challenges to firm-level intangible capital measurement, see Appendix 2.9.1.

In light of these considerations, our main measure in the paper is formed by two different components: (i) internally generated intangible capital and (ii) externally acquired intangible capital. Internally generated intangible capital in our case is obtained through the capitalization of R&D expenditure (`XRD`). We do not include organizational capital in our benchmark measure as this is normally constructed through the capitalization of a sector-dependent share of selling, general, and administrative expenses (`XSGA`).[8] This item includes many expenditures that are not inherently related to intangible capital, such as CEO wages, rents for buildings, and capital adjustment costs, among others.[9] Capitalizing such a big expenditure item would heavily downward bias our estimates of the inaction rate as this expenditure item is never zero, and even in periods of no investment in intangible capital, we would be capturing some unrelated overhead cost. Moreover, using organizational capital, we would be capitalizing a part of incurred adjustment costs, and hence we would artificially inflate our measure of intangible capital, creating conceptual issues in the estimation of the production function. Finally, the imputation of a constant fraction across firms of SG&A as intangible investments would substantially increase the concerns related to potential firm-level measurement error.

Therefore, we use the perpetual inventory method on R&D expendi-

[8]The organizational capital is used in Eisfeldt and Papanikolaou (2013), Peters and Taylor (2017), and Ewens et al. (2019).

[9]While working with Compustat data, it is often assumed that the capital adjustment costs are expensed in `XSGA`).

ture to recover a firm-level measure of knowledge capital given by

$$k_{R\&D,ft} = (1 - \delta_s)k_{R\&D,ft-1} + \texttt{XRD}_{ft}, \tag{2.1}$$

where XRD is gross investment in knowledge capital deflated by the IPP price deflator, the sector-level depreciation rate δ_s is taken from Ewens et al. (2019), and the initial stock is assumed to be zero.[10]

The second component of intangible capital is the externally acquired intangible capital, which is capitalized on the balance sheet at fair value under the variable INTAN in Compustat, according to the US GAAP under the guidelines provided in ASC 350. However, INTAN is *net* intangible capital, and to get the gross measure, to be consistent with the measurement of both tangible capital and internally produce intangible capital, we use INTAN + AM, where AM is the amortization of balance sheet intangible capital. Finally, because of measurement issues explained extensively in Appendix 2.9.1, we drop goodwill from our measure of gross balance sheet intangible capital. Hence, our final measure of balance sheet intangible capital is

$$k_{BS,ft} = \texttt{INTAN}_{ft} + \texttt{AM}_{ft} - \texttt{GDWL}_{ft}, \tag{2.2}$$

where all variables have been appropriately deflated with the IPP deflator.

Our final measure of firm-level intangible capital is given by the sum of internally produced and externally purchased intangible capital:

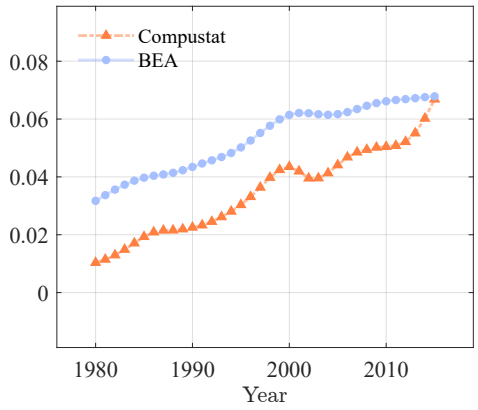
$$k_{I,ft} = k_{R\&D,ft} + k_{BS,ft}. \tag{2.3}$$

Figure 2.1 compares our total intangible capital investment share with the one reported by the Bureau of Economic Analysis (BEA) corporate nonfinancial sector, and as documented by Koh et al. (2020).[11] We focus on

[10] For all our analysis, unless differently stated, we exclude all observations in the first five years to avoid a strong dependence of our results from our assumption on the initial condition for knowledge capital. Results are not sensitive to this exclusion. Moreover, results are similar if we use a different level of initial capital—for instance, investment in the first period divided by its depreciation rate.

[11] Intangible capital investment is the sum of internal investment in knowledge capital and investment in balance sheet capital. To calculate the gross balance sheet capital investment, we assume a depreciation rate of 0.20, as is mostly done in the literature, as there are no reliable estimates for this depreciation.

Figura 2.1: Aggregate Intangible Investment Share: Compustat vs BEA



Note. The figure reports the evolution of the intangible investment share. Intangible investment share in Compustat (dashed orange line with triangles) is computed as the sum of total investment in intangible capital to the sum of total sales in a given year. Intangible investment share from the BEA corporate non-financial sector (solid light blue line with circles) is computed as the investment in intangible capital to GDP net of propriety income, taxes, and subsidies as computed by Koh et al. (2020). The data are de-trended with an HP filter with $\lambda = 6.25$.

the corporate non-financial sector as it is the most closely comparable to our Compustat dataset, given that we exclude financial firms, as explained in Appendix 2.9.1. Overall, both data sources show a similar qualitative increase over the sample period. In Appendix 2.9.1, we show additional comparisons between our firm-level measure and aggregate measures from the national accounting measured at different levels of disaggregation. In sum, we find that our firm-level measure performs reasonably well compared to national accounting data despite the data pitfalls and accounting limitations.

2.3 Empirical Analysis

This section presents the main empirical results of the paper. First, we show that the intangible capital share in production experienced a sizeable increase over the last decades and that this increase happened mostly

at the expense of the labor input share. Second, we document that the investment rate distribution of intangible capital is qualitatively different from the investment rate distribution of tangible capital, suggesting a different underlying investment process.

2.3.1 Fact 1: Fourfold Increase in Intangible Capital Share since 1980

In this section, we investigate the importance of intangible capital as a new factor of production; to do so, we estimate a production function with three inputs: tangible capital, intangible capital, and labor. Our estimates show that intangible capital is an important factor of production and that its importance has had a fourfold increase since 1980.

Production Function Estimation

We estimate the log of a firm-level Cobb-Douglas production function given by

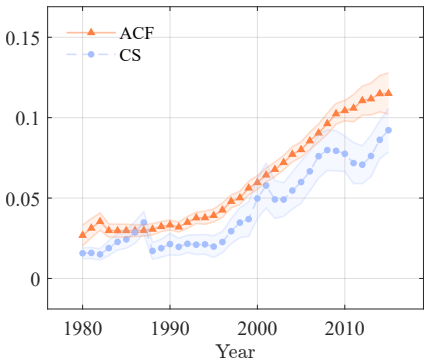
$$q_{ft} = \alpha k_{T,ft} + \nu k_{I,ft} + (1 - \alpha - \nu)\ell_{ft} + \omega_{ft} + \varepsilon_{ft}, \tag{2.4}$$

where q_{ft} is the log of output, $k_{T,ft}$ is the log of tangible capital, $k_{I,ft}$ is the log of intangible capital, ℓ_{ft} is the log of labor, ω_{ft} is the log of productivity, and ε_{ft} is the error term.[12] Given that the objective of our analysis is to estimate the variation in input shares over time, we constrain the firm-level returns to scale to 1, and assume that all the firms share a common technology. In the next section, we show that these assumptions are inconsequential for our results. Estimating firm-level production functions is notoriously difficult as firm-level productivity ω_{ft} is unobservable to the econometrician but is known to the firm at the moment of choosing its inputs. To address this endogeneity problem, we rely on two different estimation procedures proposed by the empirical industrial organization

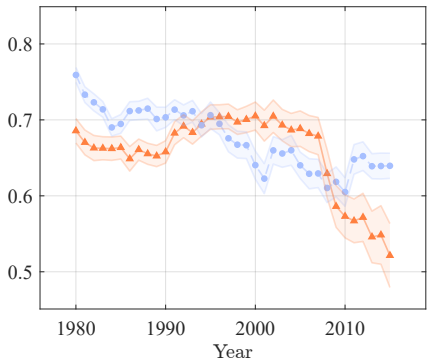
[12]Practically, as output we use the firm's sales, as tangible capital we use gross property, plant, and equipment; as intangible capital we use the measure constructed in Section 2.2.2; and as labor we use the total firm-level number of employees.

literature. In particular, we use the cost shares approach (CS), as in Foster et al. (2008), and the Ackerberg-Caves-Frazer (ACF) approach from Ackerberg et al. (2015). We provide details regarding both methodologies in Appendix 2.9.2.

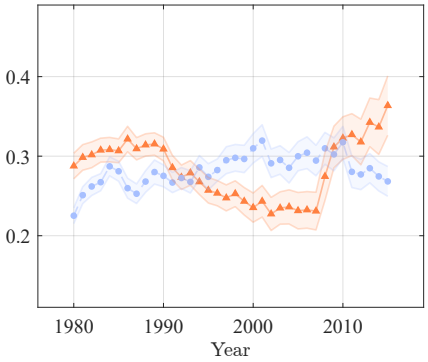
Figura 2.2: Trends in Input Shares



(a) Intangible Capital Share



(b) Labor Share



(c) Tangible Capital Input Share

Note. The figure panels present the output elasticities estimated with the cost shares (CS) approach (dashed light blue lines with circles) and with the Ackerberg-Caves-Frazer (ACF) approach (solid orange lines with triangles). The elasticities are estimating using 10-year rolling windows over time. Bands around the point estimates report the 99% confidence intervals.

To document the changes in the output elasticities of labor, intangible capital, and tangible capital in the production function, we estimate

equation (2.4) with both methodologies between 1980 and 2015 using 10-year rolling windows. Figure 2.2 presents the results. Solid orange lines with triangles report the estimates from ACF with associated 99% confidence intervals, and dashed light blue lines with circles report the estimates from CS with associated 99% confidence intervals. Regardless of the methodology chosen, all the action in the inputs share comes from intangible capital and labor, as tangible capital does not show any obvious trend over the period.

In particular, the intangible capital share with the CS approach goes from 0.016 in 1980 to 0.092 in 2015, whereas with the ACF approach it goes from 0.027 to 0.115. With the ACF approach—our preferred methodology—the intangible capital input share in 2015 is approximately four times as much as it was in 1980. The CS approach, however, estimates approximately a five-fold increase in the intangible capital input share over the same period. It is evident from these results that the Compustat firms, which represent a sizeable part of the US economy, have undergone a significant transformation in their production technology. We label this finding IBTC.

Moreover, the labor share with the CS approach goes from 0.759 to 0.639, whereas with the ACF approach it goes from 0.686 to 0.521. Therefore, we highlight that our estimates suggest a certain level of substitution between intangible capital and labor over time: while the intangible capital share has increased, the labor share has declined in the last decades. This finding is in line with the results from the literature—for instance Elsby et al. (2013), Karabarbounis and Neiman (2013), and Koh et al. (2020), among others.

Given the results documented in this section, in the subsequent part of the paper, we interpret the rise in intangible capital as an exogenous technological change in the production technology biased toward intangible capital at the expense of the labor input.

Robustness

Here we document the extent of the robustness of our results relaxing most of the assumptions imposed on the benchmark specification. In particular, we look at the following specifications: (i) we re-estimate the production function in equation (2.4) leaving returns to scale unconstrained; (ii) we estimate equation (2.4) at two digit sector-level (NAICS 2), effectively allowing for sector-specific technology; (iii) we estimate a translog production function with constant returns to scale.

Figure 2.3 shows the results from the alternative specifications. Appendix 2.9.3 explains the various specifications in detail. Overall, IBTC does not seem to be driven by the specific methodology applied and follows close patterns across the different specifications. The bottom line is that the findings from the benchmark specification are robust.

2.3.2 Fact 2: Intangible Capital More Lumpy than Tangible Capital

Bearing in mind that intangible capital is an important factor in production and that its importance is growing over time, in this section, we document the salient differences in the investment behavior of firms between tangible capital and intangible capital. The investment rate of each type of capital is defined as

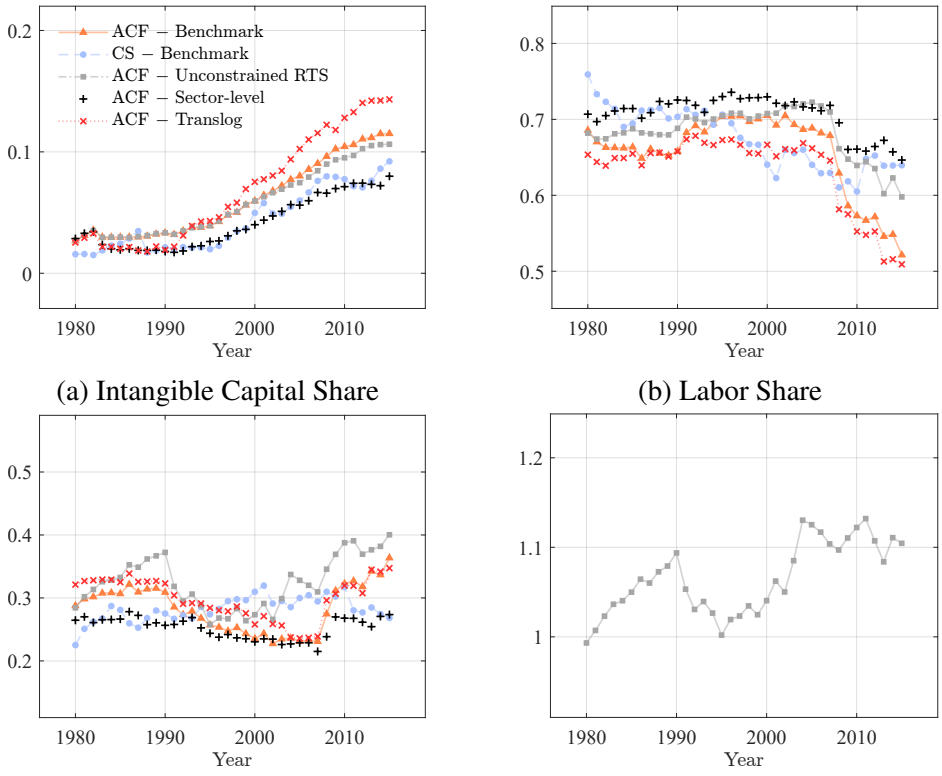
$$\frac{x_{j,ft}}{k_{j,ft-1}} \equiv \frac{k_{j,ft} - k_{j,ft-1}}{k_{j,ft-1}} + \delta_j, \quad j \in \{T, I\}, \tag{2.5}$$

where δ_j is the depreciation rate, $x_{j,ft}$ is investment, and $k_{j,ft}$ is capital.[13] Following Cooper and Haltiwanger (2006) and Clementi and Palazzo (2019), we construct a balanced panel of firms from 1980 to 1990 to study the properties of investment rates.[14] Following common practice, we also drop observations where the total value of acquisitions relative to

[13]The depreciation rate of tangible capital is 7%, whereas, the depreciation rate of intangible capital is firm dependent, as explained in Section 2.2.2.

[14]This is done to control for selection dynamics arising from entry and exit in the data.

Figura 2.3: Trends in Input Shares: Robustness



(a) Intangible Capital Share

(b) Labor Share

(c) Tangible Capital Input Share

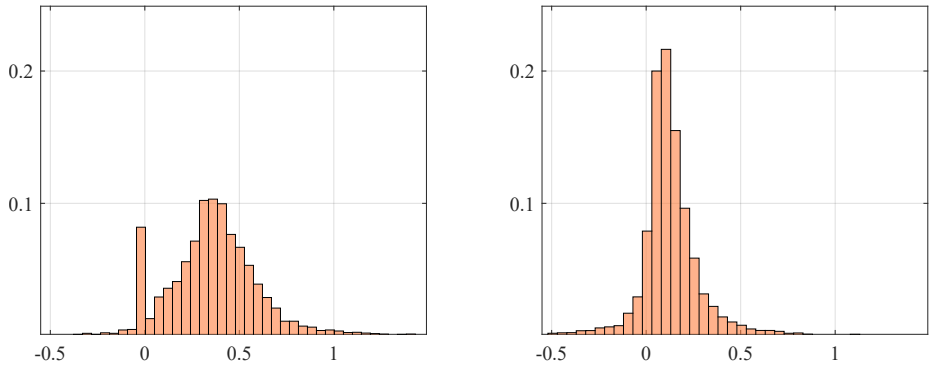
(d) Returns to Scale

Note. The figure panels present the output elasticities estimated with the cost shares (CS) approach (dashed light blue lines with circles), with the Ackerberg-Caves-Frazer (ACF) approach (solid orange lines with triangles), with the ACF approach and unconstrained returns to scale (dashed-dotted light gray lines with squares), with the sector-level ACF approach (black plus signs), and with the Translog ACF approach (dotted red lines with crosses). The elasticities are estimating using a ten-year rolling windows over time.

total assets exceeds 5%.[15] Finally, we drop those firms that have never invested in intangible capital, to prevent the overestimation of the inaction rate of intangible capital investment.

[15]This is done to avoid biases from acquisitions; that is, given the accounting standards, an acquisition would show up as a big investment for one firm but would not show up at all as a big disinvestment for the other. However, we notice that in our balanced panel, these observations represent a small share of all entries.

Figura 2.4: Investment Rate Distributions



(a) Intangible Capital (b) Tangible Capital

Note. The figure panels report the investment rate distributions of intangible and tangible capital for a balanced panel of firms between the years 1980 and 1990. Figure 2.4a shows the investment rate distribution for intangible capital. Figure 2.4b shows the investment rate distribution for tangible capital. The histograms are constructed dropping from the balanced panel all the firms that never invest in intangible capital and all the observations with investment rates above 2 or below -0.5. Results are robust to other winsorization schemes.

Figure 2.4a and Figure 2.4b plot the investment rate distribution for intangible and tangible capital, respectively. These distributions present two stark differences: first, the investment rate distribution for intangible capital presents a clear bimodality, with a lot of mass at the mean and around zero. Meanwhile, the investment rate distribution for tangible capital is almost symmetric around the mean and closely mimics the findings of Clementi and Palazzo (2019). Second, the investment rate distribution for intangible capital shows a small amount of negative investments.[16]

We summarize the main moments of the investment rate distributions in Table 2.1. First, we notice that the average investment rate is much higher for intangible capital compared to tangible capital. This partly reflects a high depreciation rate for intangible capital that pushes the level of optimal investment above that of tangible capital. Second, as anticipated above, intangible capital has a much higher inaction rate, defined as

[16]Notice that this is not by construction; that is, it is not entirely due to the capitalization of an expenditure voice such as R&D, since our measure of intangible capital indeed contains balance sheet intangibles, which, given the depreciation, allows for negatives.

the fraction of investment below 1% in absolute value; particularly, the intangible capital inaction rate is 8% compared to 3% for tangible capital. This high inactivity in intangible capital suggests some underlying non-convexity in the investment process. Third, intangible capital seems to be more serially correlated over time. The autocorrelation in intangible investment is 0.31, much higher than the 0.11 exhibited by tangible capital. This suggests that, conditional on investing in intangible capital, the investment activity goes on for longer, hinting toward some slow adjustment process in the background.

In Appendix 2.9.4, we show that the investment rate distribution exhibits the same properties across sectors, suggesting that the results are not driven by sectoral heterogeneity. Moreover, we also show that the investment rate distribution does not change over time, ruling out concerns related to potential time trends as underlying factors of the documented bimodality. Overall, we can say that the intangible capital investment process is robustly lumpy; that is, it entails long periods of inaction followed by booms of investment activity.

These findings are particularly informative on how intangible capital should be modeled as this capital appears to be neither a fixed cost nor a flexible input. Therefore, from here onward, we will think about intangible capital as a dynamic input in production that could in principle be subject to some adjustment frictions, which will be quantified in the quantitative section of the paper.

2.4 Theoretical Model

To connect the stylized facts from the previous sections, we introduce a quantitative general equilibrium model of investment dynamics with tangible and intangible capital, a rich and flexible structure of investment adjustment costs, and endogenous firm entry and exit.

Taula 2.1: Lumpiness

Investment rates	Intangible	Tangible
Average	0.35	0.13
Positive fraction, $i > 1$	0.89	0.87
Negative fraction, $i < -1$	0.03	0.10
Inaction rate	0.08	0.03
Spike rate, $\|i\| > 20$	0.75	0.25
Positive spikes, $i > 20$	0.73	0.22
Negative spikes, $i < -20$	0.02	0.03
Standard deviation	0.30	0.22
Serial correlation, Corr(i_t, i_{t-1})	0.31	0.11

Note. This table shows the moments of the investment rate distribution of intangible and tangible capital. The statistics are computed for a balanced panel of 5,687 firm-year observations between 1980 and 1990.

2.4.1 Environment

The model follows the spirit of Clementi and Palazzo (2016b). Time is discrete and indexed by $t = 1, 2, \ldots$. At time t, a positive mass of price-taking firms produce a homogeneous good by means of the production function $y = e^z \left(k_T^{\alpha} k_I^{\nu} \ell^{(1-\alpha-\nu)}\right)^{\omega}$, with α, ω, ν in $(0, 1)$, where k_T denotes tangible capital, k_I is intangible capital, ℓ is labor, and z is idiosyncratic random productivity. Idiosyncratic productivity z is driven by the stochastic process

$$z' = \rho_z z + \sigma_z \varepsilon',$$

where $\varepsilon \sim \mathcal{N}(0, 1)$. The conditional distribution of z will be denoted by $\Gamma(z'|z)$.

Firms discount future profits by means of the time-invariant discount factor $\frac{1}{R}$, $R > 1$. Tangible capital depreciates at a rate $\delta_T \in (0, 1)$, whereas intangible capital depreciates at a rate $\delta_I \in (0, 1)$. Adjusting tangible capital stock by x_T and intangible capital stock by x_I bears the cost

$$\mathcal{C}(x_T, x_I; k_T, k_I) = \frac{\gamma_T}{2}\left(\frac{x_T}{k_T}\right)^2 k_T + \frac{\gamma_I}{2}\left(\frac{x_I}{k_I}\right)^2 k_I + \mathbf{1}\{x_T \neq 0\} f_T k_T + \mathbf{1}\{x_I \neq 0\} f_I k_I,$$

where $\gamma_T, \gamma_I, f_T, f_I \in \mathbf{R}^+$. We allow for two different kinds of adjustment costs: convex and fixed. We do not allow for irreversibilities in investment in the baseline version of the model. Generally, these non-convex costs of adjustment are intended to capture indivisibilities in capital, increasing returns in the installation of new capital, and increasing returns to retraining workers and restructuring production activity. Moreover, this formulation of non-convex adjustment costs can be interpreted as a mild form of irreversibility, as disinvestment bears a cost in terms of output, which stems form the potential specific nature of capital. Specifically, if capital is tailored to some particular needs of a firm, it can in principle be difficult to resell it.[17] The convex costs capture overtime

[17] The idea that intangible capital is specific to the needs of the firm that uses it has been suggested by Haskel and Westlake (2018) and Edmond et al. (2018). However, we here move a step forward compared to those papers and test this hypothesis concretely, specifying a flexible model to be tested in the data. In principle, our model could reject this, estimating f_I to be close to zero.

costs, inventory costs, and machine setup costs. Furthermore, we assume that the capital adjustment costs are proportional to their respective capital stock; this is a common specification that is used to take care of the size effect. Finally, we assume that adjustment costs are paid in terms of final output.

We assume that the demand for a firm's output and the supply of both types of capital are infinitely elastic, and we normalize their prices to 1.[18] The supply of labor is given by $L(W) = W^{\psi}$, where $\psi > 0$ and $W \in \mathbf{R}^{+}$ is the real wage.[19]

Each period, operating firms incur a fixed cost $c_f > 0$; this cost is usually interpreted as a per-period expense that firms must incur to operate—for instance, to hire one unit of managerial activity. Firms that quit production cannot reenter the market at a later stage and recoup the undepreciated part of their capital stocks, net of the adjustment cost.

Every period there is a constant exogenous mass $m > 0$ of prospective entrants, each of which receives an initial productivity s, with $s \sim \Lambda(s)$, a Pareto distribution with scale parameter η. Conditional on entry, the distribution of the idiosyncratic shock in the first period of operation is $\Gamma(z'|s)$, strictly increasing in s. Entrepreneurs that decide to enter must pay an entry cost $c_e \geq 0$.

Finally, in each period, the stationary distribution of operating firms over the three dimensions of heterogeneity is denoted by $\Omega(z, k_T, k_I; W)$. A comprehensive picture of timing in the model is presented in Figure 2.5.

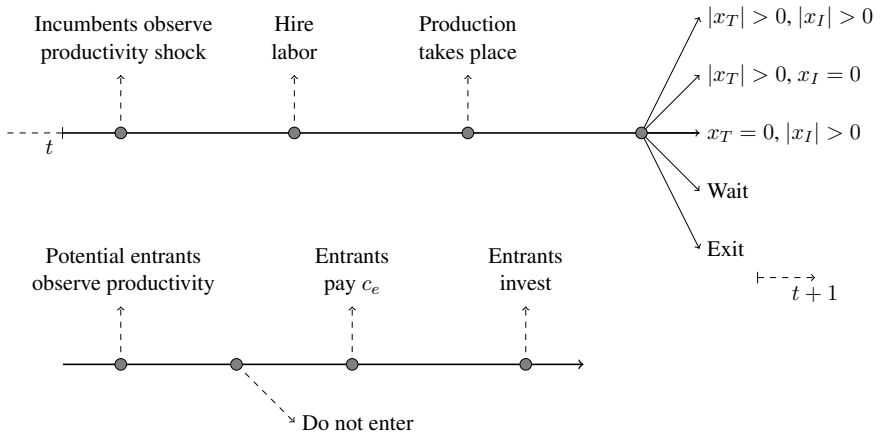
[18] This is a standard assumption in the literature; see, for example, Khan and Thomas (2008) and Clementi and Palazzo (2016b).

[19] Following Clementi and Palazzo (2016b) and Carvalho and Grassi (2019), we are assuming that the utility function of the representative household is given by

$$u(C, L) = C - \frac{L^{1+1/\psi}}{1 + 1/\psi}.$$

This is a convenient formulation that simplifies the numerical analysis while providing an elastic labor supply. However, none of our results hinge on this particular specification.

Figura 2.5: Timing in the Model



2.4.2 Problem of Incumbents

Given idiosyncratic productivity z, tangible capital k_T, and intangible capital k_I, the profits of an incumbent are given by

$$\pi(z, k_T, k_I; W) = \max_{\ell} \; e^z \left(k_T^{\alpha} k_I^{\nu} \ell^{(1-\alpha-\nu)}\right)^{\omega} - W\ell. \tag{2.6}$$

Upon exit, a firm obtains a value equal to the undepreciated portion of its tangible capital k_T and intangible capital k_I, net of the adjustment cost it incurs to dismantle them:

$$\mathcal{V}_x(k_T, k_I) = (1-\delta_T)k_T + (1-\delta_I)k_I - \mathcal{C}(-(1-\delta_T)k_T, -(1-\delta_I)k_I; k_T, k_I). \tag{2.7}$$

Then, the start-of-period value of an incumbent firm is dictated by the

function $\mathcal{V}(z, k_T, k_I; W)$, which solves the following functional equation:

$$\begin{aligned}\mathcal{V}(z, k_T, k_I; W) = \pi(z, k_T, k_I; W) \\ + \max\{\mathcal{V}_x(k_T, k_I), \tilde{\mathcal{V}}_1(z, k_T, k_I; W) - c_f, \\ \tilde{\mathcal{V}}_2(z, k_T, k_I; W) - c_f, \tilde{\mathcal{V}}_3(z, k_T, k_I; W) - c_f, \\ \tilde{\mathcal{V}}_4(z, k_T, k_I; W) - c_f\},\end{aligned} \tag{2.8}$$

where the value of investing in both types of capital is given by

$$\begin{aligned}\tilde{\mathcal{V}}_1(z, k_T, k_I; W) = \max_{k_T', k_I'} -x_T - x_I - \mathcal{C}(x_T, x_I; k_T, k_I) + \frac{1}{R}\mathbb{E}_z\mathcal{V}(z', k_T', k_I'; W), \\ \text{s.t.} \quad k_T' = (1-\delta_T)k_T + x_T, \\ k_I' = (1-\delta_I)k_I \ \ + x_I;\end{aligned} \tag{2.9}$$

the value of investing in only tangible capital is given by

$$\begin{aligned}\tilde{\mathcal{V}}_2(z, k_T, k_I; W) = \max_{k_T'} -x_T - \mathcal{C}(x_T, 0; k_T, k_I) + \frac{1}{R}\mathbb{E}_z\mathcal{V}(z', k_T', (1-\delta_I)k_I; W), \\ \text{s.t.} \quad k_p' = (1-\delta_T)k_T + x_T;\end{aligned} \tag{2.10}$$

the value of investing in only intangible capital is given by

$$\begin{aligned}\tilde{\mathcal{V}}_3(z, k_T, k_I; W) = \max_{k_I'} -x_I - \mathcal{C}(0, x_I; k_T, k_I) + \frac{1}{R}\mathbb{E}_z\mathcal{V}(z', (1-\delta_T)k_T, k_I'; W), \\ \text{s.t.} \quad k_i' = (1-\delta_I)k_I + x_I;\end{aligned} \tag{2.11}$$

and finally, the value of waiting is given by

$$\tilde{\mathcal{V}}_4(z, k_T, k_I; W) = \frac{1}{R}\mathbb{E}_z\mathcal{V}(z', (1-\delta_T)k_T, (1-\delta_I)k_I; W). \tag{2.12}$$

2.4.3 Problem of Entrants

The value of a potential entrant that draws an initial productivity s, where $s \sim \Lambda(s)$, is given by

$$\mathcal{V}_e(s;W) = \max_{k'_T,k'_I} -k'_T - k'_I + \frac{1}{R}\int \mathcal{V}(z',k'_T,k'_I;W)\Gamma(dz'|s). \quad (2.13)$$

Thus, the potential entrant will invest and start operating if and only if $\mathcal{V}_e(s;W) \geq c_e$.

2.4.4 Recursive Competitive Equilibrium

The recursive competitive equilibrium (RCE) consists of (i) value functions $\mathcal{V}(z,k_T,k_I;W)$, $\tilde{\mathcal{V}}_1(z,k_T,k_I;W)$, $\tilde{\mathcal{V}}_2(z,k_T,k_I;W)$, $\tilde{\mathcal{V}}_3(z,k_T,k_I;W)$, $\tilde{\mathcal{V}}_4(z,k_T,k_I;W)$, and $\mathcal{V}_e(s;W)$; (ii) policy functions $\ell(z,k_T,k_I;W)$, $x_T(z,k_T,k_I;W)$, $x_I(z,k_T,k_I;W)$, $k'_T(s;W)$, and $k'_I(s;W)$; and (iii) an incumbent's measure $\Omega(z,k_T,k_I;W)$ and an entrant's measure $\mathcal{E}(z,k_T,k_I;W)$ such that:

1. $\mathcal{V}(z,k_T,k_I;W)$, $\tilde{\mathcal{V}}_1(z,k_T,k_I;W)$, $\tilde{\mathcal{V}}_2(z,k_T,k_I;W)$, $\tilde{\mathcal{V}}_3(z,k_T,k_I;W)$, $\tilde{\mathcal{V}}_4(z,k_T,k_I;W)$, $\ell(z,k_T,k_I;W)$, $x_T(z,k_T,k_I;W)$ and $x_I(z,k_T,k_I;W)$ solve (2.6), (2.8), (2.9), (2.10), (2.11), and (2.12);

2. $\mathcal{V}_e(s;W)$, $k'_T(s;W)$ and $k'_I(s;W)$ solve (2.13);

3. The labor market clears: $\int \ell(z,k_T,k_I;W)d\Omega(z,k_T,k_I;W) = L(W)$;

4. For all Borel sets $\mathcal{Z}\times\mathcal{K}_T\times\mathcal{K}_I \subset \mathbf{R}^+\times\mathbf{R}^+\times\mathbf{R}^+$,

$$\mathcal{E}(\mathcal{Z}\times\mathcal{K}_T\times\mathcal{K}_I;W) = m\int_{\mathcal{Z}}\int_{\mathcal{B}_e(\mathcal{K}_T,\mathcal{K}_I;W)}\Lambda(ds)\Gamma(dz'|s),$$

   where $\mathcal{B}_e(\mathcal{K}_T,\mathcal{K}_I;W) = \{z \text{ s.t. } k'_T(s;W)\in\mathcal{K}_T, k'_I(s;W)\in\mathcal{K}_I \text{ and } \mathcal{V}_e(s;W)\geq c_e\}$;

5. For all Borel sets $\mathcal{Z}\times\mathcal{K}_T\times\mathcal{K}_I \subset \mathbf{R}^+\times\mathbf{R}^+\times\mathbf{R}^+$ and $\forall t \geq 0$,

$$\begin{aligned}\Omega(\mathcal{Z}\times\mathcal{K}_T\times\mathcal{K}_I;W) = &\int_{\mathcal{Z}}\int_{\mathcal{B}(\mathcal{K}_T,\mathcal{K}_I;W)}\Omega(dz,dk_T,dk_I;W)\Gamma(dz'|z)\\ &+ \quad \mathcal{E}(\mathcal{Z}\times\mathcal{K}_T\times\mathcal{K}_I;W),\end{aligned}$$

where $\mathcal{B}(\mathcal{K}_T, \mathcal{K}_I; W) = \big\{(z, k_T, k_I)\text{s.t.}\max\{\tilde{\mathcal{V}}_1(z, k_T, k_I; W), \tilde{\mathcal{V}}_2(z, k_T, k_I; W), \tilde{\mathcal{V}}_3(z, k_T, k_I; W), \tilde{\mathcal{V}}_4(z, k_T, k_I; W)\} - c_f \geq \mathcal{V}_x(k_T, k_I), (1-\delta_T)k_T + x_T(z, k_T, k_I; W) \in \mathcal{K}_T \text{ and } (1-\delta_I)k_I + x_I(z, k_T, k_I; W) \in \mathcal{K}_I\big\}$.

2.4.5 Output Elasticities, Adjustment Costs, and Allocative Efficiency

One of the main objects of interest for our analysis is the evolution of allocative efficiency resulting from IBTC. To define a model-consistent measure of allocative efficiency, we leverage the work of Hsieh and Klenow (2009) and define $TFPR$ in the model as

$$\begin{aligned} TFPR_{ft} &= \frac{y_{ft}}{k_{T,ft}^{\alpha} k_{I,ft}^{\nu} \ell_{ft}^{(1-\alpha-\nu)}} \\ &\propto \Big(ARPK_{T,ft}\Big)^{\alpha}\Big(ARPK_{I,ft}\Big)^{\nu}\Big(ARPL_{ft}\Big)^{(1-\alpha-\nu)}, \end{aligned} \tag{2.14}$$

where $ARPK_{T,ft} = y_{ft}/k_{T,ft}$ is the average product of tangible capital, $ARPK_{I,ft} = y_{ft}/k_{I,ft}$ is the average product of intangible capital, and $ARPL_{ft} = y_{ft}/\ell_{ft}$ is the average product of labor.[20] Therefore, our measure of allocative efficiency in the economy is defined by

$$\begin{aligned} Var(TFPR_{ft}) = \alpha^2 Var\left(ARPK_{T,ft}\right) + \nu^2 Var\left(ARPK_{I,ft}\right) \\ + 2\alpha\nu Cov(ARPK_{T,ft}, ARPK_{I,ft}) \end{aligned} \tag{2.15}$$

where $Var(\cdot)$ represents variance and $Cov(\cdot)$ is the covariance. This definition of allocative efficiency is the same as the one extensively used by the misallocation literature (Hsieh and Klenow (2009) and Hopenhayn (2014)). Notice that the allocative efficiency of this economy is independent of $ARPL$ as it is equalized across firms. This is driven by the assumption that labor input do not exhibit any adjustment costs in the

[20] Hopenhayn (2014) offers an extensive explanation of how to define $TFPR$ in models of perfect competition.

model.[21] Therefore, only the $ARPK_T$ and $ARPK_I$ are relevant to understand the evolution of allocative efficiency in our framework.

In the absence of adjustment costs to both types of capital, their average product would equalize across firms, and hence allocative efficiency as measured by the dispersion in $TFPR$ would be zero, which is by definition the highest level of allocative efficiency achievable in the model. Conversely, in the presence of adjustment costs to both types of capital, their average product no longer equalizes as the reallocation of both types of capital is slowed down by the adjustment costs themself. Therefore, the effect of the adjustment costs is to make $Var\left(ARPK_{T,ft}\right), Var\left(ARPK_{I,ft}\right) > 0$, and consequently $Var(TFPR_{ft}) > 0$.

Therefore, equation (2.15) clarifies the relation between IBTC and allocative efficiency in the model. An increase in the intangible capital share, ν, relative to the labor share, $1-\alpha-\nu$, increases the importance of $Var\left(ARPK_{I,ft}\right)$, and hence, all else equal, it increases overall dispersion in $TFPR$ and thereby lowers allocative efficiency in the model. This is just a by-product of the fact that IBTC lowers the reliance of firms on an undistorted input such as labor while increasing firms' reliance on a (potentially) highly distorted input such as intangible capital.

2.5 Quantitative Analysis

In this section, we use the structural framework presented in Section 2.4 to estimate the adjustment costs associated with tangible and intangible capital. Then, we use the following to validate our model: (i) non-targeted moments from the cross-sectional and age distribution and (ii) the empirical dispersion and responsiveness of the average revenue product of both types of capital.

[21] Adjustment costs associated with labor input may exist but remain small in comparison to intangible capital. Therefore, we focus on adjustment costs related to the both types of capital. This also helps us to reduce the number of state variables and the computational complexity.

2.5.1 Calibration

The baseline calibration jointly matches the investment behavior of tangible and intangible capital at the micro level and business dynamism in the overall US economy for the sample period 1980-1990. The parameterization proceeds in two steps. First, we fix a set of parameters that are estimated outside of the model—for instance, the parameters governing the production technology and the TFP process. Second, given the values of these fixed parameters, we choose the remaining parameters to match informative moments regarding firms' investment distribution and firms' life cycle.

Fixed parameters. A model period is one year, so we set the interest rate $R = 1.05$. The annual depreciation rate for tangible capital is $\delta_T = 0.07$, which equals the value used to perform the empirical analysis above. We set the depreciation rate for intangible capital at $\delta_I = 0.29$, which is the average firm-level depreciation rate from our data. The production function parameters comes from the estimates reported in Section 2.3.1. The returns to scale ω is set to 0.90 close to the values used in the literature.[22] Finally, the persistence of the idiosyncratic process is $\rho_z = 0.90$, and the standard deviation is $\sigma_z = 0.20$. These values are close to the empirical estimates reported in Foster et al. (2008) and in Lee and Mukoyama (2015).

Fitted parameters. We choose the remaining parameters to match some moments from Table 2.1 and some moments on business dynamism from Business Dynamics Statistics (BDS). Specifically, we use inaction rates, that is, investment rates that are within $\pm 1\%$, to discipline the parameters governing the fixed costs of investing in both tangible and intangible capital, f_T and f_I. This approach is particularly appealing since the model predicts that the fixed costs of adjusting directly influence the extensive margin of investment, that is, the amount of action and inaction in the investment of a given capital. We use the serial correlation of both

[22] Similar values have been used by Hopenhayn and Rogerson (1993) and Khan and Thomas (2008). In a perfectly competitive environment, decreasing returns to scale are necessary to ensure a well-defined firm distribution. Equivalently, one can leave returns to scale unconstrained and assume a downward-sloping demand.

investment rates to identify the convex costs of adjusting for both types of capital, γ_T and γ_I. With high convex costs, firms adjust their capital stock more slowly over time, which in turn increases the autocorrelation of investment at the firm level.[23] To identify the entry cost c_e, the operating cost c_f, and the parameter that governs the Pareto distribution of the productivity of potential entrants, η, we match the entry rate, the average size of incumbents, and the average size of entrants, respectively. Finally, we set the measure of potential entrants to m to target an equilibrium wage of 1.

The parameters are estimated using the following routine. For arbitrary values of the vector of parameters, $\mathcal{P} = (\gamma_T, \gamma_I, f_T, f_I, c_e, c_f, \eta, m)$, the model is solved and the policy functions for investment in both types of capital, for entry, and for exit are generated. Using these policy functions, the decision rules are simulated until the distribution of firms over $\{z, k_T, k_I\}$ is converged. We simulate the economy and construct a balanced panel of firms in the same spirit of the empirical analysis presented above. We compute the entry rate, the average size of entrants, and the average size of the incumbents from the stationary distribution. We compute the moments of the investment rates from the simulated balanced panel. We denote the vector of simulated moments as $\mathcal{M}(\mathcal{P})$. We estimate the fitted parameters $\widehat{\mathcal{P}}$ using a minimum distance criterion given by

$$\mathcal{L}(\mathcal{P}) = \min_{\mathcal{P}} \Big(\widehat{\mathcal{M}} - \mathcal{M}(\mathcal{P})\Big)' \mathbf{W} \Big(\widehat{\mathcal{M}} - \mathcal{M}(\mathcal{P})\Big). \tag{2.16}$$

Following Asker et al. (2014), we set the weighting matrix $\mathbf{W} = \mathbf{I}$ and use grid search to find the vector $\widehat{\mathcal{P}}$ that minimizes the objective function.

The fitted parameters from the grid search algorithm and the implied moments of the model are presented in Table 2.2. The model identifies different adjustment costs for tangible and intangible capital. Similar to Clementi and Palazzo (2019), our model imputes almost negligible fi-

[23] David and Venkateswaran (2019) explain that using the autocorrelation in investment as an identifying moment downward bias the calibration of convex adjustment costs in the presence of financial frictions. Therefore, we interpret our results as a lower bounds on the convex adjustment costs.

Taula 2.2: Parameters and Moments

Fixed	Value	Description
R	1.05	Annual interest rate
δ_T	0.07	Annual depreciation rate tangible capital
δ_I	0.29	Annual depreciation rate intangible capital
α	0.28	Tangible capital share
ν	0.03	Intangible capital share
ω	0.90	Returns to scale
ρ_z	0.90	Autocorrelation idiosyncratic productivity
σ_z	0.20	Standard deviation idiosyncratic productivity

Fitted	Value	Description	Moments	Model	Data
γ_T	0.006	Convex adj. cost k_T	corr$(x_{T,ft}, x_{T,ft-1})$	0.13	0.12
γ_I	0.135	Convex adj. cost k_I	corr$(x_{I,ft}, x_{I,ft-1})$	0.30	0.31
f_T	2.5·e-3	Fixed adj. cost k_T	Inaction rate: x_T	0.03	0.03
f_I	0.021	Fixed adj. cost k_I	Inaction rate: x_I	0.08	0.08
c_e	3·e-4	Entry cost	Entry rate	0.13	0.13
c_f	2.540	Operating cost	Avg. firm size	23.52	20.49
η	2.025	Scale parameter	Avg. entrant size	4.80	6.07
m	6.2·e-3	Measure of potential entrants	Wage	1	–

xed costs and low convex costs of investing in tangible capital, the reason being that the Compustat dataset contains disproportionately large firms.[24] Moreover, our model implies that intangible capital entails much higher adjustment costs relative to tangible capital. Therefore, the calibrated model shows that investment in intangible capital is subject to higher technological frictions and hence will be more distorted relative to a frictionless benchmark.

[24]However, contrary to our results, Cooper and Haltiwanger (2006) use a more heterogeneous sample of plants from the confidential Census database and find larger adjustment costs for tangible capital. Another point of distinction is that our analysis is at the firm level. Therefore, we want to emphasize that our estimates can be interpreted as a lower bound to these costs for both types of capital.

Finally, to validate the plausibility of our parameterization, we report that, in the model, the mean of the intangible capital investment rate and the tangible capital investment rate are, respectively, 0.32 and 0.07, close to the empirical counterparts of 0.35 and 0.14. This result is quite satisfactory as these moments are all untargeted. Moreover, we find that in the model, the standard deviation of tangible capital to sales and intangible capital to sales are respectively, 2.51 and 0.23, close to the empirical values of 2.47 and 0.36. This moment is particularly relevant as it partially identifies the persistence of the production process, as explained by Clementi and Palazzo (2016a). We also notice that the model produces an employment share for firms with $250+$ employees of 0.49 compared to 0.51 in the data.

2.5.2 Validation

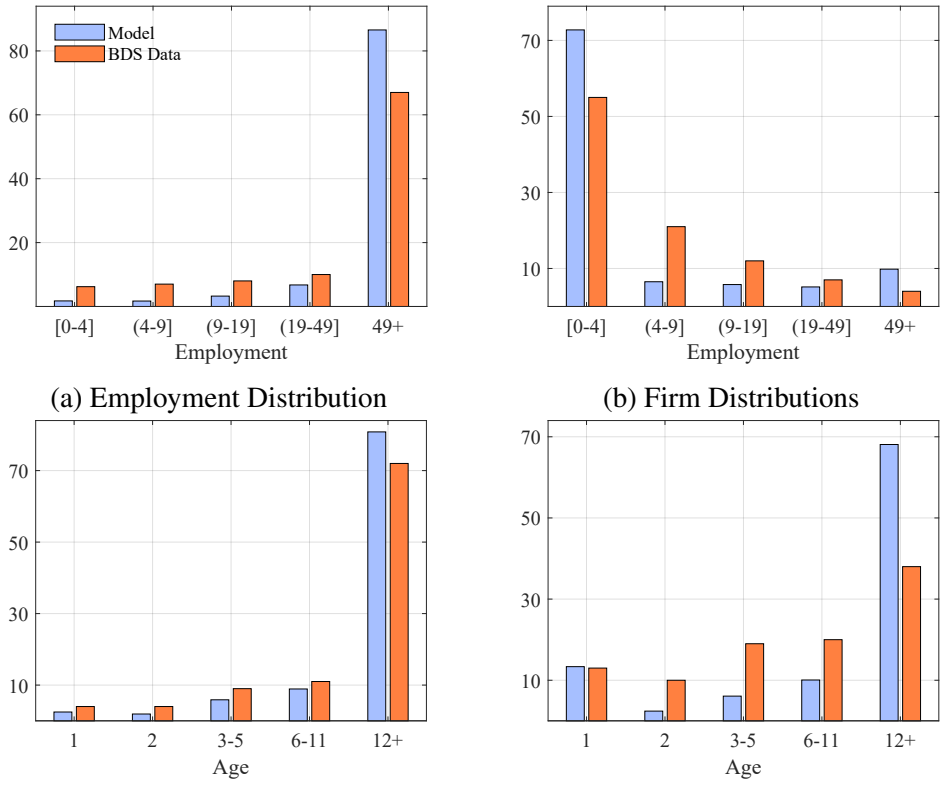
In this subsection, we use the following to validate our model: (i) non-targeted moments from the cross-sectional and age distributions (ii) the empirical dispersion and responsiveness of the average revenue product of both types of capital, and (iii) the cross-sectoral implications. Additional validation exercises are shown in Appendix 2.10.1.

Model Cross Section and Life Cycle

Here, we discuss the cross-sectional and life-cycle implications of the model. Figure 2.6 compares the distributions produced by the model with a representative empirical distribution constructed using the BDS dataset. Similar to what is documented in the previous literature on firm dynamics, the model exhibits size and age distributions that are right skewed.

Figure 2.6a shows that the model does a reasonably good job in matching the firm size distribution that is present in the data. This finding is not totally surprising as the average incumbent size and the average entrant size have been targeted in the calibration. Figure 2.6b shows that in the model, the majority of firms are small, whereas a large portion of employment is concentrated among the large firms, a feature well esta-

Figura 2.6: Size and Age Distribution



(a) Employment Distribution

(b) Firm Distributions

(c) Cohort Employment Share

(d) Age Distribution

Note. The figure panels show the size (employment) and age distribution of the firms, in both the model and the data. Orange bars show the empirical distributions; light blue bars show the distributions from the model. The top left panel shows the employment share across different employment categories. The top right panel shows the share of firms across different employment categories. The bottom left panel shows the employment share across different age bins. The bottom right panel shows the share of firms across different age bins. Empirical distributions are from the BDS data.

blished in the data and visible in Figure 2.6a. Finally, the model predicts that around 70% of the firms are operating for more than 11 years and that they account for around 80% of the employment share, which is slightly above what we observe in the data (see Figure 2.6c for cohort-wise employment shares and Figure 2.6d for the age distribution). Overall, the model does a satisfactory job in matching the empirical distributions of

size and age even though most of these distributions were not a particular target in the calibration.

Quasi-Fixed Inputs and Marginal Products

Here, we discuss the consequences of adjustment costs when firms are hit by productivity shocks. In particular, we focus on two main things: (i) the dispersion in the average revenue product of both types of capital and (ii) the responsiveness of the average revenue product of both types of capital to productivity shocks. The fact that capital, in the presence of adjustment costs and time to build, is a quasi-fixed input leads to an environment where the average revenue product of each type of capital is not equalized across firms. This happens because when a productivity shock hits, the firm cannot immediately adjust the capital stock to the desired frictionless level; therefore, the average revenue product of capital differs from the marginal cost, that is, the opportunity cost of holding the capital. Given that the calibration pointed out that intangible capital is more fixed compared to tangible capital as it subject to higher adjustment costs, the model predicts that the average revenue product of intangible capital is more dispersed as well. Moreover, this also implies that conditional on a productivity shock, intangible capital adjusts less than tangible capital, and, as a consequence, its average revenue product reacts by more.

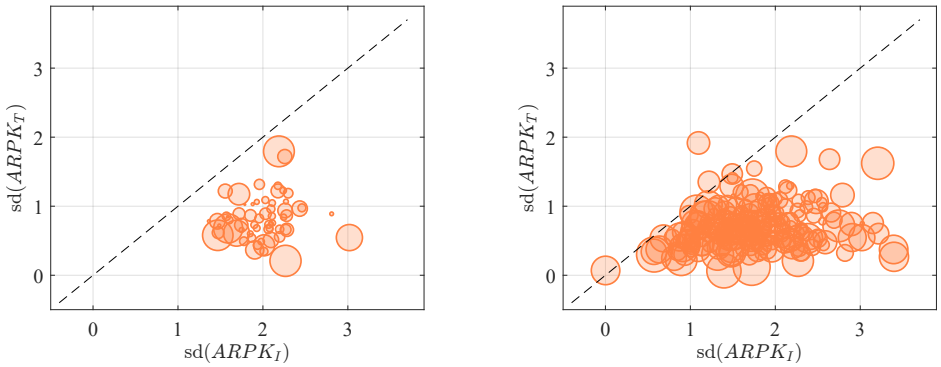
To test the aforementioned predictions of the model in the data, we need to compute the average product of both types of capital. Under the assumption that all firms share the same Cobb-Douglas production technology within a sector, we can compute in the data the log of the average revenue product of both types of capital, for firm f at time t as

$$ARPK_{j,ft} = \log(y_{ft}) - \log(k_{j,ft}), \quad j \in \{T, I\}, \tag{2.17}$$

where y_{ft} is firm-level output and $k_{j,ft}$ is firm-level capital.

We compute the dispersion in the average revenue product of capital at the SIC2 and SIC3 level. Results are presented in Figure 2.7, where we scatter plot the sector-level standard deviation of $ARPK_I$ against the sector-level standard deviation of $ARPK_T$. Both figures show that in

Figura 2.7: Sector-Level Dispersion in $ARPK_I$ and $ARPK_T$



(a) SIC2-Level (b) SIC3-Level

Note. The figures show the standard deviation of $ARPK_I$ (x-axis) and the standard deviation of $ARPK_T$ (y-axis). Standard deviations are calculated within sectors and averaged across the years. Average products are constructed as described in the text. The dashed black line shows the 45-degree line. Figure 2.7a is constructed calculating standard deviations at the SIC2 level; each circle represents a SIC2 sector, where the size of the circle is proportional to its size (sale weighted) in Compustat. Figure 2.7b is constructed calculating standard deviations at SIC3 level; each circle represents a SIC3 sector, where the size of the circle is proportional to its size (sale weighted) in Compustat.

the vast majority of sectors, if we consider both SIC2 and SIC3 levels of disaggregation, the average revenue product of intangible capital is more dispersed than that of tangible capital, as predicted by the theory.

Furthermore, we test the second prediction of the model, namely, the higher responsiveness of the average revenue product of intangible capital relative to the average revenue product of tangible capital to revenue productivity shocks. To do so, we perform the following regression:

$$ARPK_{j,ft} = \gamma_1 \varepsilon_{ft} + \gamma_2 k_{j,ft} + \gamma_1 TFPR_{ft-1} + \gamma_s + \gamma_t + \nu_{ft}, \quad j \in \{T, I\}, \tag{2.18}$$

where ε_{ft} is the innovation to log total factor productivity revenue.[25] The

[25] To compute ε_{ft} we run the following regression:

$$TFPR_{ft} = \rho TFPR_{ft} + \gamma_s + \gamma_t + \nu_{ft}.$$

Then, the firm-level innovation to revenue productivity is calculated as $\varepsilon_{ft} =$

regression coefficient of interest is γ_1. In the absence of any adjustment cost, or of time to build, the average revenue product of capital should equalize across firms and be constant; hence, the regression coefficient, γ_1, should be zero. Conversely, the more distorted an input is the higher its average revenue product response to a revenue productivity shock, and hence the higher the coefficient γ_1.

Taula 2.3: Heterogeneous Response of Average Products to $TFPR$ Shocks

	(1)	(2)	(3)	(4)
Dependent Variable	$ARPK_{T,ft}$	$ARPK_{I,ft}$	$ARPK_{T,ft}$	$ARPK_{I,ft}$
ε_{ft}	1.192***	1.592***	1.095***	1.239***
	(0.011)	(0.026)	(0.008)	(0.019))
$k_{T,ft}$			-0.111***	
			(0.001)	
$k_{I,ft}$				-0.399***
				(.001)
$TFPR_{ft-1}$			0.839***	0.940***
			(0.003)	(0.008)
Time dummies	v	v	v	v
Sector dummies	v	v	v	v
Observations	0.447	0.396	0.714	0.692
R-squared	89,967	89,967	89,967	89,967

Notes. We report the coefficients from the regressions of $ARPK_{T,ft}$ and $ARPK_{I,ft}$ on revenue productivity shock ε_{ft}. The controls include lagged revenue productivity, TFPR$_{ft-1}$, tangible capital, $k_{T,ft}$, and intangible capital, $k_{i,ft}$. Standard errors are in parentheses. *** p-value¡0.01, ** p-value¡0.05, * p-value¡0.1.

Results are presented in Table 2.3. As predicted by the theory, we find that the average product of both tangible and intangible capital reacts positively to revenue productivity shocks, as γ_1 is significantly greater than zero in all specifications. Moreover, the average revenue product of

$TFPR_{ft} - \widehat{\rho} \cdot TFPR_{ft}$.

intangible capital is more reactive to revenue productivity shocks relative to the average revenue product of tangible capital. This result is in line with the prediction of the model that firms do not adjust their intangible capital as frequently as their tangible capital because of the presence of high adjustment costs.

2.6 Intangible Capital Biased Technological Change at Work

In this section, we discuss the main mechanisms that drive our results. We describe the working of the model in detail and disentangle the partial and general equilibrium forces. Finally, we cross-validate the main mechanism in the data by exploiting cross-sector variation in the data.

2.6.1 Main Mechanism

Here, we analyze the underlying forces behind the main implications of IBTC. In the model, a rise in the output elasticity of intangible capital, at the cost of the labor elasticity, affects (i) the aggregate factor shares; (ii) the average firm size, profit rate, and concentration; and (iii) the allocative efficiency as measured by the dispersion in TFPR. This happens because when a distorted input such as intangible capital rises, it influences the demand of each input and many other equilibrium outcomes, such as equilibrium wages, firms' selection, firms' growth, and the allocation of capital across firms. Hence, the objective of this section is to uncover these forces and link them with the IBTC.

The two fundamental forces that drive the aggregate changes are: (i) change in the demand of the inputs due to the firm-level technological change and (ii) the endogenous change in the selection process of the firms due to the rise of a distorted input (intangible capital). First, IBTC makes production more intangible intensive at the expense of labor; this in turn increases the demand for intangible capital while depressing the

demand for labor. Therefore, this mechanically increases the intangible investment share while decreasing the labor share.

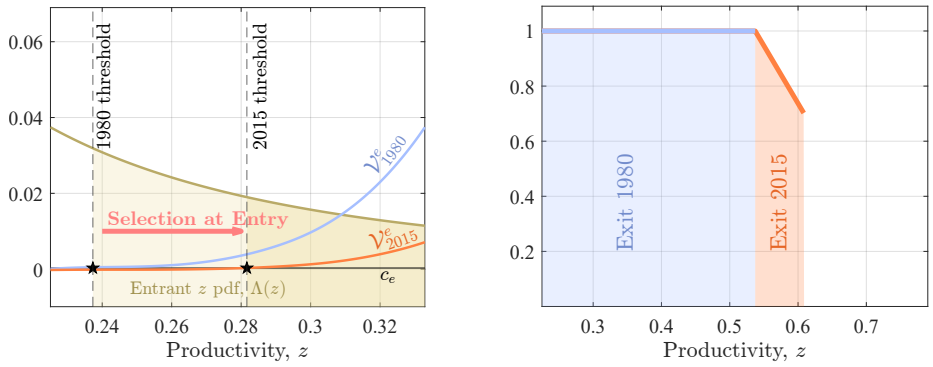
Second, this technological change commands the firm to invest more in a distorted input, subject to high adjustment costs, such as intangible capital. Only sufficiently productive firms can do this—that is, firms that are productive enough to face a positive value of operating in the new intangible-intensive economy. Therefore, this affects selection for both entrant firms and incumbent firms, as shown by Figure 2.8. Figure 2.8a shows the entry decision for potential entrants in 1980 (before IBTC) and in 2015 (after IBTC). Figure 2.8b shows the exit probability of incumbent firms in both economies.

IBTC lowers the value of entry, as shown in Figure 2.8a ($\mathcal{V}^e_{1980} > \mathcal{V}^e_{2015}$). This triggers a rightward shift of the entry threshold, implying that by 2015, only more productive firms can enter. This happens because only more productive firms can pay the entry cost c_e. Moreover, as shown by Figure 2.8b, incumbent firms are subject to a similar increase in selection. In the new economy, marginally more productive firms face a positive exit probability, as shown by the orange line. Overall, this means that IBTC increases both ex ante and ex post selection in the economy.

The rise in the aggregate intangible investment share is lower than the counterfactual increase implied by the firm-level rise in its input share in a frictionless model (without adjustment costs), as can be seen quantitatively in Section 2.7.[26] This is because despite IBTC mechanically increasing demand for intangible capital, it triggers a rise in the selection that favors more productive firms, dampening the rise of aggregate intangible capital share. In the model, high-productivity firms have a lower investment share as they expect to contract on average in the future because of the mean reversion in the productivity process. Therefore, a redistribution toward high-productivity firms translates into a redistribution toward low-investment-share firms. This composition effect dampens the rise of the aggregate intangible investment share. This same mechanism explains why the aggregate tangible investment share declines because of IBTC. In

[26] In a frictionless model, changes in the firm-level input shares uniquely pin down the change in aggregate input shares.

Figura 2.8: IBTC and Firms' Selection



(a) Entry Decision (b) Exit Probability

Note. Figure 2.8a shows graphically the entry problem of potential entrants both in the 1980 and 2015 calibrations. The 2015 calibration is shown in Section 2.7. The beige line in the background shows the productivity distribution of potential entrants, $\Lambda(z)$. The light blue and the orange curves show the value function of potential entrants for both calibrations, $\mathcal{V}^e_{1980}$ and $\mathcal{V}^e_{2015}$. The value of entry is lower in 2015 compared to 1980 because in order to grow in the intangible-intensive economy, firms have to spend more resources on high adjustment costs. The black line shows the entry cost, c_e. The two vertical dashed black lines show the exit threshold in both 1980 and 2015, that is, the productivity level that satisfies $c_e = \mathcal{V}^e_t(z),\ t \in \{1980, 2015\}$. The shaded light beige area in the background shows the ex post productivity distribution of entrants in 1980, and the shaded dark beige area in the background shows the ex post productivity distribution of entrants in 2015.

Figure 2.8b the exit probability of incumbent firms both for the 1980 and for the 2015 calibration. The light blue line shows the exit probability for incumbent firms in 1980. The orange line shows the exit probability in 2015. Firms with higher productivity in 2015 face a positive probability of exit because in the intangible-intensive economy, it is more difficult to operate as they have to spend more on adjustment costs in order to respond to productivity shocks.

this case, because the firm-level input share of tangible capital does not change over time, the composition effect drives this decline. Finally, the labor share declines only because of the change in the firm-level input share, as the composition effect has no impact on labor share, so that it is equalized across firms.

Moreover, IBTC raises the average firm size, profit rate, and concentration for two reasons: (i) it increases selection as explained above, and (ii) it favors the larger firms in the economy. IBTC makes the growth of small firms costly, as they have to incur very high adjustment costs to build their stock of intangible capital, whereas it makes it easier for large firms to shrink, as the high depreciation rate of intangible capital favors its

depletion. This mechanism, together with the above increase in selection, triggers a reallocation of sale shares toward the larger firms, reinforcing the rise in average firm size, profit rate, and industry concentration.

Finally, the model predicts that the allocative efficiency in the model declines as the intangible capital share increases. This is driven by the fact that $TFPR$ in our framework is a weighted geometric mean of the average revenue product of inputs, where the weights are proportional to their output elasticities. The presence of adjustment costs means that dispersion in $TFPR$ is driven by dispersion in average products of both types of capital. This can be seen from equation (2.15). When the output elasticity of intangible capital increases, the dispersion in $ARPK_I$ becomes the primary driver of the dispersion in $TFPR$.[27] Therefore, dispersion in $TFPR$, which is our measure of allocative efficiency (where higher dispersion in $TFPR$ means lower allocative efficiency), increases.

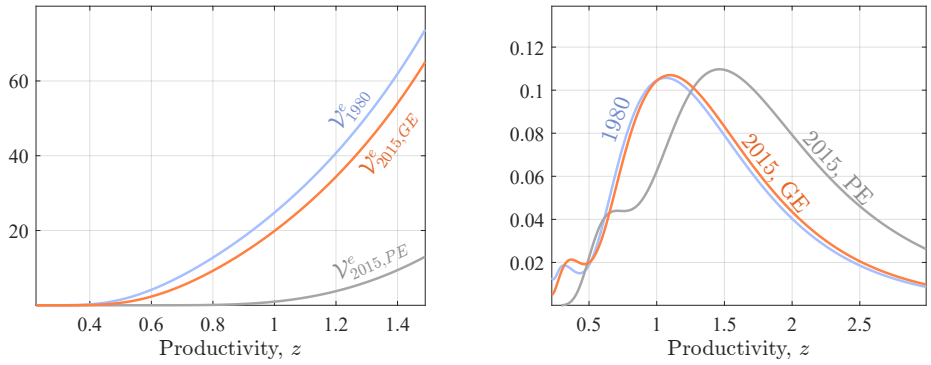
2.6.2 General Equilibrium versus Partial Equilibrium

Here, we highlight the consequences of IBTC on the economy and disentangle the partial and general equilibrium effects. To do so, we solve the model with the intangible-intensive production technology as estimated in 2015 while holding the wage constant; therefore, we only capture the partial equilibrium effects of IBTC. As discussed above, IBTC lowers the value of entry as it makes it more difficult for firms to operate. As shown in Figure 2.9a, the partial equilibrium value of entry $\mathcal{V}^e_{2015,PE}$ is significantly lower than $\mathcal{V}^e_{1980}$, thus pushing up the productivity of the marginal entrant. A similar rise in selection is also observed for exiting firms. As a result, the distribution of incumbent firms is shifted to the right, as shown in Figure 2.9b.

However, once we allow wages to adjust endogenously, the GE value of entry $\mathcal{V}^e_{2015,GE}$ increases and ends up being much higher relative to the $\mathcal{V}^e_{2015,PE}$ level because of a decline in wages that arises from an endo-

[27] Adjustment costs associated with the investment process of input do not allow its marginal product to equalize across firms and hence generate dispersion in the average revenue product.

Figura 2.9: General versus Partial Equilibrium effects



(a) Entrant Value Function (b) Incumbents Distribution

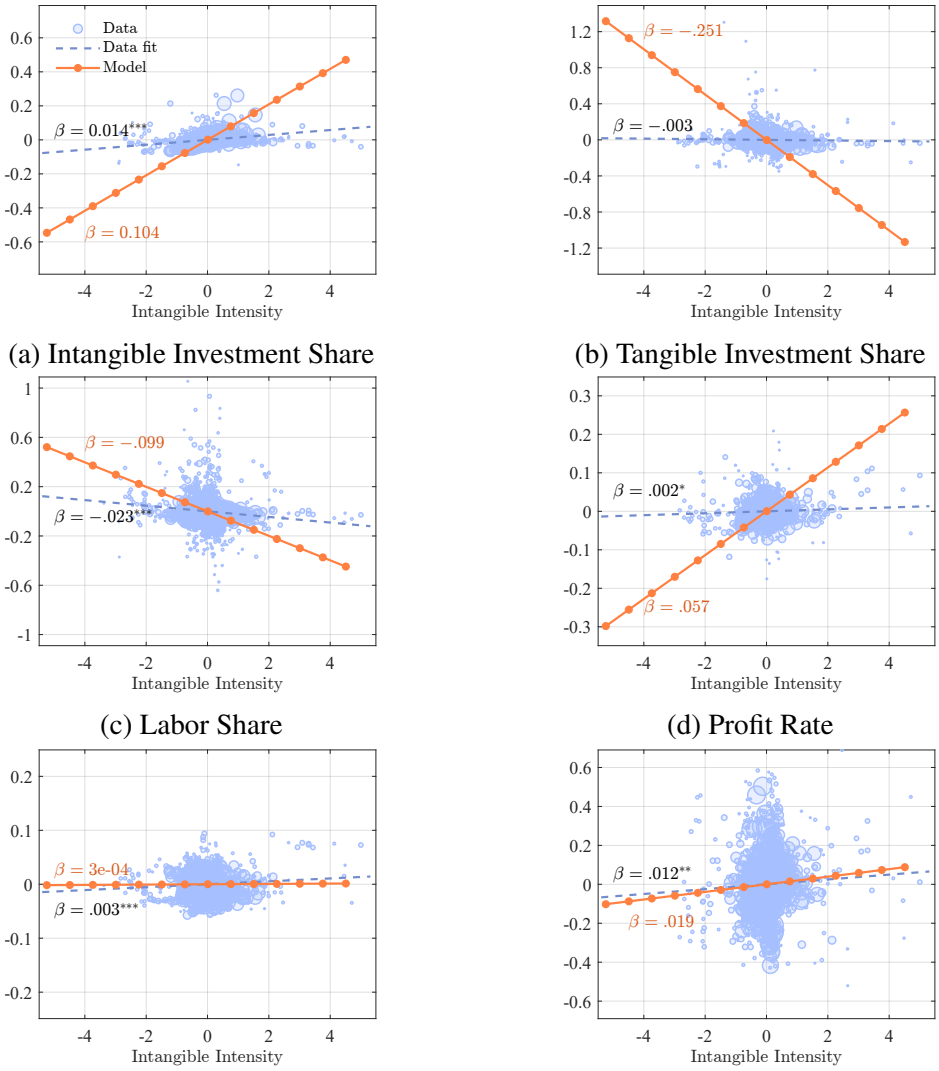
Note. Figure 2.9a shows the value of entry in 1980 and 2015 for both the general equilibrium version of the model and the partial equilibrium one. The light blue line shows the value of entry in 1980, the orange line shows the value of entry in 2015-GE, and the light grey line shows the value of entry in 2015-PE. The value of entry declines between 1980 and 2015 because in order to grow in the intangible-intensive economy, firms have to spend more resources on high adjustment costs. The value of entry declines more in PE relative to GE because in general equilibrium, the wage declines and acts like a dampening force on the effect of IBTC. Figure 2.9b shows the endogenous distribution of firms in the economy in 1980 and 2015 for both the general equilibrium version of the model and the partial equilibrium one. The light blue line shows the distribution in 1980, the orange line shows the distribution in 2015-GE, and the light grey line shows the distribution in 2015-PE. The distribution shifts to the right because of the increase in selection mentioned above. Again, the decline in wages dampens the PE effect, resulting in a milder shift of the GE distribution toward the right.

genous decline in firms' overall labor demand. This is an artifact of the reduced firm entry, reduction in the output elasticity of labor at the firm level, and the increase in the overall adjustment cost of investment faced by firms. This wage decline, which is a counterbalancing force to the selection effect from IBTC, increases the value of entry and profit rates relative to the partial equilibrium level.

This exercise highlights the importance of general equilibrium effects in pinning down the overall macroeconomic implications of IBTC, without which one would have significantly overestimated the role of IBTC in explaining the recent trends that are the main focus of this paper.

2.6.3 Cross-Sectoral Validation

Figura 2.10: Sector-Level Correlations: Model versus Data



(a) Intangible Investment Share

(b) Tangible Investment Share

(c) Labor Share

(d) Profit Rate

(e) Concentration (HHI)

(f) TFPR Dispersion

Note. The figure shows the cross-sectoral correlations between intangible intensity, $k_I/w\ell$, and various measures of interest. Light blue bubbles show the sector-year observations net of sector and time fixed effects. Sectors are defined at SIC2-level. The light blue dashed lines show the empirical fit. The solid orange lines with circles show the model-implied slope.

This section contains a validation of the mechanism described in the previous sections. Here, we test the model predictions about how different intangible capital intensities in production, defined as the ratio of the intangible capital share to the labor share, shape sector-level factor shares, concentration, and allocative efficiency. However, as pure technological intangible capital intensity in production is difficult to measure at the sector level, we use a robust model prediction and proxy it by the ratio of intangible capital to labor costs share.

Figure 2.10 shows the results. Dashed light blue lines show the data linear fit, and orange lines with circles show the model predictions.[28] We focus on six main observables of interest: (i) intangible investment share, (ii) tangible investment share, (iii) labor share, (iv) profit rate, (v) concentration, and (vi) $TFPR$ dispersion.

In the model, a rise in intangible capital intensity translates into a higher investment share of intangible capital relative to the other inputs whose shares instead decline. Therefore, the economy moves from labor, which is a highly flexible input, to intangible capital, which is highly distorted because of the presence of technological frictions. This translates into a decrease in the allocative efficiency of the economy as measured by the dispersion in $TFPR$. Finally, as investing in intangible capital is a costly activity because of the associated high adjustment costs, selection then increases, which in turn increases both the market concentration, as measured by the HHI index, and the overall profit rate. Overall, Figure 2.10 shows that all of the qualitative predictions of the model are in line with the data from the cross section of sectors.

[28] Because of the high-level non-linearities and the different disperion in intangible intensity between the model and data, to obtain the model predictions, we perturb the model around the steady state and then inferred the associated slope. We then used the inferred slopes to extrapolate the overall tendency.

2.7 Intangible Capital Biased Technological Change and Its Macroeconomics Implications

In this section, we study the quantitative implications of IBTC as documented in Section 2.3.1. First, we document the firm-level and macroeconomic implications of IBTC. Second, we document that our results are robust to alternative quantification exercises. Third, we discuss the relation between IBTC and market power and its policy implications.

2.7.1 Quantitative Implications

Here, we study the quantitative implications of IBTC documented in Section 2.3.1. In particular, we show the quantitative implications of a rise in the intangible capital share in firm-level production from 0.03 to 0.12 and of an associated decline in the labor share from 0.69 to 0.60, as estimated in the data for the period 1980-2015.[29]

Table 2.4 shows the results.[30] Looking at the firm-level moments, we can see that the IBTC explains the majority of the observed rise in the average firm size and of the rise in concentration, both as measured by the HHI and as measured by the employment share of firms with $250+$ employees. These results are driven by the exogenous technological change in the production process and the endogenous rise of selection in the model, as discussed in the previous section.

Then, we compare the quantitative implications of IBTC with the changes in factor shares documented by Koh et al. (2020). The model captures well the change in most of the factor shares in the non-financial corporate sector.[31] To study the implications of the rise of intangible capital on the decline in the labor share, we follow Koh et al. (2020) and

[29] We leave the tangible capital share unchanged as it does not show any particular trend over the period of interest.

[30] In Appendix 2.10.2, we document the evolution of the distribution of the firm-level intangible intensity and TFPR in both the model and the data.

[31] We focus on the non-financial corporate sector as it has the best mapping with the Compustat data we used in the empirical analysis.

Taula 2.4: Quantitative Implications of IBTC

			Change	
	1980 S.S.	2015 S.S.	Model	Data
Firm Distribution				
Avg. firm size	23.523	26.121	$+11\%$	$+15\%$
Concentration	7.14e-04	9.88e-04	$+38\%$	$+33\%$
Employment share				
firms with $250+$ employees	0.489	0.551	$+6$p.p.	$+6$p.p.
Aggregate Factor Shares				
Intangible				
investment share	0.014	0.055	$+4$p.p.	$+4$p.p.
Tangible				
investment share	0.078	0.070	-1p.p.	-2p.p.
Labor share	0.666	0.580	-9p.p.	-8p.p.
Labor share				
pre-revision	0.676	0.614	-6p.p.	-5p.p.
Profit rate (Compustat)	0.242	0.294	$+5$p.p.	$+3$p.p.
Profit rate (BEA)	0.242	0.294	$+5$p.p.	$+5$p.p.
Aggregate Investment Rate				
Tangible				
investment rate	0.052	0.041	-1p.p.	-2p.p.
Allocative Efficiency				
sd($TFPR$)	0.202	0.227	$+12\%$	$+38\%$
Adjusted sd($TFPR$)	0.202	0.227	$+12\%$	$+15\%$

Notes. All of the variables are calculated coherently to their definitions as used in the data. The data sources are BDS, NIPA tables, and Compustat. To calculate the empirical moments from the 1980s, we use the time window 1980-1990, whereas for the empirical moments from 2015, we simply use the values in that year. The evolution of each trend is presented in Appendix 2.9.5.

compute two different labor shares in the model, given by

$$LS \equiv \frac{WL}{Y} \quad \text{and} \quad LS \text{ pre-revision} \equiv \frac{WL}{Y - X_I}, \tag{2.19}$$

where W is the wage, L is aggregate labor, Y is aggregate output net of adjustment costs and fixed costs, and X_I is the aggregate intangible investment. The labor share pre-revision is the counterfactual labor share that would emerge if intangible capital was not counted in the overall GDP calculation. Similar to the empirical evidence in Koh et al. (2020), we find that the pre-revision labor share declines much less than the true labor share. This finding confirms the interpretation of the authors that rising intangible capital investment is quantitatively and important factor in the decline of the labor share observed in the data. Moreover, the model can satisfactorily explain the rise in the intangible capital investment share, the decline in the tangible investment share, and the rise in the profit rate. We find a lower increase in the profit rate in the data compared to De Loecker et al. (2020) because we also account for balance sheet intangible investment, as documented in Appendix 2.9.5.

Finally, when looking at the overall allocative efficiency of the economy, we see that the IBTC can explain a substantial share of its downward trend: notice that an increase in the standard deviation of $TFPR$ translates into a decline in allocative efficiency. In the model, when firms rely more on an input that is highly distorted as intangible capital relative to a flexible input such as labor, inputs become slower in reallocating toward more productive firms, and hence the overall allocation of resources worsens, as measured by an increase in the standard deviation of $TFPR$. However, we emphasize that this cannot be considered as misallocation as the economy is fully efficient and the allocation of resources coincides with the one of the social planner. Concluding, the model does a good job in matching the quantitative decline in allocative efficiency, particularly in the case of the adjusted one.[32]

[32] Adjusted allocative efficiency is measured as allocative efficiency net of a potential 60% measurement error as documented by Bils et al. (2020).

2.7.2 Robustness Checks

In this section, we perform two additional exercises to test the validity of our results from Table 1.3. In particular, we check (i) how the decline in the intangible capital investment price relative to the tangible capital investment price affects our baseline findings and (ii) how our results would change if we were to reestimate the adjustment costs with moments from the investment rate distribution computed with the final part of the sample.[33]

We estimate the relative price of intangible capital as the ratio of the intangible capital deflator to the tangible capital deflator. We find that between 1980 and 2015, this relative price experienced a decline of approximately 20%, suggesting that intangible capital is becoming cheaper relative to tangible capital. To introduce this decline into the model, we substitute into the value functions 2.9, 2.10, 2.11, 2.12, and 2.13 the relative price of intangible capital investment p such that the final intangible capital investment bill is px_I.[34]

To perform the second exercise, we just reestimate the capital adjustment costs to match moments from the investment rate distribution of both capital in the period 2000-2015. Table 2.7 in Appendix 2.9.4 shows the evolution of the investment rate distribution over time, and Table **??** in Appendix 2.10.3 shows the new calibrated parameters and the associated targeted moments.

Table 2.5 shows the results. The first column shows the benchmark results as also reported in Table 1.3. The second and third columns show the results obtained by reestimating the adjustment costs and the results obtained by accommodating the decline in the relative price of intangible capital investment. The final column reports the values from the data. Both robustness exercises show the same qualitative patterns as in the benchmark case, and results overall seem to be robust to these departures

[33] All of these additional robustness exercises are conducted in conjunction with IBTC; that is, we always re-estimate the model with IBTC together with one of the aforementioned changes.

[34] Therefore, in our quantitative experiment, we allow the price p to go from 1 in 1980 to 0.8 in 2015.

Taula 2.5: Quantitative Implications of IBTC

	Change			
	Benchmark	Alternative Adj. Costs	Decline Rel. Price k_I	Data
Firm Distribution				
Avg. firm size	$+11\%$	$+13\%$	$+5\%$	$+15\%$
Concentration (HHI)	$+38\%$	$+36\%$	$+26\%$	$+33\%$
Employment share				
firms with $250+$ employees	$+6$p.p.	$+7$p.p.	$+6$p.p.	$+6$p.p.
Aggregate Factor Shares				
Intangible				
investment share	$+4$p.p.	$+4$p.p.	$+4$p.p.	$+4$p.p.
Tangible				
investment share	-1p.p.	-1p.p.	-0p.p.	-2p.p.
Labor share	-9p.p.	-9p.p.	-9p.p.	-8p.p.
Labor share				
pre-revision	-6p.p.	-6p.p.	-5p.p.	-5p.p.
Profit rate (Compustat)	$+5$p.p.	$+5$p.p.	$+4$p.p.	$+3$p.p.
Profit rate (BEA)	$+5$p.p.	$+5$p.p.	$+4$p.p.	$+5$p.p.
Aggregate Investment Rate				
Tangible	-1p.p.	-1p.p.	-1p.p.	-2p.p.
investment rate				
Allocative Efficiency				
sd($TFPR$)	$+12\%$	$+12\%$	$+11\%$	$+38\%$
Adjusted sd($TFPR$)	$+12\%$	$+12\%$	$+11\%$	$+15\%$

Notes. All of the variables are calculated coherently to their definitions as used in the data. The data sources are BDS, NIPA tables, and Compustat. To calculate the empirical moments from the 1980s, we use the time window 1980-1990, whereas for the empirical moments from 2015, we simply use the values in that year.

from the benchmark case. Moreover, we notice that even quantitatively results do not seem to deviate significantly from these alternative specifications. Effectively, what really matters for our results is that technology

is just shifting toward an input whose sunk cost of adjusting it (or of using it) is relatively higher compared to the other inputs. Therefore, we conclude that our results are robust and that they just hinge on the main properties of estimated technology (both the production technology and the adjustment costs technology).

2.7.3 IBTC, Market Power, and Policy Implications

In our framework, as production technology becomes more intangible intensive, firms invest more in an input that entails higher adjustment costs. Although this technological change raises market concentration, firm size, and the aggregate profit rate, resources are still allocated efficiently across firms. The observed decline in allocative efficiency in the model is due to technological constraints, and therefore, the decentralized equilibrium allocation still coincides with the one provided by the social planner. Our paper suggests that a sizeable part of the macroeconomic changes that have been witnessed in the US economy are the by-product of an efficient technological change.

However, this conclusion does not exclude that other forces above and beyond the mechanism documented here are at play in the economy. For instance, consider a slightly different version of our baseline model. Instead of assuming that firms produce the same good, we could have allowed firms to produce differentiated goods aggregated à la Kimball, as in Edmond et al. (2018). In such a framework, markups would be positively correlated with firm size. Therefore, a technological change that favors larger firms would shift market shares toward high-markup firms and away from low-markup firms. The measured decline in allocative efficiency in the model would be magnified by the rise in the dispersion of markups on top of the one already generated by IBTC. Moreover, even though in this alternative framework, the decentralized allocation would not coincide with the one provided by the social planner, the implementation of the social planner allocation, through any potential optimal policy, would coincide with the allocation in our baseline framework. As a consequence, while extended frameworks could give rise to desirable policy

interventions, our work suggests that at least a significant part of many of the macroeconomic trends that we observe in the US economy could be the by-product of efficient responses to changes in the firm-level production technology.

2.8 Conclusion

In the last four decades, firm-level investment in intangible capital, such as research and development, intellectual property products, and computerized information, has dramatically increased in the US. However, little is still known about its intrinsic properties and its implications for the economy overall. In this paper, we take a step forward in the understanding of this new type of capital.

We estimate the firm-level production function, finding that intangible capital is an important input in production and that its input share has gone from 0.03 in the 1980s to 0.12 in 2015. Moreover, we document that most of this rise has happened at the expense of the labor share in production. We interpret these findings as a paradigm shift in the production process of US firms; for instance, consider the importance that software and other intellectual property products have increasingly gained in the economy. We refer to transformation in the firm-level production process as intangible capital biased technological change.

We then document some novel properties of intangible capital, particularly, the fact that this new capital entails higher adjustment costs compared to tangible capital. This is consistent with the view that investments in intangible capital are plagued by inherent indivisibilities and are often sunk.

Finally, using a structural model of firm-level investment dynamics, we document the quantitative implications of IBTC. We find that this technological change can jointly explain a sizeable fraction of the increase in average firm size, the increase in concentration, the change in the aggregate factor shares, the decline in the tangible capital investment rate, and the decline in allocative efficiency. Our findings bring intangible

capital, its properties, and its trends to the center of the macroeconomic transformations that have been witnessed in the US economy. Therefore, we hope this paper will spur new, exciting research on this topic.

2.9 Empirical Appendix

2.9.1 Data

Main Sample, Variables, and Summary Statistics

We use the Compustat dataset from 1980 to 2015. We linearly interpolate `SALE`, `COGS`, `XSGA`, `EMP`, `PPEGT`, `PPENT`, `XRD`, `INTAN`, `GDWL`, `AM`. We exclude utilities (SIC codes between 4900 and 4999) because they are heavily regulated on prices. We also exclude financial firms (SIC codes between 6000 and 6999) because their balance sheets are dramatically different from other firms.

For data quality, we interpret as mistakes if `SALE`, `PPEGT`, `PPENT`, `COGS`, `EMP`, or `XSGA` are zero, negative, or missing, and we drop those observations. Moreover, if `XSGA` is missing or negative, we drop it as well. Finally, if `XRD`, `INTAN`, `AM`, or `GDWL` are negative or missing, we treat them as zeros. To obtain a real measure of the main variables, we deflate them with the GDP deflator, and we deflate investment in tangible and intangible capital by the appropriate deflators.[35] Table 2.6 presents a few basic summary statistics for a few leading variables used in our analysis.

User Cost of Tangible and Intangible Capital

One of the challenges of using the cost shares approach to estimate the firm-level production function is that it requires a measure of the user cost of capital. To this end, we define the user cost of capital as

$$r_{j,t} = i_t - \mathbb{E}_t \pi_{t+1} + \delta_j, \quad j \in \{T, I\}, \tag{2.20}$$

where i_t equals the nominal interest rate, $\mathbb{E}_t \pi_{t+1}$ is expected inflation at time t, and δ_j is the capital-specific depreciation rate. We take the annual Moody's Seasoned Aaa Corporate Bond Yield as an empirical proxy of the nominal interest rate and use the annual growth rate of the Investment

[35] Deflators are taken from the NIPA tables.

Taula 2.6: Summary Statistics (1980-2015)

	Sales	Cost of Goods Sold	Employment	Tangible Capital Stock	Intangible Capital Stock
Mean	2,310,810	1,572,800	7,966	1,572,164	284,519
25^{th} Percentile	27,495	14,880	131	8,004	2
Median	153,005	89,241	686	51,066	3,098
75^{th} Percentile	809,728	510,199	3,625	349,551	34,060
No. Obs.	188,151	188,151	188,151	188,151	188,151

Note. Summary statistics of cleaned Compustat dataset between 1980 and 2015. All variables are in thousands of US$. Sales and Cost of Goods Sold are deflated with the GDP deflator with base year 2012, and both types of capital stock are deflated using the appropriate investment deflator with base year 2012.

Nonresidential Price Deflator to calculate expected inflation. The depreciation rate of tangible capital is calibrated to $\delta = 0.07$, and the firm-level depreciation rate of intangible capital is computed as a weighted average of the depreciation rates used to construct the intangible capital stock.[36,37,38,39]

Intangible Capital Measurement and Accounting Standards

Measuring intangible capital is a difficult task as the accounting standards (US GAAP) are insufficient to satisfactorily book the intangible assets on

[36]Moody's Seasoned Aaa Corporate Bond Yield: https://fred.stlouisfed.org/series/AAA

[37]Investment Price Deflator: https://fred.stlouisfed.org/series/A008RD3Q086SBEA

[38]We estimate an AR(1) process on the annual growth rate of the Investment Nonresidential Price deflator and define the contemporaneous expected inflation as $\mathbb{E}_t \pi_{t+1} = \mu + \rho \pi_t$.

[39]The firm-level depreciation rate of intangible capital is computed as

$$\delta_{I,ft} = \frac{k_{ft}^{R\&D}}{k_{ft}^{R\&D} + k_{ft}^{BS}} \delta_s^{R\&D} + \frac{k_{ft}^{R\&D}}{k_{ft}^{R\&D} + k_{ft}^{BS}} 0.20.$$

the balance sheets. It is well established in the corporate finance literature that intangible assets are not fully captured on firms' balance sheets because of the anachronism of the US GAAP. [40] In this section of the appendix, we explain in detail which assumptions are needed to compute intangible capital at the firm level using the balance sheet for stocks and the income statements for flows.

To introduce our main measure, we have to clarify that intangible capital is intrinsically different from tangible capital as a significant part of it is internally generated by the firms. For nearly all internally generated intangible assets, such as knowledge and organizational capital, accounting standards differ significantly from tangible assets. All purchases of tangible assets are recorded on the balance sheet at their purchased price and depreciated over their useful life. Conversly, internal intangible capital investments, such as firms' R&D expense, advertising, or training of employees, are fully expensed in the period they are incurred.[41]

For instance, the Coca-Cola Company spends several billion dollars each year to maintain and promote its products and brands, such as Coca-Cola and Dasani. These are the assets of the firm that will generate future benefits in the form of higher margins and increased sales volume. Howe-

[40] Lev and Gu (2016) write, *Revolutionary changes, shifting economies and business enterprises from the industrial to the information age, started to profoundly affect the business models, operations, and values of companies in the 1980s, yet, amazingly, triggered no change in accounting. Entire industries, which are largely intangible (conceptual industries, as Alan Greenspan called them), including software, biotech, and internet services, came into being during the 1980s and 1990s. And for all other businesses, the major value drivers shifted from property, plant, machinery, and inventories, to patents, brands, information technology, and human resources. The latter set, all missing from companies? balance sheets because accountants treat intangible investments like regular expenses (wages, or interest), thereby distorts both the balance sheet and income statement. The constant rise in the importance of intangibles in companies? performance and value creation, yet suppressed by accounting and reporting practices, renders financial information increasingly irrelevant.*

[41] However, there are some exceptions. For example, US GAAP treats the development of computer software differently from other R&D costs. Following the ASC 985 (formerly FAS 2), once a software developer has reached technological feasibility, the developer must capitalize and amortize all development costs until the product becomes available for general release to consumers.

Figura 2.11: Advertising Expenses of Coca Cola

Selling, General and Administrative Expenses

The following table sets forth the significant components of selling, general and administrative expenses (in millions):

Year Ended December 31,	2016	2015	2014
Stock-based compensation expense	$ **258**	$ 236	$ 209
Advertising expenses	**4,004**	3,976	3,499
Selling and distribution expenses	**5,177**	6,025	6,412
Other operating expenses	**5,823**	6,190	7,098
Selling, general and administrative expenses	$ **15,262**	$ 16,427	$ 17,218

ver, the Coca-Cola Company is not allowed to recognize these assets on its balance sheet. Figure 2.11 shows that Coca-Cola spent around $4 billion in advertising in 2016. We also provide the example of Google Inc., which spent around $16 billion in research and development and $12 billion in sales and marketing in 2017 (see Figure 2.12a and Figure 2.12b).

Overall, these figures prove that a lot of intangible capital investment that is simply expensed by firms, in accordance with the US GAAP, does not show up as capital on the balance sheet. To overcome this limitation in the accounting standards, we capitalize knowledge capital, as explained in Section 2.2.2.

Externally acquired intangible capital can be capitalized on firms' balance sheets at the fair value according to the US GAAP under guidelines provided from ASC 350 (formerly FAS 142) and shows up in Compustat in the variable `INTAN`. According to Ewens et al. (2019), firms and accountants follow the guidelines provided in ASC 820 (formerly FAS 157) to mark externally acquired intangible capital on the balance sheet at its fair value at the time of the acquisition. Firms can choose among different methods to compute the fair value according to the US GAAP, and firms' choices must be disclosed in the appraisal notes for intangibles in the buyer's financial statements. Firms have three options to appraise the value of intangible assets: (i) estimating the replacement cost of the asset, (ii) comparing the asset to a similar asset whose price trades on the open market, or (iii) using the Discounted Cash Flow model, where earnings or cash flows are discounted by an appropriate discount rate. In particular, acquired intangible assets can be individually capitalized with the metho-

Figura 2.12: Intangible Investments by Google

Research and Development

The following table presents our R&D expenses (in millions):

	Year Ended December 31,		
	2015	2016	2017
Research and development expenses	$ 12,282	$ 13,948	$ 16,625
Research and development expenses as a percentage of revenues	16.4%	15.5%	15.0%

R&D expenses consist primarily of:

- Compensation expenses, including SBC, and facilities-related costs for employees responsible for R&D of our existing and new products and services; and
- Depreciation and equipment-related expenses.

(a) Research and Development Expenses

Sales and Marketing

The following table presents our sales and marketing expenses (in millions):

	Year Ended December 31,		
	2015	2016	2017
Sales and marketing expenses	$ 9,047	$ 10,485	$ 12,893
Sales and marketing expenses as a percentage of revenues	12.1%	11.6%	11.6%

Sales and marketing expenses consist primarily of:

- Advertising and promotional expenditures related to our products and services; and
- Compensation expenses, including SBC, and facilities-related costs for employees engaged in sales and marketing, sales support, and certain customer service functions.

(b) Marketing Expenses

dologies reported above if and only if they are identifiable, as documented in the ASC 805 notes. An intangible asset is identifiable if it meets either (i) the separability criterion, meaning it can be separated from the entity and sold, or (ii) the contractual-legal criterion, meaning that the control of the future economic benefits arising from the intangible asset is warranted by contractual or legal rights. In other words, IIA prices reflect fair or public value rather than value specific to the post-acquisition firm. Some examples of these identifiable intangible assets include brand names, customer lists, trademarks, Internet domain names, royalty agreements, patented technologies, and trade secrets. Other intangibles with a non-zero value, such as corporate culture, advertising effectiveness, and management quality, that fail to meet these criteria for identification are

captured as goodwill on the buyer's balance sheet (GDWL in Compustat).

Figura 2.13: Coca-Cola's Externally Purchased Intangibles

Coca-Cola Co.

Balance sheet: goodwill and intangible assets

US$ in millions

	Dec 31, 2019	Dec 31, 2018	Dec 31, 2017	Dec 31, 2016	Dec 31, 2015
Trademarks	9,266	6,682	6,729	6,097	5,989
Bottlers' franchise rights	109	51	138	3,676	6,000
Goodwill	16,764	10,263	9,401	10,629	11,289
Other	110	106	106	128	164
Indefinite-lived intangible assets	**26,249**	**17,102**	**16,374**	**20,530**	**23,442**
Customer relationships	344	185	205	392	493
Bottlers' franchise rights	341	30	213	487	604
Trademarks	177	186	182	228	211
Other	55	88	94	179	97
Definite-lived intangible assets, gross carrying amount	**917**	**489**	**694**	**1,286**	**1,405**
Accumulated amortization	(400)	(321)	(432)	(688)	(715)
Definite-lived intangible assets, net	**517**	**168**	**262**	**598**	**690**
Intangible assets	**26,766**	**17,270**	**16,636**	**21,128**	**24,132**

Based on:10-K (filing date: 2020-02-24),10-K (filing date: 2019-02-21),10-K (filing date: 2018-02-23),10-K (filing date: 2017-02-24),10-K (filing date: 2016-02-25).

We give an example of Coca-Cola's externally purchased intangibles in Figure 2.13. Coca-Cola writes in their yearly report: *We classify intangible assets into three categories: (1) Intangible assets with definite lives subject to amortization, (2) intangible assets with indefinite lives not subject to amortization and (3) goodwill.* The goodwill and intangible assets with indefinite lives are subject to an impairment test every period, and their values are increased or decreased accordingly. As one can see, the balance sheet intangibles are the sum of heterogeneous assets, such as trademarks, franchise rights, and customer relationships, among others.

Internally generated intangible capital: Potential issues. The fact that a sizeable fraction of intangible capital is internally produced and cannot be capitalized on firms' balance sheets potentially implies that some possible concerns related to the double-counting of some intangible assets. For example, when firm 1 produces its own intangible capital, it will expense it in its income statement at production cost x. If this intangible capital is then sold to firm 2, this sale will not show up in the income statement of firm 1 as a negative cost (or a negative investment).

Firm 2 will, however, show this new intangible capital on its balance sheet at fair value y because it has been externally acquired. In this example, even though the overall ammount of intangible capital has not changed in the economy—just a transaction has taken place—we would potentially observe an increase in the overall stock of intangible cpaital from x to $x + y$.

Although this is a concern in theory, as a practical matter, we are confident that this situation is rare and hence of little quantitative relevance. First, we know that intangible capital is often acquired through the acquisition of an entire firm.[42] Hence, as the target firm is acquired, it exits the sample, and its intangible assets leave the sample as well, whereas now the acquiring firm will show an increase in its intangible capital on the balance sheet. Second, we also know that a lot of intangible capital is acquired as final goods from other firms (consider, for instance, software producers and advertisement/marketing companies), and in this case as well, there is no double-counting as this is final production and not internal production for firms' own use. Third, as showed in the previous section, internally produced intangible capital is a declining feature of our empirical measure, suggesting that this concern should be minor and declining over time. Therefore, we conclude that this issue is not quantitatively appealing.

Externally acquired intangible capital: Potential issues. Externally purchased intangible capital is almost often acquired through acquisitions of entire firms, and this greatly influences the way it is capitalized on the firms' balance sheets. For example, imagine firm x buys firm y at cost p^y. At the moment of the acquisition, firm x has to place the acquired assets on its balance sheet. Normally, the procedure is: (i) tangible assets are identified and capitalized at the fair value p^T, (ii) identifiable intangible assets are capitalized at the fair value p^I, and (iii) the residual value is attributed to unidentifiable intangible assets (synergies, organizational culture, etc.) and capitalized into goodwill. Therefore, in the data, we have $\texttt{GDWL} = p^y - p^T - p^I$.

If a researcher thinks that firms acquire other firms to exercise futu-

[42] Peters and Taylor (2017) and Ewens et al. (2019).

re market power (and so firms are willing to pay high prices for them), the concern can arise that these unidentifiable intangible assets are just the discounted expected sum of the value of future market power, and therefore the value of balance sheet intangibles goes up by more than its quantity. One way to address this concern is to use proper deflators, that is, to deflate intangible capital with the IPP deflator.[43] However, this only takes care of aggregate common trends and cannot account for the heterogeneity in firm-level input prices, and unfortunately more disaggregated investment deflators do not exist. We wish to emphasize that the inability to obtain firm-level investment deflators equally affects the measurement of tangible and intangible capital. Additionally, as a more appealing way to address these concerns, we remove goodwill from total balance sheet intangible capital as almost all of the potential rise in prices related to unidentifiable assets will be captured exactly by a rise in goodwill. However, we want to acknowledge that Ewens et al. (2019)—using more detailed data than we have—have shown that at least 38% of firms' goodwill is indeed true intangible capital. Therefore, we see this solution as a necessary but imperfect solution.

Accounting standards for software: A special case. The accounting standards for expenditures in internal software development or external purchases are different from those of other intangible assets. In particular, the FASB ASC subtopic 350-40 provides guidelines for the accounting of the costs for computer software developed or obtained for internal use and of the hosting arrangement obtained for internal use. The standards state that costs incurred during the development stage may be capitalized. Capitalization of the costs should cease in the post-implementation stage. The FASB ASC subtopic 985-20 provides guidelines for the accounting of the costs incurred for software meant to be sold, leased or marketed. The standards state that costs incurred subsequent to the establishment of technological feasibility may be capitalized. Capitalization of the costs should cease when the software is available for general release to customers.

To illustrate this, we provide an example of Athena Health Inc. softwa-

[43] This is standard practice in empirical work based on firm-level data.

Figura 2.14: Software Capitalization of Athena Health

6. CAPITALIZED SOFTWARE COSTS

Capitalized software consisted of the following:

	As of December 31,	
	2017	2016
Capitalized internal-use software development costs	$ 113.9	$ 122.7
Acquired third-party software licenses for internal use	53.8	47.5
Total gross capitalized software for internal-use	167.7	170.2
Accumulated amortization	(74.8)	(82.9)
Capitalized internal-use software in process	46.8	38.5
Total capitalized software costs	$ 139.7	$ 125.8

Capitalized software amortization expense totaled $71.3 million, $73.5 million, and $53.4 million for the years ended December 31, 2017, 2016, and 2015, respectively. Future amortization expense for all capitalized software placed in service as of December 31, 2017 is estimated to be:

Years ending December 31,	Amount
2018	$ 50.5
2019	25.4
2020	10.7
2021	5.5
2022	0.8

re investments (see Figure 2.14). The company has capitalized software development costs of $113.9 million in 2017 and reports, external software acquisitions of $53 million in 2015.

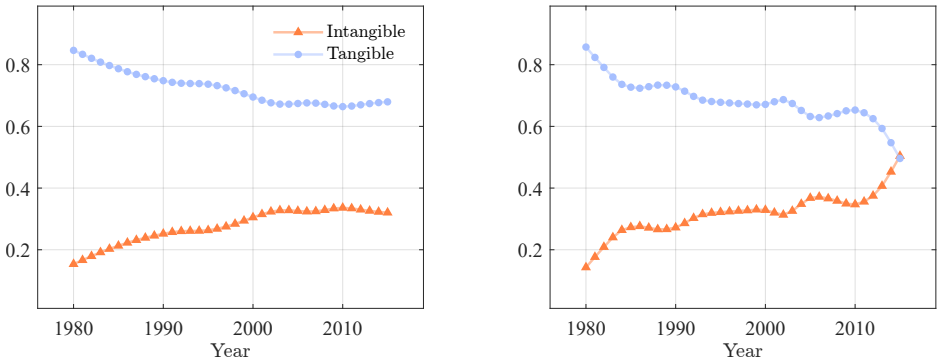
Software used in research and development is subject to the subtopic 730-10. In general, software that is purchased from others and used for research and development activities and that has alternative futures/uses should be capitalized and amortized as an intangible asset. However, the cost for software purchased from others for a particular research and development project and that has no alternative uses and therefore no separate economic value is considered a research and development cost and has to be expensed at the time it is incurred.

In any case, we would capture most of the intangible capital related to software in our measure through balance sheet intangible capital or through capitalized knowledge capital.

Additional Validations Firm-Level Intangible Capital

Here we compare some additional trends related to intangible capital investment, between aggregate data from the BEA and our measure from Compustat. Figure 2.15 compares the share of tangible capital investment into total investment and the share of intangible capital investment into total investment in both the BEA data and the Compustat data between 1980 and 2015. We can see that both data sources tell a similar story: in 1980, most of the investment was in tangible capital, whereas by 2015, tangible investment is roughly 70% of total investment in the BEA and 50% in Compustat. The two data sources tell similar stories, but they also show some discrepancies. In Compustat, the decline in the share of tangible capital investment of total investment is more pronounced; this could be due to, for instance, (i) undercapitalization of true IPP capital in BEA or (ii) selection of intangible-intensive firms in Compustat.

Figura 2.15: Investment Components Share

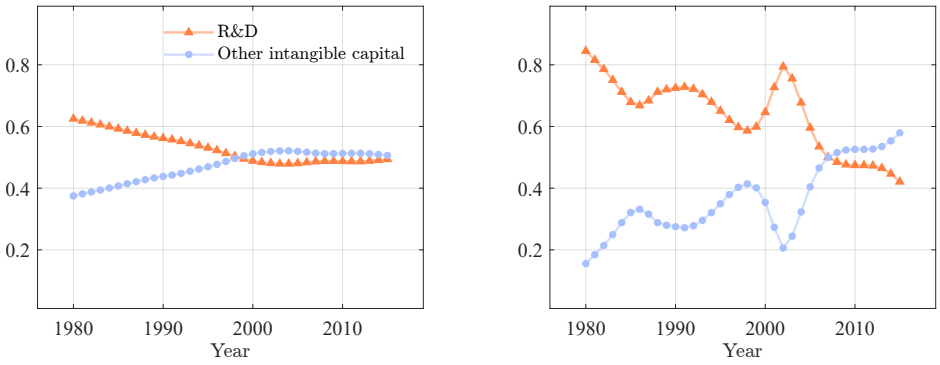


(a) BEA (b) Compustat

Note. The figure panels show the evolution of the share of tangible capital investment and of intangible capital investment over total investment in both BEA data and Compustat data for the period 1980-2015. The data are detrended with an HP filter with $\lambda = 6.25$.

Figure 2.16 shows the evolution of the different components of intangible capital investment in both the BEA and Compustat for the period 1980-2015. Again the two data sources show a similar tendency: in 1980, most of intangible capital investment was investment in research and de-

Figura 2.16: Intangible Capital Components Share

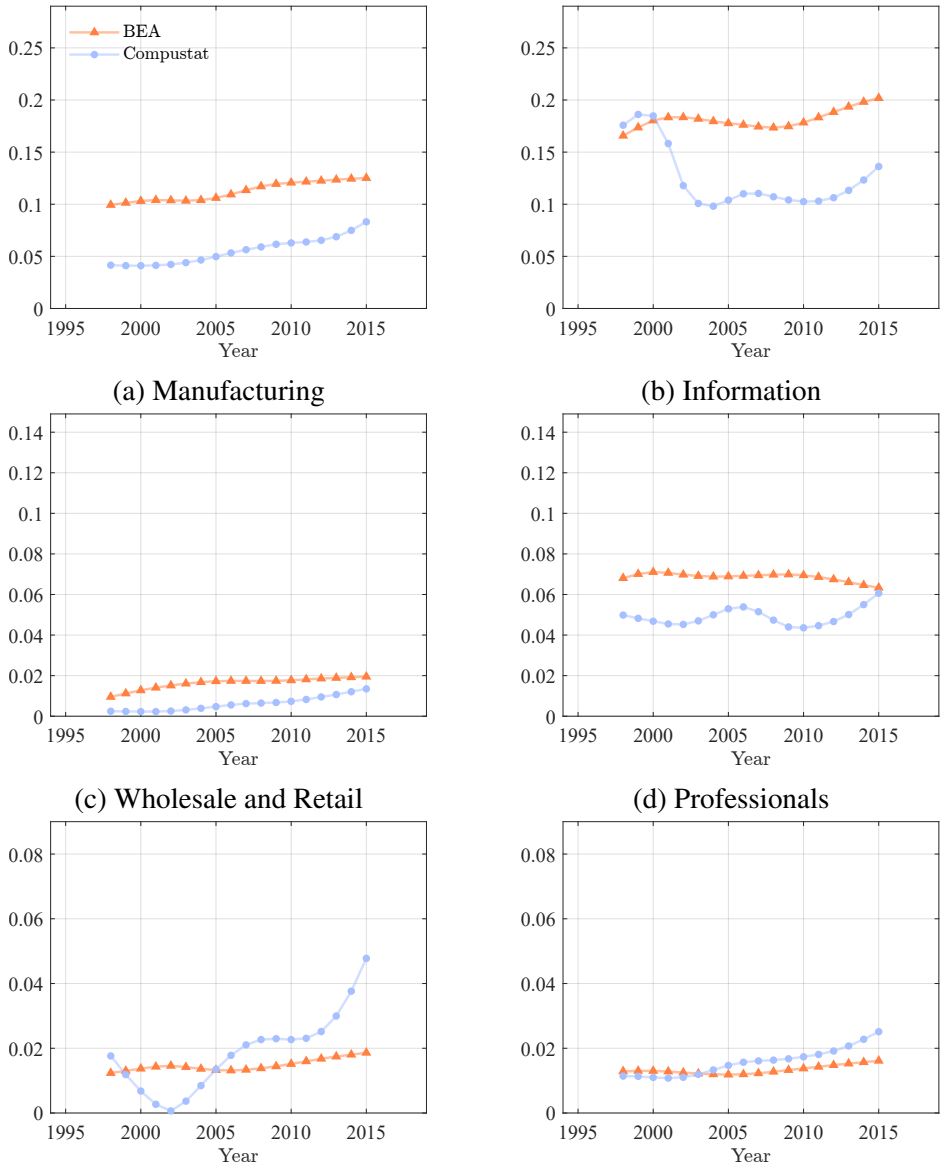


(a) BEA (b) Compustat

Note. The figure panels show the evolution of the share of knowledge capital investment (R&D) and other intangible capital investment (intangible capital investment different from R&D) over total intangible capital investment in both BEA data and in Compustat data for the period 1980-2015. The data are detrended with an HP filter with $\lambda = 6.25$.

velopment, whereas by 2015, investment in research and development accounts for less than 50% of total intangible capital investment.

Figura 2.17: Intangible Capital Components Share



(a) Manufacturing (b) Information

(c) Wholesale and Retail (d) Professionals

(e) Healthcare (f) Others

Note. The figure panels show the evolution of the intangible capital investment share across different sectors of the US economy for both BEA-KLEMS data and Compustat data between 1998 and 2015. The data are detrended with an HP filter with $\lambda = 6.25$.

Finally, in Figure 2.17 we compare the evolution of the intangible capital investment share across different sectors for both BEA data and Compustat data for the period 1998-2015. The sector-level intangible capital investment shares emerging from the Compustat data show trends similar to the one computed with the BEA data. However, we see some difference in the level within some sectors. It is difficult to know what the sources of these discrepancies are; overall, we conclude that our firm-level measure of intangibly capital does a reasonable good job in capturing the tendencies that are present in the aggregate data.

2.9.2 Production Function Estimation

To estimate the firm-level production function, we follow De Loecker et al. (2020) and use two main approaches: (i) the control function approach and (ii) the cost shares approach. Both of these approaches are popular methods used to estimate firm-level production functions. Here we review the two methodologies, emphasizing their virtues and limitations.

Ackerberg-Caves-Frazer

The control function approach has been pioneered by Olley and Pakes (1996) and developed further by Levinsohn and Petrin (2003) and Ackerberg et al. (2015). The main insight from this literature is that firm-level unobservable productivity can be proxied by some variable expenditure.

To overcome some of the criticism emphasized in Gandhi et al. (2020), we work with a structural value-added specification, as in Ackerberg et al. (2015) and De Loecker and Scott (2016), given by

$$\mathcal{Q}_{ft} = \min\left\{K_{T,ft}^{\alpha} K_{I,ft}^{\nu} L_{ft}^{1-\alpha-\nu} \exp(\omega_{ft} + \varepsilon_{ft}),\, \beta M_{ft}\right\}, \tag{2.21}$$

where $\mathcal{Q}_{ft}$ is output, $K_{T,ft}$ is tangible capital, $K_{I,ft}$ is intangible capital, L_{ft} is labor, ω_{ft} is log productivity, ε_{ft} is the error term, and M_{ft} is material. This structural value-added production function yields the following

first-order condition:

$$\mathcal{Q}_{ft} = K_{T,ft}^{\alpha} K_{I,ft}^{\nu} L_{ft}^{1-\alpha-\nu} \exp(\omega_{ft} + \varepsilon_{ft}), \tag{2.22}$$

justifying the regression of $\mathcal{Q}_{ft}$ on tangible capital, intangible capital, and labor while ignoring materials. A caveat is that, in theory, equation (2.22) may not be satisfied in certain situations. If both types of capital and labor are quasi-fixed and materials are a flexible input, then when output prices are sufficiently low relative to the price of materials, it will be better to set $M_{ft} = 0$ and not produce at all. However, given that our data only include actively producing firms, we assume that equation (2.22) always holds.[44] Therefore, under the specification in equation (2.21) the estimation of the firm-level production function reduces to

$$q_{ft} = \alpha k_{T,ft} + \nu k_{I,ft} + (1 - \alpha - \nu)\ell_{ft} + \omega_{ft} + \varepsilon_{ft}, \tag{2.23}$$

where $q_{ft} = \log(\mathcal{Q}_{ft})$, $k_{T,ft} = \log(K_{T,ft})$, $k_{I,ft} = \log(K_{I,ft})$, and $\ell_{ft} = \log(L_{ft})$. As usual, the main identification challenge to the production function estimation is the simultaneity bias induced by the unobserved time-varying firm-level productivity, ω_{ft}. We follow the control function literature, and in particular, Ackerberg et al. (2015) and De Loecker et al. (2020), to estimate the production function in (2.23) using a two-step approach based on the use of a control function for the productivity process. The identification relies on the observation that a firm's tangible capital investment demand is given by a policy function of the form $x_{T,ft} = x_T(k_{T,ft}, k_{I,ft}, \omega_{ft})$. Then, provided that the policy function is invertible, the productivity process can be proxied by a control function given by $\omega_{ft} = \omega(k_{T,ft}, k_{I,ft}, \omega_{ft})$ where $\omega(\cdot) = x_T^{-1}(\cdot)$.[45]

Therefore, in the first stage of this estimation procedure, we can clean the firm-level output value from measurement errors and unanticipated

[44] For a more detailed discussion on this issue, see Ackerberg et al. (2015).

[45] The assumptions needed to ensure the invertibility of the policy functions associated with a wide class of production functions have been discussed extensively by Pakes (1994), Olley and Pakes (1996), Levinsohn and Petrin (2003), and Ackerberg et al. (2015).

productivity shocks, regressing output on a polynomial of tangible capital, intangible capital, labor, and potential demand shifters, given by

$$q_{ft} = \mathcal{P}_t(k_{T,ft}, k_{I,ft}, \ell_{ft}, \mathbf{d}_{ft}) + \varepsilon_{ft}. \tag{2.24}$$

Then, in the second stage, using the estimate $\widehat{\mathcal{P}}_t$ from the previous stage, we can construct a measure of productivity that does not depend on the measurement error ε_{ft}, given by

$$\omega_{ft}(\alpha, \nu) = \widehat{\mathcal{P}}_t(k_{T,ft}, k_{I,ft}, \ell_{ft}, \mathbf{d}_{ft}) - \alpha k_{T,ft} - \nu k_{I,ft} - (1 - \alpha - \nu)\ell_{ft}. \tag{2.25}$$

Finally, taking advantage of the assumption that productivity follows an AR(1) process, it is possible to construct a measure of productivity innovations, given by

$$\xi(\alpha, \nu, \rho) = \omega_{ft}(\alpha, \nu) - \rho\omega_{ft-1}(\alpha, \nu). \tag{2.26}$$

Therefore, using the productivity innovations, we can construct a set of moment conditions to estimated the parameters of the production function, given by

$$\mathbb{E}(\xi(\alpha, \nu, \rho) \times \mathbf{z}_{ft}) = \mathbf{0}_{Z\times 1}, \tag{2.27}$$

where $\mathrm{Z} \geq 3$ and, under the assumption that firms react to unanticipated productivity shocks contemporaneously and that capital is predetermined, the set of admissible instruments is $\mathbf{z}_{ft} \in \{\ell_{ft}, k_{T,ft}, k_{I,ft}, \ell_{it-1}, k_{T,ft-1}, k_{I,ft-1}, \dots\}$.

Units. It is well known that most of the time, standard production data, such as Compustat, record revenues and expenditures rather than physical production and input used. In the presence of product differentiation (be it through physical attributes or location), an additional source of endogeneity presents itself through unobserved output and input prices.[46] This implies that, when bringing the model to the data, the structural value-added production function takes the following form:

$$q_{ft}+p_{ft} = \alpha(k_{T,ft}+p_t^T)+\nu(k_{I,ft}+p_t^I)+(1-\alpha-\nu)(\ell_{ft}+p_{ft}^{\ell})+\omega_{ft}+\varepsilon_{ft}, \tag{2.28}$$

[46] See De Loecker et al. (2016) for a recent treatment of these issues.

where p_{ft} is the output price, p_t^T is the common user cost of tangible capital, p_t^I is the common user cost of intangible capital, and p_{ft}^ℓ is the price of labor. This empirical specification produces the following structural error term:

$$\omega_{ft} + p_{ft} - \alpha p_t^T - \nu p_t^I - (1 - \alpha - \nu) p_{ft}^\ell. \tag{2.29}$$

We follow De Loecker et al. (2016) and let the wedge between the output and the input price (scaled by the output elasticity) be a function of the demand shifters and the productivity difference.[47] The inclusion in the control function of demand shifters $\mathbf{d}_{ft}$, constructed using measures of market shares as in De Loecker et al. (2020), should therefore capture the relevant output and input market forces that generate differences in the output and input price. As discussed in De Loecker et al. (2016), this is an exact control when output prices, conditional on productivity, reflect input price variation and when demand is of the (nested) logit form.

This is a second-best solution to address the aforementioned challenge in the estimation of the production function. Without more detailed data on output quantities, however, it is not possible to go beyond this second-best solution to the problem.

Cost Shares

The cost shares approach has been prominently adopted in Foster et al. (2008) and exploits the first-order conditions of the firm. To make fruitful use of the first-order conditions of the firm, two assumptions are needed, namely: (i) constant returns to scale in production and (ii) that all inputs are variable. Under these assumptions, the output elasticities can be calculated from cost shares. The cost shares of both inputs are defined as

$$\alpha = \text{med}\left\{ \frac{r_t^T k_{T,ft}}{w_{ft}\ell_{ft} + r_t^T k_{T,ft} + r_t^I k_{I,ft}} \right\} \quad \text{and} \quad \nu = \text{med}\left\{ \frac{r_t^I k_{I,ft}}{w_{ft}\ell_{ft} + r_t^T k_{T,ft} + r_t^I k_{I,ft}} \right\}, \tag{2.30}$$

[47] De Loecker et al. (2020) note that not observing output prices has the perhaps unexpected benefit that output price variation absorbs input price variation, thus eliminating part of the variation in the error term.

where $w_{ft}\ell_{ft}$ is the wage bill, $r_t^T k_{T,ft}$ is the rental cost of tangible capital, and $r_t^I k_{I,ft}$ is the rental cost of intangible capital. Therefore, an extra requirement to apply this method is the possibility of calculating the return on both types of capital, r_t^T and r_t^I.

2.9.3 Robustness Production Function Estimation

In this subsection of the appendix we explain the alternative specifications that we use to test the robustness of IBTC. Results are presented in Figure 2.3 in the main text.

Unconstrained Returns to Scale

To test the robustness of our results to a more flexible specification of returns to scale we estimate with the ACF approach the following production function:

$$q_{ft} = \alpha k_{T,ft} + \nu k_{I,ft} + \beta \ell_{ft} + \omega_{ft} + \varepsilon_{ft}, \tag{2.31}$$

where the only difference with equation 2.23 is that now returns to scale are unconstrained. Therefore, with this alternative specification, the set of moment conditions becomes

$$\mathbb{E}(\xi(\alpha, \nu, \beta, \rho) \times \mathbf{z}_{ft}) = \mathbf{0}_{Z\times 1}, \tag{2.32}$$

where $\mathrm{Z} \geq 4$.

Sector-Level Production Technology

One restrictive assumption of our benchmark specification is that the production technology is the same across all sectors. We relax this assumption by allowing the production technology to be sector specific. Effectively, this means that we estimate the following production function:

$$q_{ft} = \alpha_s k_{T,ft} + \nu_s k_{I,ft} + (1 - \alpha_s - \nu_s)\ell_{ft} + \omega_{ft} + \varepsilon_{ft}, \tag{2.33}$$

which is identical to the benchmark one except that now output elasticity is sector specific. Finally, with this specification, the average output elasticities will be computed using a sales-weighted average.

Translog Production Function

We also test the robustness of our results to a more flexible production function: the translog production. This production function approximates a CES production function up to a second-order. We choose a specification with constant returns to scale, given by

$$q_{ft} = \alpha k_{T,ft} + \nu k_{I,ft} + (1-\alpha-\nu)\ell_{ft} - \beta k_{T,ft}k_{I,ft} - \beta k_{T,ft}\ell_{ft} - \beta k_{I,ft}\ell_{ft} + \beta k_{T,ft}^2 + \beta k_{I,ft}^2 + \beta \ell_{ft}^2 + \omega_{ft} + \varepsilon_{ft}. \tag{2.34}$$

Therefore, with this alternative specification, the set of moment conditions becomes:

$$\mathbb{E}(\xi(\alpha,\nu,\beta,\rho) \times \mathbf{z}_{ft}) = \mathbf{0}_{Z\times 1}, \tag{2.35}$$

where Z ≥ 4. Finally, the endogenous output elasticities will be given by

$$\theta^T = \text{med}\big(\alpha - \beta k_{I,ft} - \beta \ell_{ft} + 2\beta k_{T,ft}\big), \tag{2.36}$$

$$\theta^I = \text{med}\big(\nu - \beta k_{T,ft} - \beta \ell_{ft} + 2\beta k_{I,ft}\big), \tag{2.37}$$

$$\theta^\ell = \text{med}\big(1 - \alpha - \nu - \beta k_{T,ft} - \beta k_{T,ft} + 2\beta \ell_{ft}\big). \tag{2.38}$$

2.9.4 Robustness Lumpiness

In this section of the appendix, we present some robustness analyses regarding the patterns of the investment rate distribution of intangible capital. In particular, we look at two additional dimensions that we neglected in the main analysis: the time dimension and the sector-level dimension.

Table 2.7 shows the moments of the investment rate distribution of intangible capital for different time frames: the period 1980-1999 and the period 2000-2015. The investment rate distribution of intangible capital does not show any qualitative difference over time. Overall, it seems that the most salient features of the distribution are stable over time, and hence they are not an artifact of the fact that intangible capital is rising over time and could be subject to an initial adoption phase.

Table 2.8 shows the moments of the investment rate distribution of intangible capital for different sectors. Here the analysis is complicated

Taula 2.7: Lumpiness by Period

Investment rates	1980-1999	2000-2015
Average	0.36	0.34
Positive fraction, $i > 1$	0.87	0.90
Negative fraction, $i < -1$	0.02	0.03
Inaction rate	0.11	0.07
Spike rate, $\|i\| > 20$	0.76	0.75
Positive spikes, $i > 20$	0.75	0.74
Negative spikes, $i < -20$	0.01	0.01
Standard deviation	0.29	0.30
Serial correlation, Corr(i_t, i_{t-1})	0.38	0.26

Note. This table shows the moments of the investment rate distribution of intangible and tangible capital. The statistics are computed for a balanced panel of 3,860 firm-year observations between 1980 and 1999 and for a balanced panel of 2,992 firm-year observations between 2000 and 2015.

Taula 2.8: Lumpiness by Sector

Investment rates	MIN	CON	MAN	TCU	WHO	RET	SRV
Average	0.22	0.15	0.38	0.35	0.14	0.16	0.30
Positive fraction, $i > 1$	0.65	0.79	0.91	0.93	0.69	0.81	0.87
Negative fraction, $i < -1$	0.06	0.14	0.02	0.04	0.14	0.09	0.06
Inaction rate	0.29	0.07	0.07	0.03	0.17	0.10	0.07
Spike rate, $\|i\| > 20$	0.49	0.50	0.81	0.67	0.43	0.40	0.67
Positive spikes, $i > 20$	0.48	0.38	0.80	0.64	0.38	0.35	0.64
Negative spikes, $i < -20$	0.01	0.12	0.01	0.03	0.05	0.05	0.03
Standard Deviation	0.30	0.30	0.29	0.33	0.29	0.29	0.32
Serial correlation, Corr(i_t, i_{t-1})	0.41	0.14	0.32	0.21	0.18	0.14	0.29

Note. This table shows the moments of the investment rate distribution of intangible capital across different sectors. The statistics are computed for a balanced panel between 1980 and 1990. MIN is the mining sector and has 209 firm-year observations. CON is the construction sector and has 99 firm-year observations. MAN is the manufacturing sector and has 4,455 firm-year observations. TCU is the transportation and public utilities sector and has 88 firm-year observations. WHO is the wholesale sector and has 176 firm-year observations. RET is the retail sector and has 176 firm-year observations. SRV is services sector and has 352 firm-year observations.

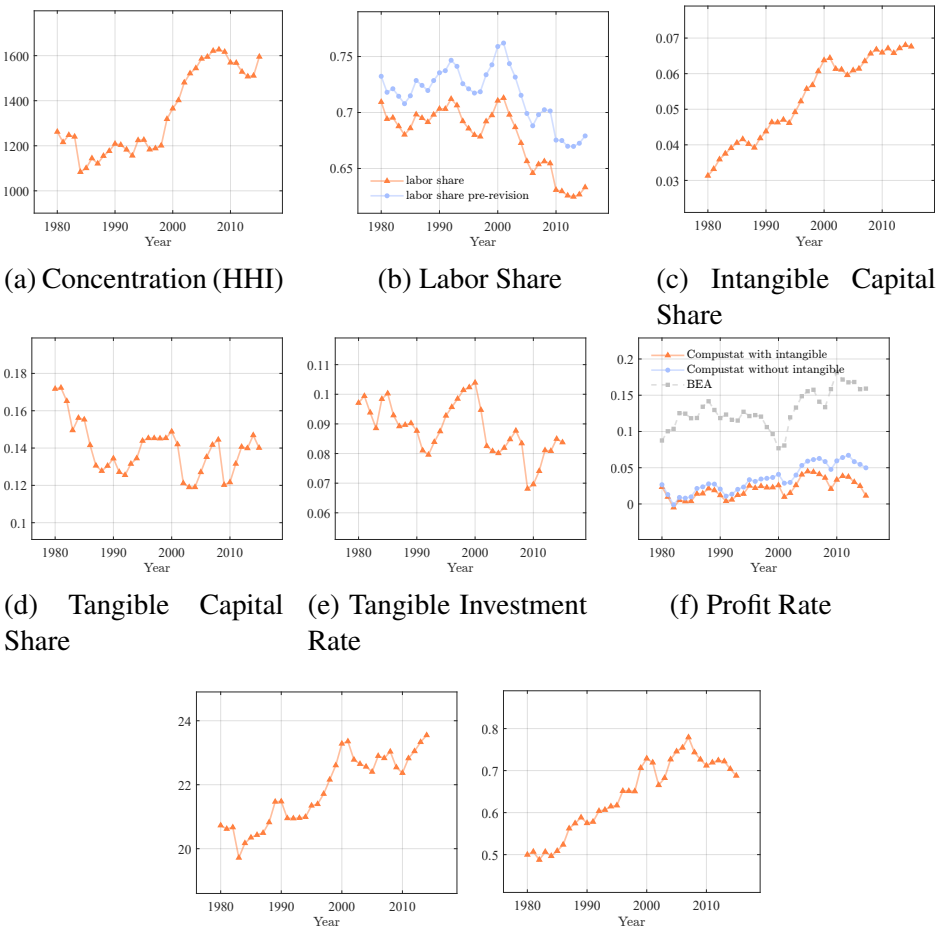
by the fact that at the sector level, the construction of a balanced panel sacrifices a lot of observation, leaving us with relatively small samples. Nonetheless, most of the salient characteristics of the investment rate distribution of intangible capital still seem to emerge from this analysis. This suggests that the investment rate distribution of intangible capital exhibits a fractal behavior overall.

Concluding, the high level of the lumpiness of the investment rate distribution of intangible capital documented in the main analysis seems robust to different time frameworks and different sectors. This suggests that this lumpiness has to come from intrinsic properties of the investment process in intangible capital.

2.9.5 Aggregate Trends

In this section, we present the evolution between 1980 and 2015 of the main trends of interest for the quantitative analysis. For the trends constructed with the Compustat data, we explain the measurement procedure; for the others, we just refer to main papers that document them. In particular, we look at: (i) the rise in concentration, (ii) the decline in the labor share, (iii) the rise in the intangible capital investment share, (iv) the decline in the tangible capital investment share, (v) the decline in the tangible capital investment rate, (vi) the rise in the profit rate, (vii) the rise in the average firm size, and (viii) the decline in the allocative efficiency of the economy, that is, the rise in the standard deviation of $TFPR$.

Figura 2.18: Aggregate Trends



(a) Concentration (HHI) (b) Labor Share (c) Intangible Capital Share

(d) Tangible Capital Share (e) Tangible Investment Rate (f) Profit Rate

(g) Avg. Firm Size (h) sd($TFPR$)

Note. Figure 2.18a replicates the evolution of the HHI index in Compustat, as documented by Grullon et al. (2019). Figure 2.18b shows the evolution of the labor share, pre- end post-revision, in the corporate non-financial sector, as reported in Koh et al. (2020). Figure 2.18c shows the evolution of the intangible capital investment share in the corporate non-financial sector, as reported in Koh et al. (2020). Figure 2.18d shows the evolution of the tangible capital investment share in the corporate non-financial sector, as reported in Koh et al. (2020). Figure 2.18e shows the evolution of the tangible capital investment rate, as reported by Crouzet and Eberly (2019). Figure 2.18f shows the evolution of the profit rate, as reported in De Loecker et al. (2020) and the profit rate adjusted for intangible capital. Figure 2.18g shows the evolution of the average firm size measured in number of employees from BDS data. Figure 2.18h shows the evolution of the standard deviation of $TFPR$ in Compustat.

We measure concentration using the HHI index, as in Grullon et al. (2019). In Compustat, the HHI index of a sector s is constructed as

$$HHI_{st} = \sum_f \left(\frac{\texttt{SALE}_{ft}}{\sum_f \texttt{SALE}_{ft}} \right). \tag{2.39}$$

Then, the aggregate concentration is simply the sales-weighted average of the sector-level concentrations.[48]

The firm-level profit rate, adjusted for intangible capital, is defined as

$$\pi_{ft} = \frac{\texttt{SALE}_{ft} - \texttt{COGS}_{ft} - (\texttt{XSGA}_{ft} - \texttt{XRD}_{ft}) - r_{T,t}k_{T,ft} - r_{I,t}k_{I,ft}}{\texttt{SALE}_{ft}}. \tag{2.40}$$

The construction of the user cost of both types of capital is described in Appendix 2.9.1. We drop `XRD` from `XSGA` to avoid double-counting research and development costs as they are both part of our measured intangible capital costs and our selling general and administrative costs. The standard unadjusted profit rate is instead defined as

$$\pi_{ft} = \frac{\texttt{SALE}_{ft} - \texttt{COGS}_{ft} - \texttt{XSGA}_{ft} - r_{T,t}k_{T,ft}}{\texttt{SALE}_{ft}}. \tag{2.41}$$

To obtain the aggregate profit rate, we use a sales-weighted average of both measures of the firm-level profit rate.

Finally, to measure the allocative efficiency of the US economy we measure the standard deviation of $TFPR$. We compute $TFPR$ as

$$TFPR_{ft} = \log \texttt{SALE}_{ft} - \alpha k_{T,ft} - \nu k_{I,ft} - (1-\alpha-\nu) \log \texttt{EMP}_{ft}, \tag{2.42}$$

where α and ν are the estimates from Section 2.3.1. Then, our measure of allocative efficiency is just the dispersion in $TFPR$ over the different years. Figure 2.18h shows the evolution of these trends between 1980 and 2015.

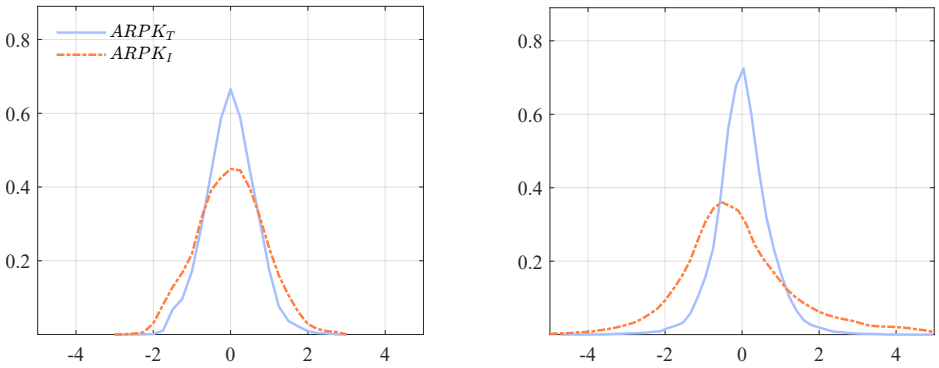
[48] We follow Grullon et al. (2019) and use the 2-digit NAICS level as the definition of sector-level.

2.10 Quantitative Appendix

2.10.1 Additional Comparisons between Model and Data in 1980

Here, we compare the distribution of the average product of tangible capital, $ARPK_T$, and the distribution of the average product of intangible capital, $ARPK_I$, from the model with the ones from the data for the 1980s.

Figura 2.19: Average Product of Tangible and Intangible Capital



(a) Model (b) Data

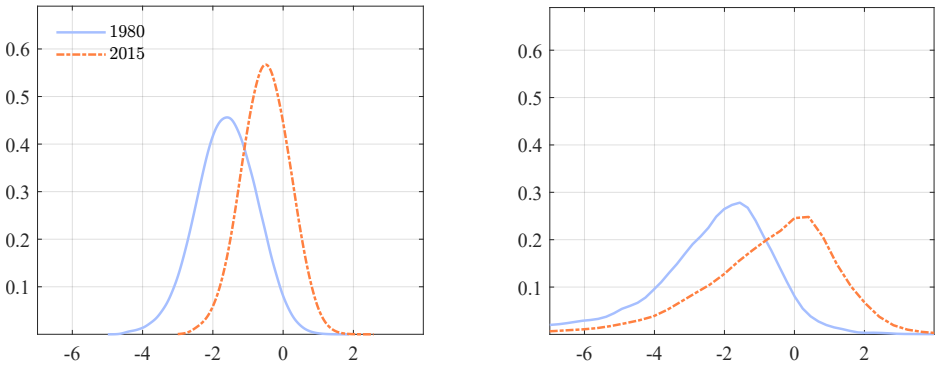
Note. Figure 2.19a shows the distribution of $ARPK_T$ (light blue solid line) and $ARPK_I$ (dashed orange line) from the model. Figure 2.19b shows the same distributions from the data. All distributions are demeaned.

Figure 2.19 shows the distributions. The distributions implied from the model capture well the main features of the distributions in the data. In particular, the model is able to capture the excess dispersion in the distribution of $ARPK_I$ relative to the distribution of $ARPK_T$. This is because intangible capital faces higher distortions from the presence of higher adjustment costs.

2.10.2 Additional Comparisons between Model and Data over Time

In this section, we document two additional implications of the model over time and compare them with the data. In particular, we look at the distribution of intangible intensity, defined as the ratio of intangible capital to the labor bill, and at the distribution of $TFPR$.

Figura 2.20: Log Intangible Intensity



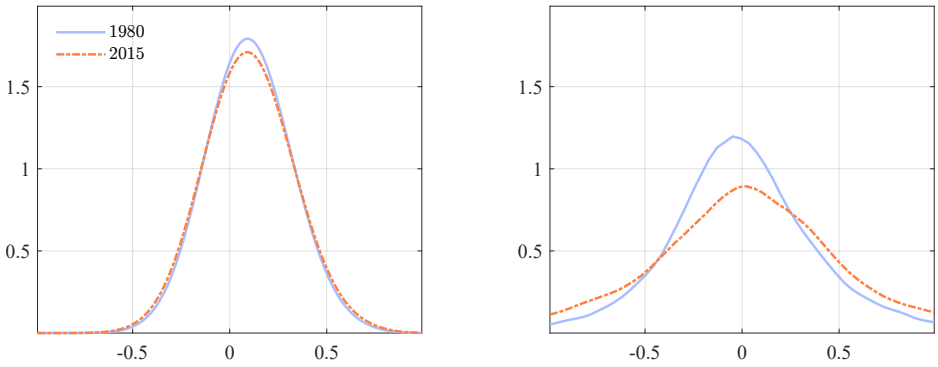
(a) Model (b) Data

Note. Figure 2.20a shows the distribution of log intangible intensity both in the 1980 (light blue solid line) and in 2015 (dashed orange line) from the model. Figure 2.20b shows the same distributions form the data.

Figure 2.20 shows the evolution over time of the distribution of intangible intensity in both the model and the data. Overall, despite some qualitative differences, both the model and the data show a shift toward the right in the distribution of intangible intensities, highlighting the fact that firms on average are using more intangible capital relative to labor.

Figure 2.21a shows the evolution, in both the model and the data, of the distribution of $TFPR$. In both the model and the data the distribution of $TFPR$ is more dispersed in 2015, highlighting a decline in allocative efficiency. This, as emphasized in the main text, is a result of firms relying more on an input that is highly dispersed because of technological constraints, which hinders a fast reallocation of inputs toward high marginal product firms.

Figura 2.21: Total Factor Productivity Revenue



(a) Model (b) Data

Note. Figure 2.21a shows the distribution of $TFPR$ in 1980 (light blue solid line) and 2015 (dashed orange line) from the model. Figure 2.21b shows the same distributions from the data. All distributions are demeaned.

2.10.3 Additional Robustness

Table 2.9 shows the parameters from the robustness exercise in which we reestimate the adjustment costs as presented in Section 2.7.2. Two things have changed relative to the calibration for the 1980 steady state. First, the intangible capital share has increased to 0.12 as we are estimating the economy in 2015. Second, all the adjustment costs associated with the investment process of both types of capital have now been changed to match the moments from the later part of our sample. The remaining parameters not associated with the production technology or with the adjustment costs are left the same as in the 1980 steady state to facilitate a comparison across the steady state and to pin down the fundamental forces underlying the results.

Taula 2.9: Parameters and Moments

Fixed	Value	Description
R	1.05	Annual interest rate
δ_T	0.07	Annual depreciation rate tangible capital
δ_I	0.29	Annual depreciation rate intangible capital
α	0.28	Tangible capital share
ν	0.12	Intangible capital share
ω	0.90	Returns to scale
ρ_z	0.90	Autocorrelation idiosyncratic productivity
σ_z	0.20	Standard deviation idiosyncratic productivity
c_e	3·e-4	Fixed to 1980 SS
c_f	2.540	Fixed to 1980 SS
η	2.025	Fixed to 1980 SS
m	6.2·e-3	Fixed to 1980 SS

Fitted	Value	Description	Moments	Model	Data
γ_T	0.012	Convex adj. cost k_T	corr$(x_{T,ft}, x_{T,ft-1})$	0.16	0.16
γ_I	0.060	Convex adj. cost k_I	corr$(x_{I,ft}, x_{I,ft-1})$	0.26	0.27
f_p	2.3·e-3	Fixed adj. cost k_T	Inaction rate: x_T	0.03	0.03
f_i	0.017	Fixed adj. cost k_I	Inaction rate: x_I	0.07	0.07

Capítol 3

HETEROGENEOUS MARKUPS CYCLICALITY AND MONETARY POLICY

Joint with Marta Morazzoni and Danila Smirnov

3.1 Introduction

Far for being a closed topic of investigation, the discussion around the cyclicality of the aggregate markup and its response to monetary policy shocks still fosters a significant volume of macroeconomic research. Parallel to that, recent contributions have brought attention on companies' heterogeneous market power, as the availability of firm-level datasets has made it easier to estimate markups from balance sheet data. However, the empirical evidence of the heterogeneity in the behavior of firm-level markups after interest rate movements is scarce, and any related quantitative analysis has not yet been provided. This project is a first step towards filling this gap. In particular, we document crucial differences in the response of markups to monetary policy shocks by firm age, and assess their macroeconomic relevance into a novel New Keynesian framework enriched with firms' heterogeneity, demand accumulation and endogenous

markups that evolve along firms' life-cycle.

We begin by estimating the behavior of markups at the company level conditional on interest rate movements, and document a significant degree of heterogeneity across old and young firms. Combining together quarterly data from Compustat with two different and exogenously identified series of monetary policy (hereafter MP) shocks for the US, we employ state-of-the-art local projection techniques to establish that the markups of firms above the median age respond more countercyclically to negative MP shocks, while for young firms the response is either mildly procyclical or insignificant. Controlling for commonly-used measures of aggregate economic activity and horse-racing our regression specifications with other firm-level characteristics, we are able to confirm that corporate age in particular influences the differential trajectory of markups upon a negative change in the interest rate. Moreover, we provide evidence that this result could indeed be related to a latent process of demand (or customer base) accumulation, for which dominant firms that are more established in their markets may have to change by less their prices in response to MP shocks, thereby leading to the stronger countercyclical response in old firms' markups documented in the data.

Next, we embed our findings into a New Keynesian (NK) framework that we enrich with firm heterogeneity, demand accumulation and endogenous markups. The model features imperfect competition among heterogeneous intermediate firms that produce using labor and choose prices to maximize profits subject to price adjustment costs à la Rotemberg. New firms can entry every period, while the exit of incumbent firms is exogenous. Our framework presents therefore two main characteristics: on the one hand, intermediate firms face a process of demand accumulation, characterized by some persistence and idiosyncratic shocks, along with a long-run mean that allows for the demand faced by companies to increase along their life-cycle. On the other hand, we assume that the final good producer combines together the intermediate inputs by means of a Kimball aggregator. As in Klenow and Willis (2016) for example, this specific choice introduces in a tractable way endogenous markups in the model, as the elasticity of substitution across intermediate goods become decrea-

sing in their relative quantity. Dominant firms will face lower elasticities of demand and, since demand is accumulated with age, older businesses will hence be able to charge higher markups.

The model is then calibrated on the US economy, following standard strategies in the literature and making use of the richness of Compustat data. In particular, the validation analysis shows that our quantitative framework is able to replicate several untargeted features of the data, such as the increasing profile of markups along the life-cycle of the firms, the fat right tail in the distribution of markups, and the growth rates of sales and employment. Moreover, the model can get realistic steady state distributions of businesses and employment shares by firm age, and replicate the elasticity of wages to firms' sales shares that we estimate in the data. This latter moment is tightly linked to the fact that dominant (and hence old) companies can increase their profits by cutting quantities and raising prices, thereby suppressing labor demand and hence wages in equilibrium.

Importantly, our NK framework enriched with firm heterogeneity, demand accumulation and endogenous markups can deliver the differential response of markups by firm age that we document in Compustat. As previously mentioned, old firms in our model economy face a lower passthrough from costs to prices due to the presence of the Kimball aggregator. When hit by a contractionary MP shock that decreases wages and puts a downward pressure on prices, dominant firms can cut prices by relatively less compared to young ones. Since markups depend on the ratio between firm prices and marginal costs, this mechanism is in turn responsible for the stronger countercyclical response in old firms' markups. In particular, we can match up to 20% of the empirically estimated relative difference in old and young firms' markups responses to a negative MP shock. In our analysis, we also show that the differential response of old firms in the model can be quantitatively decomposed to highlight the contribution of changes in aggregate variables to the overall general equilibrium impact on markups. In particular, the movements in real wages generated by a negative shock to the interest rate are found to be key in shaping the differential behavior of dominant firms' markups.

Finally, we conclude our quantitative analysis with an investigation of the shock amplification mechanisms at play in our framework, comparing our set up to a standard one-firm NK model with price rigidities. Both the presence of the Kimball aggregator and the heterogeneity of firms are shown to affect the way and extent to which MP shocks transmit in the economy, with output decreasing on average by roughly 20 percentage points (p.p.) more after a negative movement in the interest rate. Focusing on the role of the Kimball aggregator, since intermediate firms – especially old ones – temper their price drops after a negative MP shock due to the increase in their desired markup, the shock itself propagates more through quantities than through prices in our set up as opposed to the standard constant elasticity NK framework. At the same time, the Kimball aggregator alone is not sufficient to generate the observed amplification of MP shocks, as its effects on the elasticity of demand faced by firms kick in when firms are indeed heterogeneous and hence have a different passthrough from costs to prices. Firm heterogeneity is therefore key in affecting and amplifying the movements in the macroeconomic aggregates in the economy following a negative MP shock.

We see the contribution of this paper as twofold: on the one hand and to the best of our knowledge, we bring novel evidence on the remarkable heterogeneity in the response of firm markups to MP shocks based on corporate age. Specifically, while several empirical macroeconomic studies have focused on the different response of investment conditional on movements in the interest rates, we take a different perspective and explore the heterogeneous behavior of markups, the most direct measure of firms' market power. On the other hand, enriching a NK model with firm heterogeneity, demand accumulation and endogenous markups, we attempt to quantitatively study the role of firms' life-cycle in shaping the differential response of markups to changes in the interest rates, and then analyse how aggregate shocks propagate (and get amplified) in our model economy.

Related Literature. Our work builds on several macroeconomic contributions to the study of markups cyclicality. With respect to papers that have analysed the *aggregate* markup (see Gali et al. (2007), Hall (1988),

Bils et al. (2018) and Nekarda and Ramey (2020)), we focus on the heterogeneous response of *firm-level* markups to changes in the interest rate, both from an empirical and quantitative point of view. Second, in comparison to recent research on firm-level markups by Hong (2017), Burstein et al. (2020), Meier and Reinelt (2020) and Alati (2020), we do no investigate markups response to business cycle movements, but rather markups behavior conditional on monetary policy.

On the other hand, we attempt to contribute to the theoretical and quantitative macroeconomic literature that has started incorporating micro-level heterogeneity into NK frameworks and understand its implications for the transmission of monetary policy. Recent studies in this field have focused on how household-level heterogeneity affects the consumption channel of monetary policy (see, for example, McKay et al. (2016), Kaplan et al. (2018), Auclert (2019), or Wong (2019)). More in line with the spirit of Ottonello and Winberry (2020)'s investigation of firm investment, we explore the role of firm-level heterogeneity in determining differences in the response of markups to monetary policy shocks. In so doing, we also relate our work to several analyses of supply-side heterogeneities in NK set ups, such as studies on price-setting behavior (see Golosov and Lucas (2007)), market power (see Klenow and Willis (2016) and Mongey (2017)), and product life-cycle (see Bilbiie et al. (2007) and Bilbiie et al. (2012)). With respect to these papers, we present a model of firm life-cycle behavior in order to examine the endogenous response of markups to monetary policy by firm age.

Our work is also related to Gilchrist et al. (2017), who study how financial distortions can create an incentive for firms to raise prices in response to adverse financial or demand shocks. While in Gilchrist et al. (2017) the rise in markups reflects firms' decisions to preserve internal liquidity and avoid accessing external finance, the endogenous response to markups in our set up is related to the differential demand elasticities faced by firms in their life-cycle. In this sense, we see our work as closely related to Baqaee et al. (2021) from a theoretical perspective: the authors explore the first-order effect on aggregate TFP caused by the reallocation of resources triggered by a demand shock across firms with non-

uniform markups. While we answer a different research question, also in our model dominant firms tend to have both higher markups and lower pass-through from marginal costs to prices. When faced with an increase (decrease) in nominal marginal costs, high-markup firms raise (lower) their prices by less than low-markup firms in order to remain competitive.

Finally, our work is related to the macroeconomic literature pioneered by Gertler and Gilchrist (1994) that empirically documents how the effect of monetary policy can vary across firms of different characteristics. To the best of our knowledge, existing studies in this area have focused on firm-level investment, and assess how firm default risk (Ottonello and Winberry (2020)), liquidity (Jeenas (2019)) or age (Cloyne et al. (2018)) may shape the response of investment to monetary policy shocks. In a similar spirit, Fabiani et al. (2020) examine how monetary policy can influence the maturity structure of corporate debt. We also use state-of-the-art local projection techniques in our empirical analysis and explore the heterogeneous response of firms' markups to monetary policy shocks. In fact, our core contribution is to document that old firms' markups react more countercyclically to contractionary interest rate shocks than young ones, and then embed our finding into a heterogeneous firms NK framework augmented with endogenous markups formation.

The paper is organized as follows: in Section 3.2, we report and discuss the empirical evidence on the heterogeneous cyclicality of firms' markups after monetary policy shocks. Section 3.3 lays down our theoretical framework, characterized by heterogeneous firms in a NK setting with endogenous markups. Then, in Section 3.4, we illustrate the calibration and fit of the model, while in Section 3.5 we present steady state results and firm-level impulse responses to monetary policy shocks, and also discuss amplification mechanisms. In Section 3.6 we finally conclude and present the way ahead.

3.2 Empirical Analysis

In what follows, we study the heterogeneous cyclicality of firms' markups in response to monetary policy shocks. We begin by describing the sample of US firms and the monetary policy shock series on which we draw our evidence, and then illustrate how we estimate markups at the firm-level. Secondly, we document that old firms show a more countercyclical markups response after a monetary policy tightening. Finally, we briefly analyse the behavior of markups over firms' life-cycle and motivate why firm age could be a source of heterogeneity in markups responses to demand shocks.

3.2.1 Sample Construction

As previously mentioned, we make use of firm-level data from Compustat, which contains quarterly balance sheet information for North-American listed companies between 1975 and 2016. Compustat constitutes a panel of US corporations that is sufficiently high-frequency to be used to study monetary policy, and long enough to exploit within-firm variation. However, it comes at the expenses of representing the universe of publicly-listed incorporated firms only, even though these companies are estimated to make up for 30% of private sector employment. In terms of coverage, Compustat reports details on firm performance indicators and outcomes, including sales, liquid assets, financing sources, total assets, and production costs. It also reports the industry sector (SIC codes) where the business operates and firm age, which is the crucial dimension of heterogeneity in our analysis. Importantly, the age variable contained in the original dataset counts the years since incorporation, but we provide a robustness considering the establishment year for each firm in our sample.

Following standard practices in the literature, we restrict our attention to firms that are incorporated in the US and our final sample excludes utilities companies (SIC codes 4900-4999), financial entities (SIC codes 6000-6999), as well as corporations for which the industry code, or the information on sales, assets and production costs is missing. Whenever

Taula 3.1: Summary Statistics

	Sales	Cogs	Assets	Leverage	Liquidity	Age
mean	447.69	303.17	4919.69	0.45	0.17	9.46
p25	6.06	3.31	37.83	0.04	0.02	4
p50	31.01	17.18	229.50	0.18	0.07	8
p75	164.58	100.60	1118.33	0.39	0.22	14
N	715,874	715,874	685,784	641,316	683,696	715,874

Notes: the first three columns are measured in millions of real 2012 $, while column (4) and (5) are ratios and column (6) is measured in years. *Cogs* is the cost of good sold, which includes production expenditures.

applicable, we deflate variables using a GDP-deflator from the NIPA tables. 3.1 reports summary statistics for the variables of interest.

Our final sample of firms is then merged with two different interest rates datasets: first, we take the quarterly monetary policy shock series from Gürkaynak et al. (2005), who build a measure of interest rate surprises based on the % change in the FED Funds Futures rate in 30-minute windows around the policy announcement. Secondly, we also and primarily make use of the quarterly monetary policy shocks from Jarociński and Karadi (2020), a pure interest rate surprises series that removes from the estimation any component attributed to the provision of private FED information on the state of the economy to private agents through policy announcements. The common identifying assumption on the exogeneity in the variation of the policy rate is that nothing else occurs within this 30-minutes time window that could drive both private sector behavior and monetary policy decisions. Both series are available for the years and quarters between 1990Q1 and 2016Q4.

Before describing the estimation of markups at the firm level, we briefly recap on other important variables that we further employ as controls in our regressions. First, using balance sheet data, we compute the leverage and the holdings of liquid assets for the companies in our sample. With respect to the former, we take the ratio of corporate total debt divided by total assets in each period, both measured at book values and where debt

is the sum of short term and long term debt. Parallel to that and to provide a measure of corporate liquidity, we compute the ratio of cash and short-term investments to total assets. Our main regression specifications also include firm size as a control, which is measured as the log of total assets (at book value).[1] Finally, we complement our firm-level data with general indicators of economic activity at quarterly level. In particular, we include the GDP growth rate, the Consumer Price Index (CPI) growth rate, the Excess Bond Premium (EBP), and the 1-Year Treasury rate change, all taken from the Federal Reserve of St.Louis (FRED) series.[2]

3.2.2 Markups Estimation

Firm-level markups are a common measure of whether companies are able to set their prices above marginal costs. To estimate them, we follow recent works by De Loecker and Warzynski (2012) and De Loecker et al. (2020), which are based on the production function approach pioneered by Hall (1988) on industry-level data. Their estimation strategy is grounded on firm's optimizing behavior with respect to production costs-minimization, and delivers an estimate of markups at the firm-level without specifying an explicit demand system. In fact, consider a firm i that employs a production technology given by:

$$Q_{i,t} = F_{i,t}(\boldsymbol{X}_{i,t}, K_{i,t}, \omega_{i,t})$$

where $\boldsymbol{X}$ is a vector of variable inputs, K is the predetermined input and ω is firm-specific productivity. The cost minimization problem for each producer can be hence expressed as follows:

$$\min_{\{\boldsymbol{X}_{i,t}, K_{i,t}\}} \{\boldsymbol{P}'_{i,t}\boldsymbol{X}_{i,t} + R_t K_{i,t} + \lambda_{i,t}(Q_{i,t} - Q(\cdot)\}$$

where $\boldsymbol{P}_{i,t}$ is the vector of prices for variable inputs, R_t is the price of the predetermined input, and $\lambda_{i,t}$ is the Lagrangian multiplier associated

[1] To eliminate seasonality, variables can be measured as the rolling means in the previous 4 quarters as in Jeenas (2019).

[2] Even if the identified monetary policy shock series are exogenous, macro controls are typically included for robustness.

to the firm's cost minimization problem. One can then compute the first order condition (FOC) for a generic variable input $X^{\nu} \in \boldsymbol{X}$, which is given by:

$$\frac{\partial \mathcal{L}(\cdot)}{\partial X^{\nu}_{i,t}} = P^{\nu}_{i,t} - \lambda_{i,t} \frac{\partial Q(\cdot)}{\partial X^{\nu}_{i,t}} = 0 \tag{3.1}$$

Notice that the Lagrangian multiplier $\lambda_{i,t}$ can be also interpreted as the marginal cost of producing at a given level of output. 3.1 can be further rearranged as:

$$\frac{\partial Q(\cdot)}{\partial X^{\nu}_{i,t}} \frac{X^{\nu}_{i,t}}{Q_{i,t}} = \frac{1}{\lambda_{i,t}} \frac{P^{\nu}_{i,t} X^{\nu}_{i,t}}{Q_{i,t}}$$

Defining the markup as price over marginal costs, $\mu_{i,t} \equiv \frac{P_{i,t}}{\lambda_{i,t}}$, it is possible to rearrange the FOC for a generic variable input $X^{\nu} \in \boldsymbol{X}$ such that it yields:

$$\mu_{i,t} = \theta^{\nu}_{s,t} \frac{P_{i,t} Q_{i,t}}{P^{\nu}_{i,t} X^{\nu}_{i,t}} \tag{3.2}$$

where $\theta^{\nu}_{s,t}$ is the elasticity of output with respect to the variable input X^{ν}. The computation of markups can hence be implemented using firms' financial statements only. To estimate this theoretical expression in Compustat, we make use of both sales and cost of good sold data for each firm and in each quarter, which map to the denominator and numerator of 3.2 according to:

$$\hat{\mu}_{i,t} = \widehat{\theta}^{\nu}_{s,t} \frac{\texttt{Sales}_{i,t}}{\texttt{Cogs}_{i,t}}$$

where we use the estimates of the sectoral output-input elasticity $\hat{\theta}^{\nu}_{s,t}$ from De Loecker et al. (2020).

3.2.3 Heterogeneous Markups Cyclicality

We then proceed to use Compustat quarterly balance-sheet data to investigate cross-sectional differences in the response of markups to interest

rate policies. The main goal of our analysis is to estimate how firm i's markup $\mu_{i,t+h}$, at horizon $h \geq 0$, behaves in response to a monetary policy shock at time t, conditional on firm i's age just before the shock. To this end, we borrow the empirical strategies of Jeenas (2019) and Ottonello and Winberry (2020), and use a panel version of the Jordà (2005)'s local projections (hereafter: LP) to regress the cumulative difference in firm markups at different horizons on the interaction term between firm age at time $t-1$ and the monetary policy shock at time t, alongside a set of control variables. This flexible specification enables us to estimate impulse response functions on our firm-level panel data using the identified monetary shocks as instruments for changes in the policy interest rate. In particular, we estimate by OLS the following set of equations:

$$\begin{aligned}\Delta_h \log \mu_{i,t+h} = \sum_{x \in \mathcal{X}} \left(\alpha_{x,h} + \beta_{x,h} \Delta Y_{t-1} + \sum_{k=-\kappa}^{h} \gamma_{x,h}^{k} \varepsilon_{t+k}^{m} \right) \times 1_{i \in \mathcal{I}^x} \\ + \sum_{\ell=1}^{L} \boldsymbol{\delta_h'} \boldsymbol{X_{i,t-\ell}} + \varphi_{i,h} + \varphi_{s,t,h} + \boldsymbol{\vartheta_h t} + u_{i,t+h}\end{aligned} \tag{3.3}$$

with horizons $h = 0, 1, \ldots, H$ and $H = 20$ quarters. The dependent variable is the cumulative change in markups for any firm i at horizon h, given by:

$$\Delta_h \log \mu_{i,t+h} \equiv \log \mu_{i,t+h} - \log \mu_{i,t-1}$$

Focusing on our regressors, $1_{i \in \mathcal{I}^x}$ is an indicator that takes a value of 1 if $i \in \mathcal{I}^x$, namely if the firm i is above the $median$ in one or more dimensions of the vector $\mathcal{X} = \{age, leverage, liquidity, assets\}$. The main coefficient of interest is $\gamma_{age,h}$, which captures the relative response of old companies (compared to young ones)[3] to a variation in the FED short-term policy rate.[4] Note that we horse-race our main regressor – corporate

[3] We note that the median corporate age – since incorporation – in our Compustat sample is 8 years old.

[4] In our preferred specification, we therefore adopt a non-parametric estimation approach by using dummies instead of linear interactions. We show robustness checks following instead a parametric approach at the end of the section.

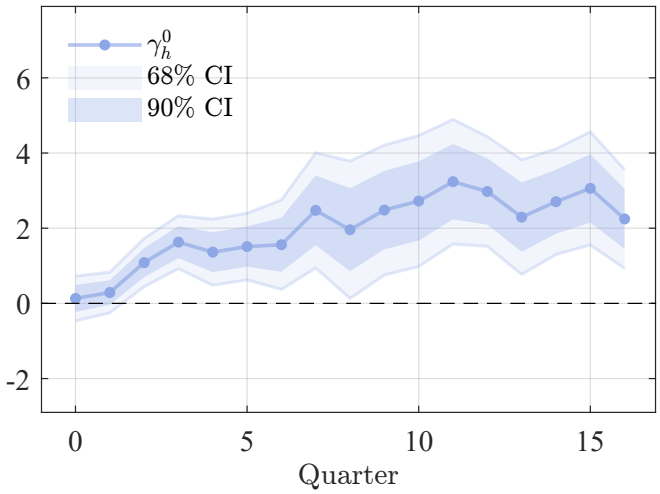
age – against other layers of firm heterogeneity typically studied by the literature, such as size, leverage and liquidity (see Jeenas (2019)). We also interact the vector of firm-level regressors x with ΔY_{t-1}, which is the previous quarter's GDP growth, to control for the differential sensitivity of firm markups to the business cycle, following Ottonello and Winberry (2020).

Furthermore, $\boldsymbol{\varepsilon_t^m} \equiv \sum_{k=-4}^{h} \varepsilon_{t+k}^m$ is the series of monetary policy shocks from Jarociński and Karadi (2020), while $\boldsymbol{X_{i,t}}$ is a vector of controls that includes firm-level variables such as sales growth and overhead costs to sales, and macro-level controls like GDP and CPI growth, 1-year treasury rate change, EBP, and fiscal quarter dummies to account for seasonality. Following standard practices in the literature, we include control variable lags (up to 4) and measure the controls and the variables in vector x at the end of the quarter before the arrival of the monetary policy shock to ensure exogeneity with respect to it. We then allow for firm ($\varphi_{i,h}$), and sector-time ($\varphi_{s,t,h}$) fixed effects (FE) to control for the unobserved time-invariant heterogeneity at the level of the firm and to absorb time-varying shocks that are common to all firms in a given industry. We also include a linear and quadratic trend ($\boldsymbol{\vartheta_h t}$). Saturating the regression in 3.3 with these FE implies that, first, our coefficients of interest are identified by within-firm variation over time, namely by changes in the markups response of an otherwise identical firm when it is old compared to when it was young. Secondly, our estimation fully exploits the cross-sectional variation across firms in a given industry. Finally, we cluster the standard errors at the firm and quarter level to account for correlation in the error term.[5]

As mentioned in the previous paragraph, our main coefficient of interest is given by $\gamma_{age,h}$, which captures the differential h-quarter growth

[5]Clustering at the firm level allows for a fully flexible dependence in the error terms across time within each company. Clustering by time is necessary whenever firm-level shocks are correlated within a quarter and if this effect may go potentially above the co-movement caused by industry-level shocks already captured by the sector-quarter dummies. We note that the confidence intervals on estimates would be significantly lower without clustering at the quarter-level.

Figura 3.1: Markups Response to a Monetary Policy Tightening



Notes: Within each quarter, firms' markups are winsorized at the 1% and 99% cutoff, to avoid any outlier to drive our results. Confidence intervals at 90% and 68%, which approximates one standard deviation.

of markups for firms above the median age after a 25 basis point hike in the interest rate (which corresponds to a rise of a quarter of a percent). Since we are including quarter FE and hence controlling for the time variation of the shock, the coefficient $\gamma_{age,h}$ can precisely identify the excess cyclicality of older firms' markups. In particular, 3.1 reports the impulse response function obtained from the OLS-estimation of $\gamma_{age,h}$ in 3.3, along with standard confidence intervals around the point estimates. The magnitude of the $\gamma_{age,h}$ coefficient suggests that being above the median age before a contractionary monetary policy shock hits can imply up to a +3% statistically significant difference in the subsequent response of firm markups.

Interestingly, older firms' markups respond more countercyclically to a monetary policy tightening, with the cumulative effect lasting for at least 16 quarters after the shock. It is important to stress that we control for firm's FE and for other crucial determinants of between-firms heterogeneity studied by the literature, namely size, leverage and liquidity. Yet,

none of the interactions between these three firm-level variables and the monetary policy shock are statistically significant predictors of markups heterogeneous response to interest rate changes, as further reported in the Appendix. Moreover, as in Cloyne et al. (2018), we note that firm age is pre-determined and cannot vary as a result of changes in monetary policy. In contrast, size, leverage and liquidity endogenously respond to shocks and vary over the business cycle, which can in turn affect the ranking of firms in the distribution of these variables. In this sense, even if there was any, it would be hard to interpret markups (ex-post) heterogeneity as being driven by ex-ante differences in these specific firm characteristics. Contrary to that, we establish that firm age can significantly determine the differential response of producers' markups to MP shocks, above and beyond other relevant firm characteristics.

The relative response of markups of old companies estimated through 3.3 does not allow to understand the separate response of markups of firms in different age categories to monetary policy shocks. In particular, the regression specification in 3.3 is saturated with industry and time FE that span out completely the time-series variation common across all firms. Hence, we proceed to estimate the following regression specification for firms above and below the median age:

$$\Delta_h \log \mu_{i,t+h} = \varphi_{i,h} + \boldsymbol{\vartheta_h t} + \sum_{\ell=1}^{L} \boldsymbol{\delta_h'} \boldsymbol{X_{i,t-\ell}} + \sum_{k=-\kappa}^{h} \gamma_h^k \varepsilon_{t+k}^m + u_{i,t+h} \tag{3.4}$$

with horizons $h = 0, 1, \ldots, H$ and $H = 20$ quarters. Note that the dependent variable is the cumulative change in markups for any firm i at horizon h, which is defined as:

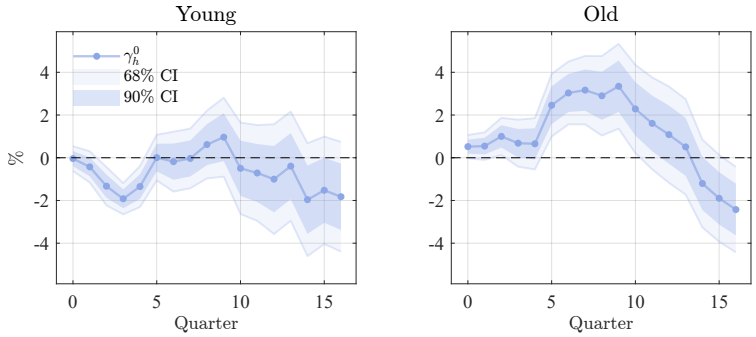
$$\Delta_h \log \mu_{i,t+h} \equiv \log \mu_{i,t+h} - \log \mu_{i,t-1}$$

Hence, in this second specification, we simply exploit the time-variation and look at the absolute change in markups after a change in the interest rate set by the FED for firms of different age categories, while the coefficient of interest γ_h is estimated for each age group separately. Note that

$\boldsymbol{\varepsilon_t^m} \equiv \sum_{k=-4}^{h} \varepsilon_{t+k}^m$ is again the series of monetary policy shocks from Jarociński and Karadi (2020). Moreover, $\boldsymbol{X_{i,t}}$ is a vector of controls that include firm-level variables such as sales growth and overhead costs to sales, leverage, liquidity and assets, as well as macro-level controls like GDP growth, CPI growth, 1-year treasury rate change, EBP, and fiscal quarter dummies. Importantly, we also include control variable lags (up to 4). We allow for firm's FE ($\varphi_{i,h}$) to account for time-invariant firm-heterogeneity, and also add a linear and quadratic trend ($\boldsymbol{\vartheta_h t}$). Finally, we cluster our robust standard errors at the firm and quarter level to account for correlation in the error term.

The results of our estimation are shown in 3.2: more specifically, the left panel documents the cumulative response of markups for firms below the median age to a negative movement in the FED interest rate, while the right panel focuses on the markups response of companies above the median age. This second estimation strategy further strengthens the insight from 3.1, by documenting that older firms present a pronounced and statistically significant countercyclical response in their markups after a monetary policy tightening, while young firms' markups move procyclically, albeit the statistical significance of the estimated coefficient is much lower. Importantly, note that this second specification estimates a dynamic regression without the sector-time fixed effects and still shows that the above-median age firms' response in markups is nonetheless persistent, peaking 8 to 10 quarters after the shock. Taken together, these findings seem to suggest that dominant companies do adjust upwards their markups, whereas young firms' markups are generally less sensitive to monetary policy or tend to be adjusted downwards following a negative change in the FED interest rates. Finally, we check that our results hold more generally when we focus on a different partitioning of the age distribution, by for example considering as "old"those firms that are above the third quartile and as "young"firms all the others. Moreover, to have a further understanding of the possible interaction between corporate age and other firm-level characteristics, we also split old and young companies according to their position in the distribution of leverage, liquidity and size, which are other dimensions of firm's heterogeneity typically

Figura 3.2: Firms' Markups Response to a Monetary Policy Shock by Age Category



Notes: Within each quarter, firms' markups are winsorized at the 1% and 99% cutoff, to avoid any outlier to drive our results.

investigated in the literature that we have always controlled for in our regression analysis. Corporate age is the crucial dimension determining the heterogeneous response of markups to monetary policy shocks, while other firm's characteristics – such as leverage, liquidity or size – are less powerful or even insignificant predictors of the differential behavior of markups at the company-level.

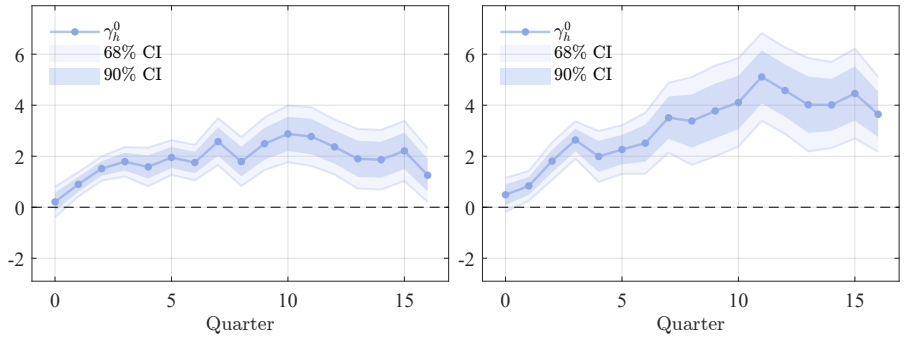
3.2.4 Discussion of Results

In what follows, we discuss the main robustness checks to further confirm the evidence on the role of age in shaping markups response to MP shocks at the firm-level. First, we run alternative regression specifications that present minor differences with respect to our baseline case. In particular, relative to the way we define the regressor of interest – namely corporate age – our main result is robust to consider age groups by industry and quarter, and also to interact the interest rate shock series with firm's age in a linear fashion, thereby adopting a parametric estimation strategy (similarly, we also linearly interact the MP shock series with the leverage, size and liquidity of the firms). The results of these alternative

specifications are reported in 3.12 in the Appendix and both confirm that older firms present a stronger countercyclical response of markups to a monetary policy tightening.

Secondly, we check that our insights are not driven by the specific time span considered, in particular, by running again our estimation procedure on a sub-sample of the dataset that extends until the 2009 crisis. This is due to the fact that the Great Recession was indeed a period of exceptional financial conditions and, at the same time, the post-2009 era was characterized by a lower variation in the interest rate policy, with the federal funds rate often hitting the zero lower bound. However, as reported in the right panel of 3.3, our results acquire a stronger statistical significance when the post-2009 era is excluded, and are hence not driven by specific period conditions only. Moreover, we also replicate our estimation using

Figura 3.3: Using GSS Shocks (left) and Focusing on pre-2009 period (right)



Notes: Within each quarter, firms' markups are winsorized at the 1% and 99% cutoff, to avoid any outlier to drive our results.

the monetary policy shock series from Gürkaynak et al. (2005) (hereafter GSS), which does not remove the informational component when measuring the interest rate surprises based on the 30-minute windows around FED policy announcements. As it is possible to check from the left panel of 3.3, the coefficient on the interaction between the MP shock and firm's age – $\gamma_{age,h}$ – is economically relevant and significant, confirming that

firms above the median age present an excess counter-cyclicality in their markups response to a monetary policy tightening, and that this differential effect lasts for an average horizon of 16 quarters after the shock.

Finally, our findings are robust to excluding future shocks from the estimation, as well as sector-quarter fixed effects, as reported in 3.13. In particular, future shocks were included among the regressors to control for the presence of auto-correlation and to increase the estimation precision, despite of the fact that the monetary policy shock series we have used should already be clear from confounding factors of this sort. Taking our results together, we argue that corporate age is a robust driver of firm's heterogeneity in the cyclicality of markups response to a monetary policy shock, and we hence proceed to briefly analyse the behavior of markups over firms' life-cycle.

3.2.5 Markups and Firm's Life-Cycle

After having estimated the heterogeneous response of firms' markups to MP shocks, we provide a further discussion on why old and young companies may possibly show such stark differences in cyclical behavior of their respective markups. As mentioned before, corporate age has been investigated to be an important element of firm's employment and leverage dynamics over the business cycle by the works of Haltiwanger et al. (2013), Dinlersoz et al. (2018), and Pugsley et al. (2019). Interestingly, Cloyne et al. (2018) have studied how corporate age can determine investment heterogeneity across firms, especially in response to interest rate changes. Specifically, by documenting that the investment and the borrowing of younger firms paying no-dividends exhibit a large and significant decline in response to a tightening of the monetary policy, the authors argue that such companies are more likely to face financial frictions. In their view, this can also rationalize why the borrowing of young and non-dividend paying firms is far more sensitive to fluctuations in collateral values compared to other businesses, for which their results turn less significant.

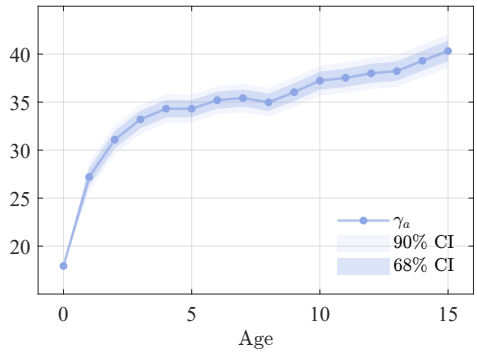
In a similar in spirit, we argue instead that firm age matters signifi-

cantly for the profile of markups and their response to a MP shock. In particular, it is reasonable to assume that corporate age may capture how *established* is a firm in her (unobservable) product market. Older firms, by means of having competed and produced in their given markets for a longer period, may be able to charge higher prices to consumers and hence be less subject to the downward pressure exerted on prices by a monetary policy tightening. In fact, according to the theoretical expression of firm-level markups, a negative interest rate shock puts a negative pressure on both input costs and prices. However, if older firms are able to decrease their prices by relatively less by taking advantage of their established position within a market, this may rationalize a more countercyclical markups response to a monetary tightening. To provide suggestive evidence of how corporate age is related to firm's established position in a given market, we examine the profile of markups ($\mu_{i,t}$) and selling expenditure ($sr_{i,t}$) over the business life-cycle using Compustat data. In particular, we run the following regressions:

$$\log \mu_{i,t} = \alpha + \sum_{a=2}^{\mathcal{A}} \gamma_a 1_{\{age_{i,t}=a\}} + \varphi_{s,t} + \varepsilon_{i,t}$$

where $\varphi_{s,t}$ are sector and quarter fixed effects. Not only old firms are on average big, but they most importantly tend to have higher markups, as documented in 3.4. The main takeaway from 3.4 is that firms are able to charge higher markups (hence higher prices) as they grow older. We interpret this suggestive evidence as an indicator that older companies may have already secured their customer base enough to be less inclined to drastically reduce prices in response to a negative monetary policy shock, resulting in the stronger markups counter-cyclicality documented in the previous paragraphs. Since we argue that firm age is a proxy for how established a production unit is in her given market, we will then rationalize our findings into a NK model with heterogeneous firms, demand accumulation and endogenous markups that not only will generally differ across producers but that will be further allowed to grow according to the firm's life-cycle.

Figura 3.4: Markups over Firm’s Life-Cycle



3.3 The Model

In this section, we outline our theoretical framework and discuss how each assumption relates to and can deliver qualitative predictions in line with the evidence from the data. In particular, we enrich a relatively standard NK to accommodate three main novelties: first, we allow for full heterogeneity on the supply side of the economy, by including heterogeneous intermediate firms that produce a different variety of input used in the final good sector. Secondly, we introduce a simple form of demand accumulation that makes the demand for the good of a given firm increase with the time that the firm survives on the market. Third, we embed endogenous and variable markups in the economy, which differ across companies according to the quantities they produce, and that also evolve along with the life-cycle of the firms. We now proceed to present the model in full details below.

3.3.1 Household’s Side

Time is continuous. The model features a representative household that optimizes the discounted flow of utility from consumption and labor over an infinite lifetime horizon, where we indicate the discount factor as $\rho \geq 0$. We assume that the utility of the agent is strictly increasing and con-

cave in consumption, and strictly decreasing and convex in the amount of hours worked respectively. Preferences are time-separable and the infinite stream of household's utility is hence given by:

$$\int_0^\infty e^{-\rho t}\left(\frac{C_t^{1-\nu}}{1-\nu}-\frac{L_t^{1+\gamma}}{1+\gamma}\right)$$

where ν represents the risk aversion in the CRRA utility function over consumption, whereas γ is the inverse of the Frisch labor elasticity. Moreover, $L_t \in [0,1]$ are the hours supplied as a fraction of the time endowment (normalized to 1), while C_t denotes the aggregate consumption good. In each period, the household can borrow in bonds B_t at the real interest rate r_t. Finally, the household owns all the firms in the economy, while labor supply, aggregate consumption and bond investment paths are chosen as a result of a value maximization problem subject to a standard budget constraint:

$$\mathcal{V} = \max_{\left\{C_t, L_t, \dot{B}_t\right\}} \int_0^\infty e^{-\rho t}\left(\frac{C_t^{1-\nu}}{1-\nu}-\frac{L_t^{1+\gamma}}{1+\gamma}\right) dt \tag{3.5}$$

$$\text{s.t.} \quad C_t + \dot{B}_t = W_t L_t + r_t B_t + D_t \tag{3.6}$$

where we denote by D_t the dividends from the firms and by W_t the wage earned by the household in real terms. As we will explain below, r_t will be determined by the monetary policy and Fisher equation, while W_t is determined by the market clearing condition for labor. Solving for the optimal value of consumption and labor, we get the following standard Euler and labor supply equations:

$$r = \rho + \nu\frac{\dot{C}}{C} \tag{3.7}$$

$$L^\gamma C^\nu = \frac{W}{\varphi} \tag{3.8}$$

3.3.2 Final Good Producer

A competitive representative final-good producer aggregates a continuum of intermediate inputs indexed by $i \in [0,1]$ according to the following

expression:

$$\int_0^1 \mathcal{K}\left(a_{i,t}\frac{y_{i,t}}{Y_t}\right)di = 1 \tag{3.9}$$

where we assume that intermediate inputs denote by y_t are aggregated using the Kimball aggregator $\mathcal{K}$, with $\mathcal{K}'(\cdot) > 0$, $\mathcal{K}''(\cdot) < 0$, and $\mathcal{K}(1) = 1$. Notice that the CES aggregator obtains as a special case of the Kimball aggregator, and namely when $\mathcal{K}(q) = q^{\frac{\sigma-1}{\sigma}}$ for an elasticity of substitution $\sigma > 1$. Importantly, $a_{i,t}$ is a stochastic demand process that will be explained in due details in the next paragraph. For the moment, taking the prices $p_{i,t}$ of any intermediate input i as given and normalizing the price of the final good to 1, the final good producer minimizes production costs subject to 3.9. The optimality condition of this problem gives rise to the *inverse demand* function for good i:

$$p_{i,t} = \mathcal{K}'\left(a_{i,t}\frac{y_{i,t}}{Y_t}\right)a_{i,t}\mathcal{D}_t \tag{3.10}$$

where:

$$\mathcal{D}_t = \left(\int_0^1 \mathcal{K}'\left(a_{i,t}\frac{y_{i,t}}{Y_t}\right)a_{i,t}\frac{y_{i,t}}{Y_t}di\right)^{-1} \tag{3.11}$$

is a *demand index*. In the CES case $\mathcal{K}(q) = q^{\frac{\sigma-1}{\sigma}}$ this index is a constant, so that $\mathcal{D}_t = \frac{\sigma}{\sigma-1}$ and 3.10 reduces to the familiar constant elasticity demand curve given by $p_{i,t} = \left(a_{i,t}\frac{y_{i,t}}{Y_t}\right)^{\frac{-1}{\sigma}}$. To keep the exposition concise, further derivations related to 3.10 and 3.11 are contained in the Appendix. Moreover, we use the Klenow and Willis (2016) specification for $\mathcal{K}(q)$ given by:

$$\mathcal{K}(q) = 1 + (\sigma - 1)\exp\left(\frac{1}{\omega}\right)\omega^{\frac{\sigma}{\omega}-1}\left[\Gamma\left(\frac{\sigma}{\omega}, \frac{1}{\omega}\right) - \Gamma\left(\frac{\sigma}{\omega}, \frac{q^{\omega/\sigma}}{\omega}\right)\right] \tag{3.12}$$

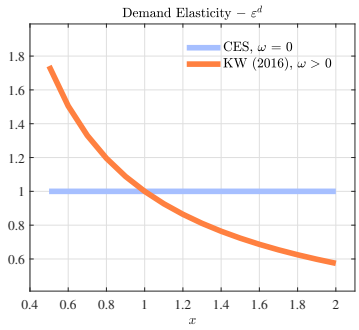
where $\sigma > 1$, $\omega \geq 0$ and $\Gamma(s, x)$ is the upper incomplete Gamma function such that $\Gamma(s, x) := \int_x^\infty t^{s-1}e^{-t}dt$. In particular, ω is the *super elasticity*,

which is 0 in the CES aggregator. Finally, we can derive an analytical expression for the elasticity of demand ε_i^d that is faced by a producer of any good variety i as a function of the relative quantity of good i in the economy, which is given by:

$$\varepsilon_i^d = \sigma \left(a_{i,t} \frac{y_{i,t}}{Y_t} \right)^{-\frac{\omega}{\sigma}}, \quad \omega \geq 0$$

As already pointed out, the standard CES case is recovered when $\omega = 0$ and hence the elasticity of demand $\varepsilon_i^d = \sigma$ is constant across producers. In contrast, in the case of Kimball aggregator the elasticity of substitution is lower for firms with higher relative quantity $x = a\frac{y}{Y}$, implying that larger firms can choose higher markups, in a similar spirit to the different set up adopted in Atkeson and Burstein (2008) and as further made clear in 3.5. When $\omega > 0$, the extent to which a firm's markup increases with its relative size is determined by the ratio σ/ω, which will also be shown to quantitatively matter in shaping how markups change with monetary policy later on in the analysis.

Figura 3.5: Kimball Aggregator



3.3.3 Intermediate Good Producers

Each intermediate good i is produced by a monopolistically competitive firm using effective units of labor $\ell_{i,t}$ in the production process and

according to the technology given by:

$$y_{i,t} = \ell_{i,t}^{1-\alpha} \tag{3.13}$$

with $\alpha \in [0, 1]$. In each time t, firms hire labor at wage W_t in a competitive labor market. As already mentioned, intermediate producers are monopolistic competitors on their respective markets and each one of them faces a demand function which can be written explicitly from 3.10 as:

$$y_{i,t} = \left(1 - \omega \log\left(\frac{\sigma}{\sigma - 1}\frac{1}{a_{i,t}}\frac{p_{i,t}}{\mathcal{D}_t}\right)\right)^{\sigma/\omega} \frac{Y_t}{a_{i,t}} \tag{3.14}$$

Moreover, each intermediate firm i is characterized by a process of demand accumulation given by a_i, which shows some persistence ρ_a and an idiosyncratic risk component given by $\xi_a d\mathcal{W}$ (as we work in continuous time, note that $d\mathcal{W}$ is a standard Wiener process). We also include a drift $\bar{a}$ that allows for the demand to grow over time, generating a realistic life-cycle profile. It is important to stress that we load the heterogeneity across firms that we see in the data in this specific process, which is meant to capture in a reduced-form the fact that markups and size increase with the firm's life-cycle. Such demand process may actually rationalize some underlying form of customer accumulation or, alternatively, a latent phenomenon of consumers habit formation. In other words, one can think about it in the sense that the more consumers experience the good of a given firm i, the more inelastic their demand for that specific item would consequently be.

Intermediate firms in this economy, characterized by the demand process a, maximize the discounted stream of profits by choosing prices. Hence, at each instant in time, the state of the economy is given by the joint distribution $\lambda_t(da, dp)$. Finally, intermediate producers discount future profits at the rate $r_t + \delta$, where δ is the exogenous Poisson intensity that determines firm's exit. Exiters are replaced by new firms with an initial a_0 drawn from a log-normal distribution of mean $\overline{a}_{entry}$ and standard deviation $\xi_{a,entry}$, which will be further discussed in the calibration exercise. Moreover, intermediate firms bear Rotemberg adjustment costs

when changing prices, which we assume to be proportional to their sales and quadratic. We can summarize the problem of a given firm i as follows:

$$\mathcal{J}_{i,0} = \max_{\{\dot{p}_{i,t},\ell_{i,t},y_{i,t}\}_{t\geq 0}} \mathbb{E}_0 \int_0^\infty e^{-\int_t^\infty (r_t+\delta)dt} \left\{ p_{i,t}y_{i,t} - W_t\ell_{i,t} - \frac{\vartheta}{2}\left(\pi_t + \frac{\dot{p}_{i,t}}{p_{i,t}}\right)^2 p_{i,t}y_{i,t} \right\} dt \tag{3.15}$$

$$\text{s.t.} \quad y_{i,t} = \left(1 - \omega \log\left(\frac{\sigma}{\sigma-1}\frac{1}{a_{i,t}}\frac{p_{i,t}}{\mathcal{D}_t}\right)\right)^{\sigma/\omega} \frac{Y_t}{a_{i,t}} \tag{3.16}$$

$$y_{i,t} = \ell_{i,t}^{1-\alpha} \tag{3.17}$$

$$\dot{a}_{i,t} = \rho_a(\overline{a} - a_{i,t})dt + \xi_a d\mathcal{W}_{i,t} \tag{3.18}$$

$$p_{i,0} \text{ and } a_{i,0} \text{ given} \tag{3.19}$$

Importantly, the initial price set by entrant firms p_0 is the one that maximizes the expected value $\mathcal{J}_{i,0}$ for a given initial value of firm's productivity $a_{i,0}$. Note that, in the solution process, the demand process given by $\dot{a}_{i,t}$ is exponentiated. Intermediate firms take as given equilibrium paths for the real wage $\{W_t\}_{t\geq 0}$ and the interest rate $\{r_t\}_{t\geq 0}$. In steady state, the recursive solution to this problem consists of decision rules for labor $\ell(a, p; \mathcal{S})$ and output $y(a, p; \mathcal{S})$, with $\mathcal{S} := (r, W, Y, \mathcal{D}, \pi)$. These rules in turn also imply optimal drifts for prices, and together with a stochastic process for a, induce a stationary joint distribution of firms given by $\lambda(da, dp; \mathcal{S})$ and characterized by a standard Kolmogorov forward equation. Out of the steady state, each of these objects is time-varying and depends on the time path of prices and policies: $\{\mathcal{S}_t\}_{t\geq 0} := \{r_t, W_t, Y_t, \mathcal{D}_t, \pi_t\}_{t\geq 0}$.

3.3.4 Monetary Authority

Our model economy features a monetary authority that sets the nominal interest rate according to a standard Taylor rule, penalizing deviations from the optimal inflation rate π^* in the following way:

$$i_t = \phi_\pi(\pi_t - \pi^*) + \rho + \varepsilon_t^m \tag{3.20}$$

where $\phi_\pi > 1$, ρ is the discount factor and ε_t^m is the monetary policy shock that can be mapped directly to the series from either Jarociński and Karadi (2020) or Gürkaynak et al. (2005) that we have used in the empirical analysis of the paper. Note that $\varepsilon_t^m = 0$ in steady state: one of our main quantitative exercises will be precisely to study the economy's adjustment after an unexpected temporary monetary shock, namely after a change in ε_t^m. Finally, given inflation π_t and the nominal interest rate i_t, the real return on bonds r_t is determined by the Fisher equation $r_t = i_t - \pi_t$.

3.3.5 Equilibrium Condition

An equilibrium in this economy is defined as a set of paths for individual household's $\{C_t, L_t\}_{t\geq 0}$ and firm's decisions $\{\dot{p}_{i,t}, \ell_{i,t}, y_{i,t}\}_{t\geq 0}$, input prices $\{W_t\}_{t\geq 0}$, the return on bonds $\{r_t\}_{t\geq 0}$, the inflation rate $\{\pi_t\}_{t\geq 0}$, the distribution of firms $\{\lambda_t\}_{t\geq 0}$, the demand index $\{\mathcal{D}_t\}_{t\geq 0}$, and aggregate quantities such that, at every t: (i) the household and the firms maximize their objective functions taking as given equilibrium prices and aggregate quantities; (ii) the sequence of distributions satisfies aggregate consistency conditions; (iii) all markets clear. There are three markets in our economy: the bond market, the labor market, and the goods market. The bond market clears when the following holds:

$$B_t = 0 \tag{3.21}$$

Moreover, the labor market clears when:

$$L_t = \int \ell_t(a,p) d\lambda_t \tag{3.22}$$

Finally, the goods market clears according to:

$$C_t = Y_t - \int \frac{\vartheta}{2}\left(\pi_t + \frac{\dot{p}_t(a,p)}{p}\right)^2 py(a,p)d\lambda_t \tag{3.23}$$

where C_t is the total real expenditure in consumption, Y_t is aggregate output, and the last term is the sum of adjustment costs to prices paid by intermediate firms.

3.4 Quantification

In what follows, we proceed to explain the quantification of our model, including the calibration strategy and the overall fit of both targeted and untargeted moments computed from available US data. In particular, we discuss the ability of our theoretical framework to replicate salient features of the markups and firms' distribution, which is a crucial property needed to provide a link with the empirical analysis of the previous sections. Once quantified, the model is then used in 3.5 to study and analytically decompose the impulse response functions of firms' markups after a negative monetary policy shock. Moreover, in 3.5, we also compare the amplification mechanism implied in our framework with respect to a standard representative firm New Keynesian model.

3.4.1 Calibration

A model period in one quarter. Of the 14 parameters we need to calibrate, 8 are fixed outside of the model, for which we pick common values used in the literature. In particular, we set the risk aversion $\nu = 2$ and the disutility of labor $\gamma = 2$, while the discount factor $\rho = 0.012$ is specified to deliver a yearly interest rate of 5% in equilibrium. With respect to the parameters related to firms' life-cycle, technology and pricing behavior, we fix the quarter exit rate $\delta = 0.024$ to imply that 10% of the firms exit each year, and the returns to scale $\alpha = 0.33$ such that the labor share is around 0.6 in equilibrium. Moreover, it is important to specify that we normalize at 1 the mean demand $\overline{a}_{entry}$ faced by entrant intermediate firms, while the demand dispersion $\xi_{a,entry}$ at entry is set to be equal to the dispersion of the demand process faced by incumbents.[6] Finally, the monetary policy coefficient $\phi_\pi = 1.5$ in the Taylor rule is chosen to replicate a similar strategy as in Taylor (1999) and Galí (2015). The full list of both fixed and fitted parameter, as well as targeted moments, is presented in 3.2.

[6] Our results do not depend on this choice, which is just a simplification for the sake of the estimation procedure.

Taula 3.2: Estimated Parameters and Targeted Moments

Fixed	Value	Description			
ρ	0.012	Discount factor			
ν	1	Risk aversion			
γ	2	Inverse Frisch elasticity			
α	0.33	Production function curvature			
δ	0.024	Exit rate			
$\overline{a}_{entry}$	1	Mean demand entrants			
$\xi_{a,entry}$	0.11	Demand dispersion entrants			
ϕ_π	1.5	Taylor rule coefficient			
Fitted	Value	Description	Moments	Model	Data
θ	20	Price adjustment cost	Avg. cost change prices over sales	0.11	0.09
σ	4	Elasticity of demand	Avg. markup	1.68	1.68
ω	5.1	Superelasticity of demand	Elasticity markups to sale shares	0.11	0.10
$\overline{a}$	2	Mean demand	Median markup	1.37	1.30
ξ_a	0.11	Demand dispersion	Markups standard deviation	1.23	1.22
ρ_a	0.02	Demand mean reversion	Markups growth between age 0-5	0.24	0.22

Note: Empirical estimates for fitted parameters from Compustat Data (1990Q1-2016Q4). For the fixed parameters, see text.

In addition to that, we need to endogenously assign values to the remaining 6 parameters, for which we match as many salient moments from available US data. To begin with, we set the price adjustment cost factor $\theta = 20$ such that the average ratio between the cost paid by firms to change prices and their sales is the same in the model and in the data.[7] As standard in the literature, we set the elasticity of demand $\sigma = 4$ to match an average markup of 1.68 computed in the sample of Compustat firms:[8] this parameter determines the level of substitutability across the output of different producers in the model, and hence influences the average market power in the economy. Moreover, the superelasticity of demand ω is fitted

[7]Estimates for vary between 0.04 for physical costs and 0.09 for customer costs, see for example Levy et al. (1997) and Zbaracki et al. (2004). As in Golosov and Lucas (2007) and Baley and Blanco (2019), we choose a value in between those.

[8]We use Compustat Data between 1990Q1 and 2016Q4. For the empirical definition of markups, see 3.2.

such that the elasticity of markups to sale shares in the model is the same as in the data. In particular, our choice is motivated by the fact that the parameter ω in the Kimball aggregator is tightly linked to the relationship between the relative size of the firms and their markups: if ω was to be 0, such relationship would be null because all firms would have the same markup independently of their size. On the contrary, for $\omega > 0$, the higher the ω the higher the dependence of markups on sales shares. To this end, using Compustat firm-level data, we empirically estimate the elasticity of (log) markups to (log) sales shares according to:

$$\log \mu_{i,t} = \beta * \log(\textit{sales shares})_{i,t} + \varphi_{s,t} + \varepsilon_{i,t} \tag{3.24}$$

where $\varphi_{s,t}$ are sector-time FE and the coefficient β precisely informs by how much markups are linked to firms' sales shares. In the model, we use the theoretical definitions of markups and sales shares.

Finally, turning to the parameters related to the demand accumulation process, the mean demand is set to match the median markup in the US economy, as $\overline{a}$ identifies the distance between the average demand faced by entrants and incumbents, and hence relates to the skewness of the markup distribution. Furthermore, the dispersion in the demand process faced by incumbent firms ξ_a is identified from the standard deviation of markups, while the mean reversion in the demand process ρ_a is picked to match the growth of markups for firms between age 0 to 5. In particular, a higher mean reversion in the demand process impacts how fast firms grow, and therefore relates to the trajectory of markups over the firm's life-cycle.

3.4.2 Quantitative Fit

In the following paragraphs, we present and discuss our main validation exercises, which provide a overview of the quantitative fit of our framework with respect to empirical moments and data features that have not been targeted in the calibration. In particular, we first discuss the cross-sectional and life-cycle characteristics of firms in our model, and how they compare to their empirical counterparts from Compustat. Secondly,

we dig into the properties of the markup distribution and then analyse markups dynamics over the firm's life-cycle. Finally, we conclude with a note on the model and data-implied elasticity of wages to sales and relate it to the behavior of markups under the Kimball aggregator case and in imperfect competition, following similar lines as in Edmond et al. (2018).

Implications for Markups in Steady State

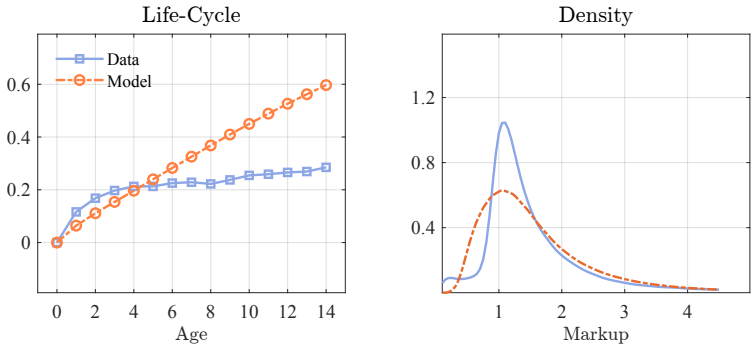
One of our main validation exercises is to look at the properties of markups in the data and compare them with the ones implied by our quantitative framework. Importantly, in 3.2, we have shown that markups increase with the age of the firm, and argued that such behavior may in principle be due to the fact that, as businesses advance along their life-cycle, they are also able to establish their position in their respective markets and progressively accumulate demand for their products. This in turn allows producers to progressively charge higher prices and hence set higher markups. Accordingly, the left panel in 3.6 reports the pattern of markups over firms' life-cycle both in the model and in the data. In particular, we remind the reader that the empirical series has been computed using Compustat data between 1990Q1 and 2016Q4 and netting out sector and time FE.

On the one hand, the model slightly underestimates the rapid increase of markups in the first 5 years of a firm's life, whereas it tends to modestly overestimate their subsequent growth in the next years.[9] On the other hand, our calibrated framework can replicate qualitatively the growth of markups over firm age and match more than half of the quantitative features of the relationship between markups and the life-cycle of producers. Importantly, it needs to be stressed that the ability of the model to imply life-cycle markups' properties consistent with the empirical observations will prove crucial when assessing the differential response of firms to interest rate shocks. In fact, as documented in 3.2, old firms' markups show

[9]The fit is very precise during the first years of business operations which is due to the fact that, in our calibration, we target the mean reversion in the demand process ρ_a to match the growth of markups for firms aged 0 to 5.

a more countercyclical response after a negative MP shock: absent the fit of the life-cycle profile of markups, our model would then not be able to replicate the heterogeneous response of markups to a MP shock according to firms' relative age.

Figura 3.6: Markups Steady State Properties



Secondly, as illustrated in the right panel of 3.6, we can reasonably match the entire distribution of markups estimated from Compustat data. In particular, while a couple of distributional properties have been indeed targeted in the calibration, the model itself delivers a fat right tail in the distribution of markups consistent with our empirical observations and with the analysis of De Loecker et al. (2020). As reported in 3.3, our quantitative framework implies that the bottom 25% firms in the distribution have an average markup of 1.15, against an empirical value of 1.03 computed in the data, while a similar fit holds for the top 75% firms. Matching the right proportions of high and low-markup firms' will prove crucial when comparing the response of firms' markups to a monetary policy shock across companies that are below or above the median age.

The Link between Wages and Sales

While the superelasticity parameter ω has been identified by computing the elasticity of markups to sales shares, our calibrated model has also a testable prediction on the relationship between the wage bill and the sales

Taula 3.3: Distributional Properties of Markups

	Model	Data
Bottom 25% Firms	1.15	1.03
Top 75% Firms	1.79	1.86

of firms, which we can match as an untargeted dimension. In particular, recall that markups are a measure of whether firms can set prices above their marginal costs. Similarly to Edmond et al. (2018), in our theoretical set up the salaries paid by firm i hence depend on its sales and markup according to a simple expression given by:

$$\text{wage bill} = \frac{\text{sales}}{\text{markup}}$$

Moreover, if the superelasticity ω in the Kimball aggregator was equal to zero as in the standard NK model, markups would not increase with firm sales and, in turn, the wage bill shares would increase one-for-one with sales shares. But when ω is strictly positive, as in our framework, markups do increase with firms sales, implying that the wage bill increases less than one-for-one with sales. In this sense, both empirically and quantitatively, the extent to which the wage bill share of firms increases with their sales shares can therefore be linked to the extent to which markups increase with producers' size. A small caveat to keep in mind is that Compustat does not report a precise measure for firms' wage bills but only a balance sheet item related to the cost of goods sold. This variable comprises the cost of all variable inputs used in production, included (but not exclusively) labor. Nevertheless, we exploit the available data to run the following regression:

$$\log(\textit{wage bill shares})_{i,t} = \beta * \log(\textit{sales shares})_{i,t} + \varphi_{s,t} + \varepsilon_{i,t} \quad (3.25)$$

where $\varphi_{s,t}$ are sector-time FE and the coefficient β precisely informs by how much variable input costs are linked to firms' sales shares. A value

of the elasticity $\beta < 1$ confirms the fact that, absent perfect competition – as in our model –, firms increase sales by increasing prices, thereby suppressing produced quantities. In turn, this mechanism implies that growing firms also demand less employment, which creates a wedge such that wage bill shares do not move one to one with sales shares. The results of the empirical estimation and quantitative fit are reported in 3.4.

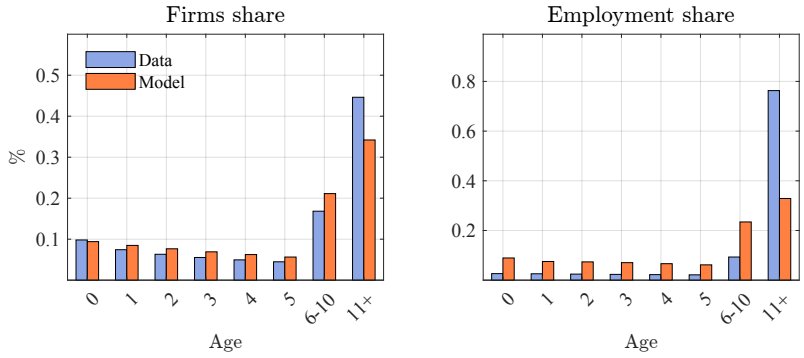
Taula 3.4: Estimated Relationship between Wages and Sales

	Model	Data
Elasticity of Wage Bill Shares to Sales Shares	0.87	0.88

Cross-sectional and Life-Cycle Properties

In our last exercise, we analyse the distribution of firms by age and the life-cycle profile of both employment and sales growth rates for the businesses in our model economy. In 3.7, we report the distribution of firms and employment shares by age, comparing the empirical ones from Compustat (1990Q1-2016Q4) to the ones obtained in our quantified framework. Note that none of these distributions was targeted in the calibration of the model, and hence both comparisons are to be considered as a pure validation exercise. First, focusing on the left panel, one can observe that our framework succeeds in replicating the distribution of firms over their age, and only partially underestimates the share of businesses that are 11+ years old. In this sense, as most of our empirical analysis is highly focused on markups' properties over the life-cycle of firms, it is remarkably important that we are able to capture the correct number of firms per age bin. In fact, the share of companies in each age bin influences the heterogeneous response of markups' to monetary policy shocks, and hence is relevant to get a correct quantitative fit of the empirically estimated dynamics of markups by firms' age.

Figura 3.7: Distributions of Firms and Employment Shares by Age

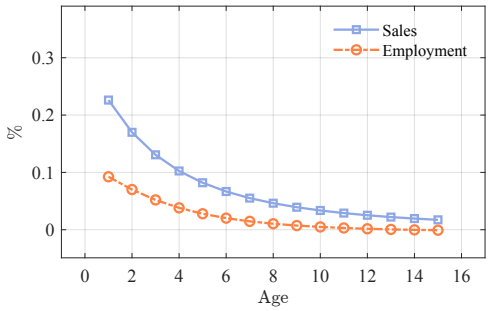


Secondly, in the right panel of 3.7, we plot the distribution of employment shares over firm age, comparing the empirical ones with their model-implied counterparts. As it becomes clear from the graph, our framework is able to match only up to half of the right tail in the employment share distribution. This is precisely due to the fact that, in the model economy, big firms (and hence old firms), find optimal to increase sales by increasing prices, thereby suppressing produced quantities and employment demand. This mechanism is a key characteristic of our set up in which companies operate in an environment with imperfect competition, and it is hence responsible for the fact that 11+ years old firms in the model generate a lower employment share compared to their empirical counterpart. Nonetheless, from this particular validation exercise we are still able to get a satisfactory fit of both firms and employment share distributions over the age of the businesses.

As a final note, in 3.8 we plot the average employment and sales growth rates over the life-cycle of firms. Understandably, both measures decrease over time, as companies become old and hence slow down in their growth processes: this means that growth rates are unconditionally negatively correlated with age, as empirically noted in Dunne et al. (1989). However, sales grow relatively more than employment, which is indeed consistent with the early discussion related to the employment share distribution depicted in the right panel of 3.7. In particular, as argued

in the previous paragraphs and due to the presence of the Kimball aggregator, markups do increase with firms sales, implying that the wage bill increases less than one-for-one with sales, depressing the labor demand by firms and resulting in lower employment growth rates compared to the growth rate of firm’s sales. In other words, due to market power, companies can increase sales by raising prices and decreasing output, which lowers their demand of labor and hence employment growth.

Figura 3.8: Employment and Sales Growth Rates



3.5 Results

In the following section, we begin by discussing the response of firm markups to interest rate shocks, and compare the relative response of old and young firms in the model with the ones obtained in the data and reported in 3.2. Secondly, having assessed how much of the heterogeneity in response of markups to interest rates by firm age our model is able to replicate, we also illustrate by how much the changes of aggregate variables such as output and wages after a MP shock contribute to the differential response of markups of old firms with respect to young ones. Finally, we conclude by analysing the amplification of shocks at work in our framework compared to a one-firm NK model.

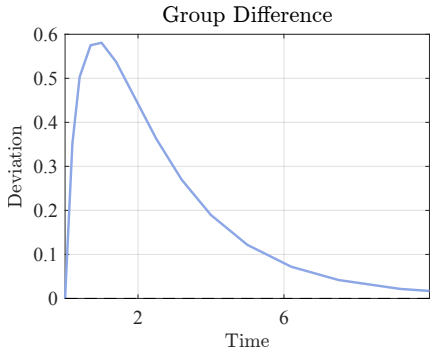
3.5.1 Response of Markups to Monetary Policy Shocks

We proceed to illustrate the dynamics of the economy after the arrival of a negative MP shock and to compare the relative response of old firms and young firms' markups to the ones obtained from the data and discussed in 3.3. As standard in frameworks characterized by nominal rigidities, a negative MP shock features an increase in the nominal interest rate and implies a downwards pressure on the labor cost W. Parallel to that, both employment, consumption and output decrease on impact and slowly recover as the shock fades away, while the downwards pressure on prices determines a deflationary episode. Moreover, the aggregate markup increases as a result of decreasing labor costs, and hence shows a countercyclical behavior in response to negative shocks to the nominal interest rate. The aggregate response of our calibrated economy hence resembles qualitatively the one of a standard NK textbook model, as in Galí (2015). However, the aggregate pattern of markups masks a noticeable degree of heterogeneity at the firm-level which we explore in what follows.

To obtain a comparable set up to our empirical analysis, we first categorize firms in our model economy by their age decile and then classify all businesses above the median age as "old"and below the median age as "young". We simulate the hit of a negative MP shock, otherwise defined as an exogenous increase in the nominal interest rate. Similarly to the empirical analysis in 3.1, we then compute the differential response of markups to a MP shock for firms above the median age compared to firms below the median age. 3.9 plots the differential response of markups by firm age over a horizon of several quarters and in deviation from the mean response. Clearly, firms above the median age respond more countercyclically to a negative MP shock compared to businesses below the median age, consistent with the empirical evidence documented in Compustat data.

Moreover, the differential response of old firms markups upon a negative MP shock in the model peaks at a value of 0.6%, while empirically it goes up to 3%. Importantly then, our quantitative framework is able to replicate 20% of the excess counter-cyclicality of old firms' markups to

Figura 3.9: Markups IRFs After a Negative MP Shock



MP shocks that we have estimated in Compustat. In turn, this represent a satisfactory quantitative validation of our framework, which is hence able to replicate both qualitatively and quantitatively the heterogeneity in the response of markups by firm age to MP shocks that has been documented in the data.

3.5.2 Decomposing the Differential Response of Markups

Analytical Result under Flexible Prices. In what follows, we show that the combination of the total derivatives of the demand function and the desired markup respectively gives the opportunity to understand the heterogeneous response of firm prices to changes in aggregate output Y, wage W and the demand index $\mathcal{D}$. For the sake of analytical tractability, we first carry out such decomposition in a version of the model without price adjustment costs. As derived in 3.3, the demand function in our theoretical framework is given by:

$$y = \left(1 - \omega \log\left(\frac{\sigma}{\sigma - 1}\frac{1}{a}\frac{p}{\mathcal{D}}\right)\right)^{\sigma/\omega}\frac{Y}{a} \tag{3.26}$$

while the desired markup can be written as follows:

$$\frac{\alpha p}{W y^{\frac{1}{\alpha}-1}} = \frac{\sigma\left(\frac{y}{Y}a\right)^{-\omega/\sigma}}{\sigma\left(\frac{y}{Y}a\right)^{-\omega/\sigma} - 1} \equiv \mu(a) \tag{3.27}$$

where $\mu(a)$ denotes the markup and increases in the demand faced by the firm. From these two equations, we can derive a set of expressions linking the change in firm prices (and similarly markups) to changes in aggregates W, Y and $\mathcal{D}$ and model parameters (see the derivations in the 3.8):

$$\frac{\partial \log p}{\partial \log Y} = \frac{\frac{1}{\alpha} - 1}{1 + \left(\frac{1}{\alpha} - 1\right) \frac{\mu(a)}{\mu(a)-1} + \frac{\omega}{\sigma}\mu(a)} \tag{3.28}$$

$$\frac{\partial \log p}{\partial \log W} = \frac{1}{1 + \left(\frac{1}{\alpha} - 1\right) \frac{\mu(a)}{\mu(a)-1} + \frac{\omega}{\sigma}\mu(a)} \tag{3.29}$$

$$\frac{\partial \log p}{\partial \log \mathcal{D}} = \frac{\left(\frac{1}{\alpha} - 1\right) \frac{\mu(a)}{\mu(a)-1} + \frac{\omega}{\sigma}\mu(a)}{1 + \left(\frac{1}{\alpha} - 1\right) \frac{\mu(a)}{\mu(a)-1} + \frac{\omega}{\sigma}\mu(a)} \tag{3.30}$$

where the standard CES equivalent, in the case of perfect competition, can be obtained setting the Kimball superelasticity parameter $\omega = 0$. Notice that all derivatives are positive, which means that the negative MP shock negatively affects W, Y and $\mathcal{D}$. Moreover, the first two derivatives are decreasing in $\mu(a)$ and the third one increases in $\mu(a)$. The same observations hold true if we were to write the derivative of firm's markup with respect to Y, W and $\mathcal{D}$. At the same time, the second derivatives with respect to the demand faced by the firm are given by:

$$\frac{\partial^2 \log p}{\partial \log Y \partial a} < 0, \quad \frac{\partial^2 \log p}{\partial \log W \partial a} < 0, \quad \frac{\partial^2 \log p}{\partial \log \mathcal{D} \partial a} > 0 \tag{3.31}$$

The signs of these second derivatives imply that the prices of firms facing higher demand decline less after the MP shock due to the effects coming from the decline in W and Y, whereas prices decline more due to the decline in $\mathcal{D}$. The aggregate effect prevailing in GE will then depend on the specific parametrization. It is important to stress that, while these derivatives have been taken with respect to the demand a face by firms, there is a strong correlation and direct mapping between the accumulation of demand and firm age progression. This ensures that we can safely

interpret the above results as the effects of the changes in Y, W, $\mathcal{D}$ after a MP shock on the prices of relatively older firms.

Benchmark Economy. The same decomposition is then carried out in practice in the quantitative model with nominal rigidities. In particular, we first compute numerically the general equilibrium response of the economy to a negative MP shock. Then, taking as given the equilibrium paths for the aggregate variables $Y, W, \mathcal{D}, r, \pi$ we look at the partial responses of old firms' markups to each of the shocks separately. Before commenting on the quantitative results, we follow the same spirit as in Kaplan et al. (2018) and provide intuition for the channels at play in our fully-fledged economy with heterogeneous firms and endogenous markups. Let us first write the difference between the average markups of old firms and the average markups of young firms as a function of the equilibrium prices, quantities, and inflation. We collect these terms in the vector $\{\mathcal{S}_t\}_{t\geq 0}$, with $\mathcal{S}_t = \{r_t, W_t, Y_t, \mathcal{D}_t, \pi_t\}$, and define the above-mentioned difference $\widehat{\mathcal{M}}(\{\mathcal{S}_t\}_{t\geq 0})$ induced by the path of the monetary shock $\{\varepsilon_t\}_{t\geq 0}$ from its initial hit until it fully reverts to zero as:

$$\begin{aligned}\widehat{\mathcal{M}}(\{\mathcal{S}_t\}_{t\geq 0}) := &\int \mu_t(p,a;\{\mathcal{S}_t\}_{t\geq 0}) 1\{g_t(p,a)\geq \overline{a}\} d\lambda_t \\ &- \int \mu_t(p,a;\{\mathcal{S}_t\}_{t\geq 0}) 1\{g_t(p,a)< \overline{a}\} d\lambda_t.\end{aligned} \tag{3.32}$$

where $\mu_t(p,a;\{\mathcal{S}_t\}_{t\geq 0})$ is the firm markup, $g_t(p,a)$ is a mapping between firm's states and its age, $\overline{a}$ is the median firms' age, and $d\lambda_t(p,a;\{\mathcal{S}_t\}_{t\geq 0})$ is the joint distribution of prices and idiosyncratic demand. Totally differentiating 3.32, we decompose the difference in the average markup response between old and young firms at time $t=\tau$ as:

$$d\widehat{\mathcal{M}}_\tau = \underbrace{\int_\tau^\infty \frac{\partial \widehat{\mathcal{M}}_\tau}{\partial r_t} dr_t dt}_{\text{direct effect}} + \underbrace{\int_\tau^\infty \left(\frac{\partial \widehat{\mathcal{M}}_\tau}{\partial W_t} dW_t + \frac{\partial \widehat{\mathcal{M}}_\tau}{\partial Y_t} dY_t + \frac{\partial \widehat{\mathcal{M}}_\tau}{\partial \mathcal{D}_t} d\mathcal{D}_t + \frac{\partial \widehat{\mathcal{M}}_\tau}{\partial \pi_t} d\pi_t \right) dt}_{\text{indirect effect}} \tag{3.33}$$

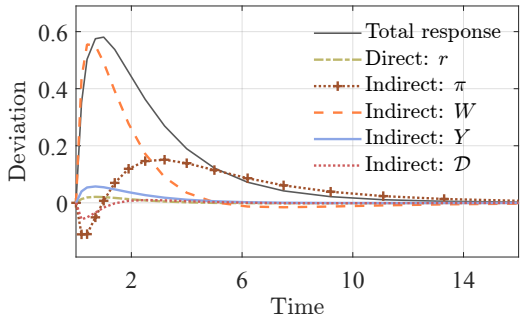
where the first term reflects the direct effect of a change in the interest rate, which enters the Euler equation of the agents, holding the other variables

of interest constant. The remaining terms in the decomposition reflect the indirect effects of changes in inflation, the real wage, real output and the demand index that arise in general equilibrium after the hit of the MP shock. In practice, we need to compute each of these components numerically. For example, the formal definition of the first term in 3.33, which is the direct effect of changes in the real interest rate $\{r_t\}_{t\geq 0}$, is:

$$\int_{\tau}^{\infty} \frac{\partial \widehat{\mathcal{M}}_{\tau}}{\partial r_t} dr_t dt = \int_{\tau}^{\infty} \frac{\partial \widehat{\mathcal{M}}(\{r_t, \overline{W}, \overline{Y}, \overline{\mathcal{D}}, \overline{\pi}\}_{t\geq 0})}{\partial r_t} dr_t dt. \tag{3.34}$$

This term is the *partial-equilibrium* response of the difference in the average markups between old and young firm that face a time-varying real interest rate path $\{r_t\}_{t\geq 0}$, but holding the paths for the real wage $\overline{W}$, the real output $\overline{Y}$, the demand index $\overline{\mathcal{D}}$, and nominal inflation rate $\overline{\pi}$ constant at their steady-state values. We calculate this term from the model by feeding these time paths into the firms' (and household's) optimization problem, computing the policy function and their markups for each firms, and aggregating across firms using the corresponding distribution. The other terms in the decomposition are computed in a similar fashion.

Figura 3.10: Decomposing the Differential Response of Markups



The results of the decomposition exercise are depicted in 3.10. All effects are to be intended as p.p. deviations from the mean response across all firms in the economy. The outer dark line represents the total GE effect of a negative MP shock on the differential impulse response of markups

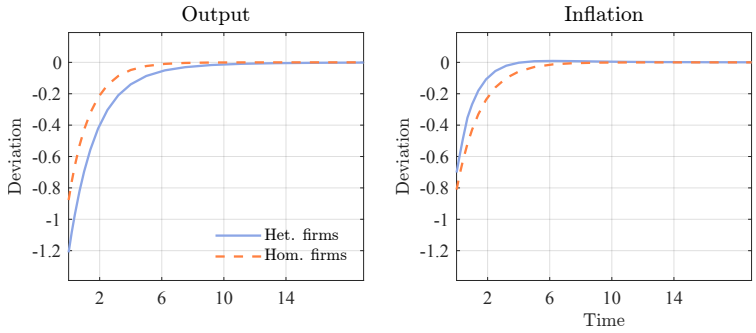
for old firms compared to young ones. Also, note that the GE effect is not a direct sum of the partial effects due to non-linearities in aggregation. Most of the resulting effect on the differential response of old firms' markups is to be attributed to changes in the aggregate W, hence to changes in the cost of labor after a negative shock to the interest rate. Since our model features heterogeneous and endogenous markups in the presence of a Kimball aggregator, big firms (and hence old firms) have a lower passthrough from production costs to prices. In this sense, a negative MP shock in the economy puts a downward pressure on the labor input cost W, but old firms' sales react less than proportionally, as dominant companies do not decrease prices as much. Since markups are the ratio between business sales and costs, the resulting effect on markups is positive, leading to the observed stronger countercyclical response of old firms' markups to a negative MP shock.

3.5.3 Amplification Mechanism

In what follows, we conclude our quantitative analysis by studying the shock amplification mechanism at work in our economy, comparing our calibrated framework with a standard one-firm NK model. As pointed out in Mongey (2017), in economies where real rigidities are present, shocks have a strong propagation through quantities, which we set to verify in our case. Moreover, we proceed to also explain to which extent both firm heterogeneity and the Kimball aggregator that characterize our model can be responsible for greater swings in macro aggregates after a negative MP shock.

To ensure we are working with two comparable economies, we first calibrate the one-firm NK economy to have the same size as in our heterogeneous firms framework (hereafter the FDNK) in terms of overall output produced. Moreover, since the standard NK model features perfect competition, we set the elasticity of substitution σ in its CES aggregator to match an aggregate markup of 1.68, which is the value targeted in the FDNK model under the Kimball aggregator. With the two models at hand, we simulate a negative MP shock and solve for the response of the

Figura 3.11: Comparing Output and Inflation Responses



main macroeconomic aggregates in the two economies. In particular, we analyse the trajectories of inflation π and output Y over an 16-quarters period, and hence compare the relative percentage deviation from steady state values of both prices and quantities. The results of this exercise are depicted in 3.11.

Comparing output and inflation responses across the two models, it is clear that a negative MP shock produces a bigger drop in output and a milder decline in prices in our FDNK set up compared to a standard one-firm NK model. The negative change in the interest rate decreases output by on average 20 p.p. more in the economy characterised by heterogeneous firms and endogenous markups, with the effect lasting for more than 10 quarters after the shock hits. At the same time, prices and hence inflation drop by relatively more in the one-firm NK model, which implies that the presence of the Kimball aggregator and the differential passthrough that characterize our model economy mitigate the downward pressure exerted by the negative MP shock on firm prices.

On the one hand, as argued in Klenow and Willis (2016), the presence of the Kimball aggregator adds a source of real frictions in the NK model, represented by a higher degree of concavity in the firm's profit function with respect to its relative price. Under the Kimball aggregator, sellers face a price elasticity of demand that is increasing in their good's relative price. For instance, when a repricing producer faces lower labor costs after a negative MP shock, it will temper its price drop because of

the endogenous increase in its desired markup, and this effect would be stronger the lower the elasticity of demand faced by the producer. Since the presence of a real rigidity makes firms more reluctant to change prices, firms do not pass marginal cost shocks as fully onto their prices as they would in a standard NK model with a CES aggregator. Hence, in our FDNK set up, MP shocks propagate more through quantities than prices, and decrease aggregate output by relatively more.

On the other hand, without heterogeneity on the firm's side, the presence of the Kimball aggregator alone does not automatically imply the amplification of shocks in our setup: in fact, the effects of the real rigidity introduced by the Kimball aggregator kick in only when businesses are indeed heterogeneous and hence characterized by different passthroughs from costs to prices with respect to one another. If all firms were to be equal (as in the representative-firm NK model), they would also be equal to the average firm in the economy and have identical sales shares. Specifically, focusing on 3.10, the elasticity of demand faced by producers would not vary across firm, and their response to MP shocks would be identical. On the contrary, in our FDNK set up, since big firms (hence old firms) respond more countercyclically than small ones and decrease their prices by less, the propagation of a negative shock gets strengthened. Hence, the heterogeneity of firms, combined with the real rigidity introduced by the Kimball aggregator set up, delivers the amplification mechanism at work in the present model.

3.6 Conclusion

In this paper, we have taken an empirical and theoretical approach to the study of firm heterogeneity in the response of markups to MP shocks. In order to carry out our data analysis, we have merged exogenously-identified monetary policy shocks series with a rich quarterly dataset comprising publicly-listed companies based in the US between 1990Q1 and 2016Q4. Next, we have documented that old firms' markups tend to increase after a monetary policy tightening, while young firms' markups show

a mildly procyclical behavior after a negative interest rate shock. Moreover, our empirical investigation seems to also suggest that the differential response of markups by firm's age could be related to the accumulation of customers and demand over time, which enables older firms to change by relatively less their prices thanks to an established position in their markets.

In our quantitative analysis, we have embedded our findings into a NK model, augmented with heterogeneous firms and a process of demand accumulation, and in which markups arise endogenously and evolve over the life-cycle of the companies. Our calibrated framework can replicate the life-cycle profile of firms' markups and growth rates, and the distribution of companies and employment shares by corporate age. Moreover, we were able to explain up to a fifth of the empirically estimated excess counter-cyclicality in the markups of firms above the median age after a negative monetary policy shock. Finally, we have shown that both firms' heterogeneity and endogenous markups generate amplification in the response of aggregate quantities to contractionary interest rate movements, which further distinguishes our set up from standard frameworks with nominal rigidities. In the future, we aim to further study optimal monetary policies in the presence of imperfect competition, demand accumulation, and heterogeneity in the passthrough from costs to prices.

3.7 Data Appendix

Figura 3.12: Alternative Specification for Corporate Age

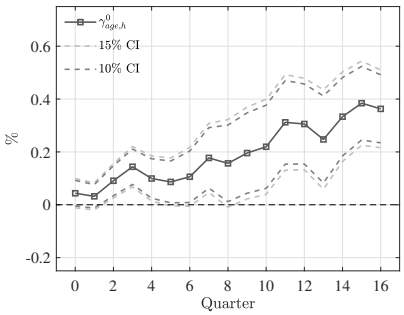
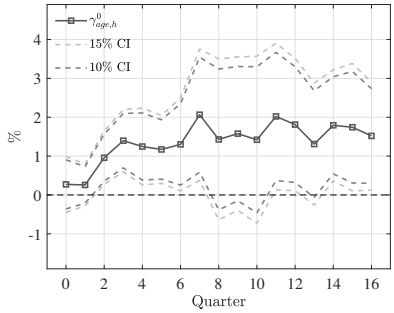
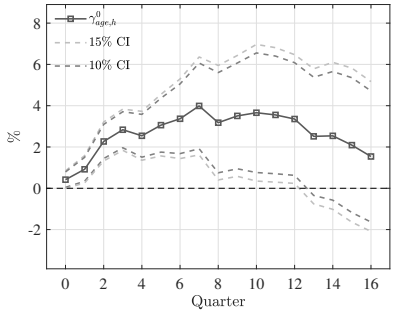


Figura 3.13: Excluding Future Shocks (left) and Sector-Quarter FE (right)





3.8 Quantitative Appendix

3.8.1 Decomposition Exercise: Derivations

The demand function in our model is given by:

$$y = \left(1 - \omega \log\left(\frac{\sigma}{\sigma - 1}\frac{1}{\xi(a)}\frac{p}{\mathcal{D}}\right)\right)^{\sigma/\omega}\frac{Y}{\xi(a)}$$

Which has the total derivative:

$$dlog\frac{y}{Y}\xi(a) = -\frac{\sigma(dlogp - dlog\mathcal{D})}{1-\omega\log\left(\frac{\sigma}{\sigma-1}\frac{1}{\xi(a)}\frac{p}{\mathcal{D}}\right)} = -\sigma\left(\frac{y}{Y}\xi(a)\right)^{-\omega/\sigma}(dlogp - dlog\mathcal{D})$$

The desired markup is instead defined as follows:

$$\frac{\alpha p}{W y^{\frac{1}{\alpha}-1}} = \frac{\sigma\left(\frac{y}{Y}\xi(a)\right)^{-\omega/\sigma}}{\sigma\left(\frac{y}{Y}\xi(a)\right)^{-\omega/\sigma} - 1}$$

By taking the total derivative it is possible to get:

$$\frac{\alpha}{W y^{\frac{1}{\alpha}-1}}dp - \frac{\alpha p}{W y^{\frac{1}{\alpha}-1}}dlogW + \left(1-\frac{1}{\alpha}\right)\frac{\alpha p}{W y^{\frac{1}{\alpha}}}dy =$$
$$-\frac{1}{\left(\sigma\left(\frac{y}{Y}\xi(a)\right)^{-\omega/\sigma} - 1\right)^2}\sigma\left(-\frac{\omega}{\sigma}\right)\left(\frac{y}{Y}\xi(a)\right)^{-\omega/\sigma-1}d\frac{y}{Y}\xi(a)$$

Substituting in the above expression $dy = \frac{Y}{\xi(a)}d\frac{y}{Y}\xi(a) + \frac{y}{Y}dY = y(dlog\frac{y}{Y}\xi(a) + dlogY)$ it is possible to obtained the following equation:

$$\frac{\alpha p}{W y^{\frac{1}{\alpha}-1}}dlogp - \frac{\alpha p}{W y^{\frac{1}{\alpha}-1}}dlogW + \left(1-\frac{1}{\alpha}\right)\frac{\alpha p}{W y^{\frac{1}{\alpha}-1}}(dlog\frac{y}{Y}\xi(a) + dlogY) =$$
$$-\frac{1}{\left(\sigma\left(\frac{y}{Y}\xi(a)\right)^{-\omega/\sigma} - 1\right)^2}\sigma\left(-\frac{\omega}{\sigma}\right)\left(\frac{y}{Y}\xi(a)\right)^{-\omega/\sigma}dlog\frac{y}{Y}\xi(a)$$

which in turn implies:

$$dlogp - dlogW + \left(1-\frac{1}{\alpha}\right)(dlog\frac{y}{Y}\xi(a) + dlogY) = \frac{1}{\left(\sigma\left(\frac{y}{Y}\xi(a)\right)^{-\omega/\sigma} - 1\right)}\left(\frac{\omega}{\sigma}\right)dlog\frac{y}{Y}\xi(a)$$

$$dlogp - \left(\left(\frac{1}{\alpha}-1\right) + \frac{1}{\left(\sigma\left(\frac{y}{Y}\xi(a)\right)^{-\omega/\sigma} - 1\right)}\left(\frac{\omega}{\sigma}\right)\right)dlog\frac{y}{Y}\xi(a) = dlogW + \left(\frac{1}{\alpha}-1\right)dlogY$$

Substituting $dlog\frac{y}{Y}\xi(a)$ from the total derivative of the demand function we get:

$$dlogp + \left(\left(\frac{1}{\alpha} - 1\right) + \frac{1}{\left(\sigma\left(\frac{y}{Y}\xi(a)\right)^{-\omega/\sigma} - 1\right)}\left(\frac{\omega}{\sigma}\right)\right)\sigma\left(\frac{y}{Y}\xi(a)\right)^{-\omega/\sigma}(dlogp - dlog\mathcal{D}) =$$

$$dlogW + \left(\frac{1}{\alpha} - 1\right)dlogY$$

$$dlogp + \left(\left(\frac{1}{\alpha} - 1\right)\frac{\mu(a)}{\mu(a) - 1} + \frac{\omega}{\sigma}\mu(a)\right)(dlogp - dlog\mathcal{D}) = dlogW + \left(\frac{1}{\alpha} - 1\right)dlogY$$

$$dlogp = \frac{dlogW + \left(\frac{1}{\alpha} - 1\right)dlogY + \left(\left(\frac{1}{\alpha} - 1\right)\frac{\mu(a)}{\mu(a)-1} + \frac{\omega}{\sigma}\mu(a)\right)dlog\mathcal{D}}{1 + \left(\frac{1}{\alpha} - 1\right)\frac{\mu(a)}{\mu(a)-1} + \frac{\omega}{\sigma}\mu(a)}$$

Where $\mu(a)$ is the markup. The above expression can be rearranged to get the relative contributions of Y, W and D to the change in prices and markups, reported in the main text.

Bibliografia

Andrew B Abel and Janice C Eberly. A unified model of investment under uncertainty. *American Economic Review*, 84:1369–1384, 1994.

Andrew B Abel and Janice C Eberly. Optimal investment with costly reversibility. *The Review of Economic Studies*, 63(4):581–593, 1996.

Daniel A Ackerberg, Kevin Caves, and Garth Frazer. Identification properties of recent production function estimators. *Econometrica*, 83(6): 2411–2451, 2015.

Hassan Afrouzi, Andres Dernik, and Ryan Kim. Growing by the masses. revisiting the link between firm size and market power. *Working Paper*, 2020.

Philippe Aghion, Antonin Bergeaud, Timo Boppart, Peter J Klenow, and Huiyu Li. A theory of falling growth and rising rents. *NBER Working Paper*, 2019.

Ajay Agrawal, Joshua Gans, and Avi Goldfarb. *Prediction machines: the simple economics of artificial intelligence*. Harvard Business Press, 2018.

Ufuk Akcigit and Sina T Ates. Ten facts on declining business dynamism and lessons from endogenous growth theory. *American Economic Journal: Macroeconomics*, 13(1):257–98, 2021.

Andrea Alati. *Essays on firms heterogeneity and business cycles*. PhD thesis, The London School of Economics and Political Science (LSE), 2020.

Andrea Alati. Initial aggregate conditions and heterogeneity infirm-level markups. *Working Paper*, 2021.

John Asker, Allan Collard-Wexler, and Jan De Loecker. Dynamic inputs and resource (mis) allocation. *Journal of Political Economy*, 122(5): 1013–1063, 2014.

Andrew Atkeson. Alternative facts regarding the labor share. *Review of Economic Dynamics*, 37:S167–S180, 2020.

Andrew Atkeson and Ariel Burstein. Pricing-to-market, trade costs, and international relative prices. *American Economic Review*, 98(5):1998–2031, 2008.

Andrew Atkeson and Patrick J Kehoe. Modeling and measuring organization capital. *Journal of political Economy*, 113(5):1026–1053, 2005.

Adrien Auclert. Monetary policy and the redistribution channel. *American Economic Review*, 109(6):2333–67, 2019.

David Autor, David Dorn, Lawrence F Katz, Christina Patterson, and John Van Reenen. The fall of the labor share and the rise of superstar firms. *The Quarterly Journal of Economics*, 135(2):645–709, 2020.

Isaac Baley and Andrés Blanco. Firm uncertainty cycles and the propagation of nominal shocks. *American Economic Journal: Macroeconomics*, 11(1):276–337, 2019.

David Baqaee, Emmanuel Farhi, and Kunal Sangani. The supply-side effects of monetary policy. Technical report, National Bureau of Economic Research, 2021.

Simcha Barkai. Declining labor and capital shares. *Stigler Center for the Study of the Economy and the State New Working Paper Series*, 2, 2016.

Robert J Barro. Double-counting of investment. Technical report, National Bureau of Economic Research, 2019.

Juliane Begenau, Maryam Farboodi, and Laura Veldkamp. Big data in finance and the growth of large firms. *Journal of Monetary Economics*, 97:71–87, 2018.

Florin O Bilbiie, Fabio Ghironi, Marc J Melitz, Virgiliu Midrigan, and Julio J Rotemberg. Monetary policy and business cycles with endogenous entry and product variety [with comments and discussion]. *NBER Macroeconomics Annual*, 22:299–379, 2007.

Florin O Bilbiie, Fabio Ghironi, and Marc J Melitz. Endogenous entry, product variety, and business cycles. *Journal of Political Economy*, 120 (2):304–345, 2012.

Mark Bils, Peter J Klenow, and Benjamin A Malin. Resurrecting the role of the product market wedge in recessions. *American Economic Review*, 108(4-5):1118–46, 2018.

Mark Bils, Peter J Klenow, and Cian Ruane. Misallocation or mismeasurement? Technical report, National Bureau of Economic Research, 2020.

Nicholas Bloom, Luis Garicano, Raffaella Sadun, and John Van Reenen. The distinct effects of information technology and communication technology on firm organization. *Management Science*, 60(12): 2859–2885, 2014.

Gideon Bornstein. Entry and profits in an aging economy: the role of consumer inertia. *Working Paper*, 2018.

Bart J Bronnenberg, Jean-Pierre H Dubé, and Matthew Gentzkow. The evolution of brand preferences: evidence from consumer migration. *American Economic Review*, 102(6):2472–2508, 2012.

Erik Brynjolfsson, Daniel Rock, and Chad Syverson. The productivity j-curve: How intangibles complement general purpose technologies. *American Economic Journal: Macroeconomics*, 13(1):333–72, 2021.

Kenneth Burdett and Melvyn G Coles. Steady state price distributions in a noisy search equilibrium. *Journal of Economic Theory*, 72(1):1–32, 1997.

Kenneth Burdett and Kenneth L Judd. Equilibrium price dispersion. *Econometrica: Journal of the Econometric Society*, pages 955–969, 1983.

Kenneth Burdett and Guido Menzio. The (q, s, s) pricing rule. *The Review of Economic Studies*, 85(2):892–928, 2018.

Ariel Burstein, Vasco M Carvalho, and Basile Grassi. Bottom-up markup fluctuations. Technical report, National Bureau of Economic Research, 2020.

Luis Cabral and Jose Mata. On the evolution of the firm size distribution: facts and theory. *American Economic Review*, 93(4):1075–1090, 2003.

Andrea Caggese and Ander Perez-Orive. How stimulative are low real interest rates for intangible capital? *Finance and Economics Discussion Series*, 9, 2020.

Mr Yan Carriere-Swallow and Mr Vikram Haksar. *The economics and implications of data: an integrated perspective*. International Monetary Fund, 2019.

Vasco M Carvalho and Basile Grassi. Large firm dynamics and the business cycle. *American Economic Review*, 109(4):1375–1425, 2019.

Raj Chetty, Adam Guren, Day Manoli, and Andrea Weber. Are micro and macro labor supply elasticities consistent? a review of evidence on the intensive and extensive margins. *American Economic Review*, 101(3): 471–75, 2011.

Andrea Chiavari. The macroeconomics of rising returns to scale: Costumer acquisition, markups, and dynamism. 2021.

Andrea Chiavari and Sampreet Goraya. The rise of intangible capital and the macroeconomic implications. *Working Paper*, 2021.

Gian Luca Clementi and Berardino Palazzo. On the calibration of competitive industry dynamics models. *Unpublished working paper*, 2016a.

Gian Luca Clementi and Berardino Palazzo. Entry, exit, firm dynamics, and aggregate fluctuations. *American Economic Journal: Macroeconomics*, 8(3):1–41, 2016b.

Gian Luca Clementi and Berardino Palazzo. Investment and the cross-section of equity returns. *The Journal of Finance*, 74(1):281–321, 2019.

James Cloyne, Clodomiro Ferreira, Maren Froemel, and Paolo Surico. Monetary policy, corporate finance and investment. Technical report, National Bureau of Economic Research, 2018.

Alex Coad. *The growth of firms: a survey of theories and empirical evidence*. Edward Elgar Publishing, 2009.

Wesley M Cohen. Fifty years of empirical studies of innovative activity and performance. *Handbook of the Economics of Innovation*, 1:129–213, 2010.

Wesley M Cohen and Steven Klepper. The anatomy of industry r&d intensity distributions. *The American Economic Review*, pages 773–799, 1992.

Russell W Cooper and John C Haltiwanger. On the nature of capital adjustment costs. *The Review of Economic Studies*, 73(3):611–633, 2006.

Carol Corrado, Charles Hulten, and Daniel Sichel. Intangible capital and us economic growth. *Review of income and wealth*, 55(3):661–685, 2009.

Carol A Corrado and Charles R Hulten. How do you measure a"technological revolution"? *American Economic Review*, 100(2):99–104, 2010.

Nicolas Crouzet and Janice C Eberly. Intangible capital and the investment-q relation. *Proceedings of the 2018 Jackson Hole Symposium*, pages 87–148, 2019.

Joel M David and Venky Venkateswaran. The sources of capital misallocation. *American Economic Review*, 109(7):2531–67, 2019.

Steven J Davis, John Haltiwanger, Ron Jarmin, Javier Miranda, Christopher Foote, and Eva Nagypal. Volatility and dispersion in business growth rates: publicly traded versus privately held firms. *NBER Macroeconomics Annual*, 21:107–179, 2006.

Jan De Loecker and Paul T Scott. Estimating market power evidence from the us brewing industry. *NBER Working Paper*, 2016.

Jan De Loecker and Frederic Warzynski. Markups and firm-level export status. *American economic review*, 102(6):2437–71, 2012.

Jan De Loecker, Pinelopi K Goldberg, Amit K Khandelwal, and Nina Pavcnik. Prices, markups, and trade reform. *Econometrica*, 84(2): 445–510, 2016.

Jan De Loecker, Jan Eeckhout, and Gabriel Unger. The rise of market power and the macroeconomic implications. *The Quarterly Journal of Economics*, 135(2):561–644, 2020.

Jan De Loecker, Jan Eeckhout, and Simon Mongey. Quantifying market power and business dynamism in the macroeconomy. *NBER Working Paper*, 2021.

Maarten De Ridder. Market power and innovation in the intangible economy. *Working Paper*, 2019.

Ryan Decker, John Haltiwanger, Ron Jarmin, and Javier Miranda. The role of entrepreneurship in us job creation and economic dynamism. *Journal of Economic Perspectives*, 28(3):3–24, 2014.

Ryan A Decker, John Haltiwanger, Ron S Jarmin, and Javier Miranda. Declining business dynamism: what we know and the way forward. *American Economic Review*, 106(5):203–07, 2016.

Ryan A. Decker, John Haltiwanger, Ron S. Jarmin, and Javier Miranda. Changing business dynamism and productivity: shocks versus responsiveness. *American Economic Review*, 110(12): 3952–90, December 2020. doi: 10.1257/aer.20190680. URL https://www.aeaweb.org/articles?id=10.1257/aer.20190680.

Emin Dinlersoz, Sebnem Kalemli-Ozcan, Henry Hyatt, and Veronika Penciakova. Leverage over the life cycle and implications for firm growth and shock responsiveness. Technical report, National Bureau of Economic Research, 2018.

Emin M Dinlersoz and Mehmet Yorukoglu. Information and industry dynamics. *American Economic Review*, 102(2):884–913, 2012.

Avinash K Dixit and Joseph E Stiglitz. Monopolistic competition and optimum product diversity. *The American Economic Review*, 67(3): 297–308, 1977.

Mark Doms and Timothy Dunne. Capital adjustment patterns in manufacturing plants. *Review of economic dynamics*, 1(2):409–429, 1998.

Jean-Pierre Dubé, Günter J Hitsch, and Peter E Rossi. State dependence and alternative explanations for consumer inertia. *The RAND Journal of Economics*, 41(3):417–445, 2010.

Timothy Dunne, Mark J Roberts, and Larry Samuelson. The growth and failure of us manufacturing plants. *The Quarterly Journal of Economics*, 104(4):671–698, 1989.

Jonathan Eaton, Marcela Eslava, Maurice Kugler, and James R Tybout. *8. Export dynamics in colombia: firm-level evidence*. Harvard University Press, 2009.

Chris Edmond, Virgiliu Midrigan, and Daniel Yi Xu. How costly are markups? Technical report, National Bureau of Economic Research, 2018.

Jan Eeckhout. *The profit paradox: how thriving firms threaten the future of work*. Princeton University Press, 2021.

Liran Einav, Peter J. Klenow, Jonathan D Levin, and Raviv Murciano-Goroff. Customers and retail growth. *Working Paper*, 2020.

Andrea L Eisfeldt and Dimitris Papanikolaou. Organization capital and the cross-section of expected returns. *The Journal of Finance*, 68(4): 1365–1406, 2013.

Michael WL Elsby, Bart Hobijn, and Ayşegül Şahin. The decline of the us labor share. *Brookings Papers on Economic Activity*, 2013(2):1–63, 2013.

Michael Ewens, Ryan H Peters, and Sean Wang. Acquisition prices and the measurement of intangible capital. *NBER Working Paper*, (w25960), 2019.

Andrea Fabiani, Luigi Falasconi, and Janko Heineken. Monetary policy and corporate debt maturity. 2020.

John G Fernald. Productivity and potential output before, during, and after the great recession. *NBER Macroeconomics Annual*, 29(1):1–51, 2015.

Lucia Foster, John Haltiwanger, and Chad Syverson. Reallocation, firm turnover, and efficiency: selection on productivity or profitability? *American Economic Review*, 98(1):394–425, 2008.

Rosemary R Fullerton, Cheryl S McWatters, and Chris Fawson. An examination of the relationships between jit and financial performance. *Journal of Operations Management*, 21(4):383–404, 2003.

Jordi Galí. *Monetary policy, inflation, and the business cycle: an introduction to the new Keynesian framework and its applications*. Princeton University Press, 2015.

Jordi Gali, Mark Gertler, and J David Lopez-Salido. Markups, gaps, and the welfare costs of business fluctuations. *The review of economics and statistics*, 89(1):44–59, 2007.

Amit Gandhi, Salvador Navarro, and David A Rivers. On the identification of gross output production functions. *Journal of Political Economy*, 128(8):2973–3016, 2020.

Wei Gao and Matthias Kehrig. Returns to scale, productivity and competition: empirical evidence from us manufacturing and construction establishments. *Working Paper*, 2017.

Mark Gertler and Simon Gilchrist. Monetary policy, business cycles, and the behavior of small manufacturing firms. *The Quarterly Journal of Economics*, 109(2):309–340, 1994.

Simon Gilchrist, Raphael Schoenle, Jae Sim, and Egon Zakrajšek. Inflation dynamics during the financial crisis. *American Economic Review*, 107(3):785–823, 2017.

Avi Goldfarb and Daniel Trefler. Ai and international trade. *NBER Working Paper*, 2018.

Mikhail Golosov and Robert E Lucas. Menu costs and phillips curves. *Journal of Political Economy*, 115(2):171–199, 2007.

Francois Gourio and Leena Rudanko. Customer capital. *Review of Economic Studies*, 81(3):1102–1136, 2014.

Basile Grassi. I-o in io: size, industrial organization, and the input-output network make a firm structurally important. *Working Paper*, 2017.

Zvi Grilliches. R&d and productivity: Econometric results and measurement issues. *Handbook of Economics of Innovation and Technological Change, Oxford*, pages 52–89, 1995.

Gustavo Grullon, Yelena Larkin, and Roni Michaely. Are us industries becoming more concentrated? *Review of Finance*, 23(4):697–743, 2019.

Refet S Gürkaynak, Brian Sack, and Eric Swanson. The sensitivity of long-term interest rates to economic news: Evidence and implications for macroeconomic models. *American economic review*, 95(1):425–436, 2005.

Germán Gutiérrez and Thomas Philippon. Investment-less growth: An empirical investigation. Technical report, National Bureau of Economic Research, 2016.

Robert E Hall. The relation between price and marginal cost in us industry. *Journal of Political Economy*, 96(5):921–947, 1988.

Robert E Hall. Quantifying the lasting harm to the us economy from the financial crisis. *NBER Macroeconomics Annual*, 29(1):71–128, 2015.

John Haltiwanger, Ron S Jarmin, and Javier Miranda. Who creates jobs? small versus large versus young. *Review of Economics and Statistics*, 95(2):347–361, 2013.

Jonathan Haskel and Stian Westlake. *Capitalism without capital: the rise of the intangible economy*. Princeton University Press, 2018.

Sungki Hong. Customer capital, markup cyclicality, and amplification. *FRB St. Louis Working Paper*, (2017-33), 2017.

Hugo Hopenhayn and Richard Rogerson. Job turnover and policy evaluation: A general equilibrium analysis. *Journal of political Economy*, 101(5):915–938, 1993.

Hugo Hopenhayn, Julian Neira, and Rish Singhania. The rise and fall of labor force growth: implications for firm demographics and aggregate trends. *NBER Working Paper*, 2018.

Hugo A Hopenhayn. Entry, exit, and firm dynamics in long run equilibrium. *Econometrica: Journal of the Econometric Society*, pages 1127–1150, 1992.

Hugo A Hopenhayn. Firms, misallocation, and aggregate productivity: A review. *Annu. Rev. Econ.*, 6(1):735–770, 2014.

Andreas Hornstein, Per Krusell, and Giovanni L Violante. Frictional wage dispersion in search models: a quantitative assessment. *American Economic Review*, 101(7):2873–98, 2011.

Herbert Hovenkamp. Is antitrust's consumer welfare principle imperiled? *J. Corp. L.*, 45:65, 2019.

Herbert Hovenkamp. Antitrust: what counts as consumer welfare. *Working Paper*, 2020a.

Herbert Hovenkamp. On the meaning of antitrust's consumer welfare principle. *Revue Concurrentialiste (Jan. 17, 2020), U of Penn, Inst for Law & Econ Research Paper*, (20-16), 2020b.

Chang-Tai Hsieh and Peter J Klenow. Misallocation and manufacturing tfp in china and india. *The Quarterly journal of economics*, 124(4): 1403–1448, 2009.

Chang-Tai Hsieh and Esteban Rossi-Hansberg. The industrial revolution in services. *NBER Working Papers*, 2019.

Marek Jarociński and Peter Karadi. Deconstructing monetary policy surprises-the role of information shocks. *American Economic Journal: Macroeconomics*, 12(2):1–43, 2020.

Priit Jeenas. Firm balance sheet liquidity, monetary policy shocks, and investment dynamics. 2019.

Charles I Jones and Christopher Tonetti. Nonrivalry and the economics of data. *American Economic Review*, 110(9):2819–58, 2020.

Òscar Jordà. Estimation and inference of impulse responses by local projections. *American economic review*, 95(1):161–182, 2005.

Greg Kaplan and Guido Menzio. The morphology of price dispersion. *International Economic Review*, 56(4):1165–1206, 2015.

Greg Kaplan and Piotr Zoch. Markups, labor market inequality and the nature of work. Technical report, National Bureau of Economic Research, 2020.

Greg Kaplan, Benjamin Moll, and Giovanni L Violante. Monetary policy according to hank. *American Economic Review*, 108(3):697–743, 2018.

Loukas Karabarbounis and Brent Neiman. The global decline of the labor share. *The Quarterly journal of economics*, 129(1):61–103, 2013.

Fatih Karahan, Benjamin Pugsley, and Ayşegül Şahin. Demographic origins of the startup deficit. *NBER Working Paper*, 2019.

Matthias Kehrig and Nicolas Vincent. The micro-level anatomy of the labor share decline. *The Quarterly Journal of Economics*, 136(2):1031–1087, 2021.

Aubhik Khan and Julia K Thomas. Idiosyncratic shocks and the role of nonconvexities in plant and aggregate investment dynamics. *Econometrica*, 76(2):395–436, 2008.

Lina M Khan. Amazon's antitrust paradox. *Yale lJ*, 126:710, 2016.

Miles S Kimball. The quantitative analytics of the basic neomonetarist model. *Journal of Money, Credit and Banking*, 27(4):1241–1277, 1995.

Peter J Klenow and Jonathan L Willis. Real rigidities and nominal price changes. *Economica*, 83(331):443–472, 2016.

Tor Jakob Klette and Frode Johansen. Accumulation of r&d capital and dynamic firm performance: a not-so-fixed effect model. In *The Economics and Econometrics of Innovation*, pages 367–397. Springer, 2000.

Tor Jakob Klette and Samuel Kortum. Innovating firms and aggregate innovation. *Journal of political economy*, 112(5):986–1018, 2004.

Dongya Koh, Raül Santaeulàlia-Llopis, and Yu Zheng. Labor share decline and intellectual property products capital. *Econometrica*, 88(6): 2609–2628, 2020.

Anton Korinek, Ding Xuan Ng, and Johns Hopkins. Digitization and the macro-economics of superstars. *Working Paper*, 2018.

Kyle Kost, Jeremy Pearce, and Liangjie Wu. Market power through the lens of trademarks. *Working Paper*, 2019.

Danial Lashkari, Arthur Bauer, and Jocelyn Boussard. Information technology and returns to scale. *Working Paper*, 2021.

Yoonsoo Lee and Toshihiko Mukoyama. Productivity and employment dynamics of us manufacturing plants. *Economics Letters*, 136:190–193, 2015.

Baruch Lev and Feng Gu. *The end of accounting and the path forward for investors and managers*. John Wiley & Sons, 2016.

James Levinsohn and Amil Petrin. Estimating production functions using inputs to control for unobservables. *The Review of Economic Studies*, 70(2):317–341, 2003.

Daniel Levy, Mark Bergen, Shantanu Dutta, and Robert Venable. The magnitude of menu costs: direct evidence from large us supermarket chains. *The Quarterly Journal of Economics*, 112(3):791–824, 1997.

Ernest Liu, Atif Mian, and Amir Sufi. Low interest rates, market power, and productivity growth. *Econometrica forthcoming*, 2020.

Ioana Marinescu and Herbert Hovenkamp. Anticompetitive mergers in labor markets. *Ind. LJ*, 94:1031, 2019.

Joseba Martinez. Automation, growth and factor shares. *Working Paper*, 2018.

Ellen R McGrattan and Edward C Prescott. Technology capital and the us current account. *American Economic Review*, 100(4):1493–1522, 2010a.

Ellen R McGrattan and Edward C Prescott. Unmeasured investment and the puzzling us boom in the 1990s. *American Economic Journal: Macroeconomics*, 2(4):88–123, 2010b.

Ellen R McGrattan and Edward C Prescott. A reassessment of real business cycle theory. *American Economic Review*, 104(5):177–82, 2014.

Alisdair McKay, Emi Nakamura, and Jón Steinsson. The power of forward guidance revisited. *American Economic Review*, 106(10): 3133–58, 2016.

Matthias Meier and Timo Reinelt. Monetary policy, markup dispersion, and aggregate tfp. 2020.

Guido Menzio and Shouyong Shi. Block recursive equilibria for stochastic models of search on the job. *Journal of Economic Theory*, 145(4): 1453–1494, 2010.

Guido Menzio and Shouyong Shi. Efficient search on the job and the business cycle. *Journal of Political Economy*, 119(3):468–510, 2011.

Guido Menzio and Nicholas Trachter. Equilibrium price dispersion with sequential search. *Journal of Economic Theory*, 160:188–215, 2015.

Guido Menzio and Nicholas Trachter. Equilibrium price dispersion across and within stores. *Review of Economic Dynamics*, 28:205–220, 2018.

Espen R Moen. Competitive search equilibrium. *Journal of Political Economy*, 105(2):385–411, 1997.

Simon Mongey. Market structure and monetary non-neutrality. *Job market paper. http://www. simonmongey. com/uploads/6/5/6/6/65665741/mongey_market_structure_monetary_nonneutrality_draft.pdf*, 2017.

Monica Morlacco and David Zeke. Monetary policy, customer capital, and market power. *Journal of Monetary Economics*, 2021.

Masao Nakamura, Sadao Sakakibara, and Roger Schroeder. Adoption of just-in-time manufacturing methods at us-and japanese-owned plants: some empirical evidence. *IEEE transactions on engineering management*, 45(3):230–240, 1998.

Christopher J Nekarda and Valerie A Ramey. The cyclical behavior of the price-cost markup. *Journal of Money, Credit and Banking*, 52(S2): 319–353, 2020.

Nathan Newman. Search, antitrust, and the economics of the control of user data. *Yale J. on Reg.*, 31:401, 2014.

G Steven Olley and Ariel Pakes. The dynamics of productivity in the telecommunications equipment industry. *Econometrica*, 65(6):1263–1297, 1996.

Jane Olmstead-Rumsey. Market concentration and the productivity slowdown. *Working Paper*, 2019.

Pablo Ottonello and Thomas Winberry. Financial heterogeneity and the investment channel of monetary policy. *Econometrica*, 88(6):2473–2502, 2020.

Luigi Paciello, Andrea Pozzi, and Nicholas Trachter. Price dynamics with customer markets. *International Economic Review*, 60(1):413–446, 2019.

Ariel Pakes. *Dynamic structural models, problems and prospects: mixed continuous discrete controls and market interactions*, volume 2 of *Econometric Society Monographs*, pages 171–274. Cambridge University Press, 1994. doi: 10.1017/CCOL0521444608.005.

Michael Peters. Heterogeneous markups, growth, and endogenous misallocation. *Econometrica*, 88(5):2037–2073, 2020.

Michael Peters and Conor Walsh. Declining dynamism, increasing markups and missing growth: the role of the labor force. *Working Paper*, 2019.

Ryan H Peters and Lucian A Taylor. Intangible capital and the investment-q relation. *Journal of Financial Economics*, 123(2):251–272, 2017.

Thomas Philippon. *The great reversal*. Harvard University Press, 2019.

Annette Ptok, Rupinder P Jindal, and Werner J Reinartz. Selling, general, and administrative expense (sga)-based metrics in marketing: conceptual and measurement challenges. *Journal of the Academy of Marketing Science*, 46(6):987–1011, 2018.

Benjamin W Pugsley, Petr Sedlacek, and Vincent Sterk. The nature of firm growth. *Available at SSRN 3086640*, 2019.

Pau Roldan-Blanco and Sonia Gilbukh. Firm dynamics and pricing under customer capital accumulation. *Journal of Monetary Economics*, 2020.

Kim J Ruhl and Jonathan L Willis. Convexities, nonconvexities, and firm export behavior. In *Manuscript, Midwest Macro Conference, Philadelphia*, 2008.

Edouard Schaal. Uncertainty and unemployment. *Econometrica*, 85(6): 1675–1721, 2017.

Chad Syverson. Market structure and productivity: a concrete example. *Journal of Political Economy*, 112(6):1181–1222, 2004.

John B Taylor. The robustness and efficiency of monetary policy rules as guidelines for interest rate setting by the european central bank. *Journal of Monetary Economics*, 43(3):655–679, 1999.

Joshua Weiss. Intangible investment and market concentration. *Working Paper*, 2019.

Arlene Wong. Refinancing and the transmission of monetary policy to consumption. *Unpublished manuscript*, 20, 2019.

Mark J Zbaracki, Mark Ritson, Daniel Levy, Shantanu Dutta, and Mark Bergen. Managerial and customer costs of price adjustment: direct evidence from industrial markets. *Review of Economics and statistics*, 86(2):514–533, 2004.

Lichen Zhang. Intangibles, concentration, and the labor share. *Mimeo*, 2019a.

Lichen Zhang. Intangible-investment-specific technical change, concentration and labor share. *Working Paper*, 2019b.