Entrepreneurship Across Cities: Uncovering Policy Implications*

Marko Irisarri †

November 18, 2024
(Preliminary and Incomplete)
Latest version: here

Abstract

Are entrepreneurs and their capital allocated optimally across space? Should governments employ place-based entrepreneurial policies? To study these questions, we first develop a dynamic spatial quantitative framework featuring financial frictions, dynamic capital accumulation, occupational and location choice and agglomeration forces. We then take this model to the largest 20 Urban Areas (UA) of the Spanish State by relying on rich administrative and balance sheet data. A key prediction of the model, which the data supports, is that there are heterogeneous returns to capital across space, and more productive UAs are more capital constrained. Intuitively, financial frictions, albeit symmetric, disproportionately hinder more productive UAs from reaching their production frontier. Second, we provide an efficient solution method by exploiting the parallel nature of GPUs in CUDA. Speed-ups in the range of 60 to 20,000 are obtained compared to standard methods. Third, the policy analysis suggests that, compared to a spatially neutral policy, targeting a subset of the most productive UAs achieves greater welfare and production gains. However, these policies pose a trade-off to policymakers between aggregate gains and increased regional disparities.

Keywords: Spatial Dynamics, Place-based Policies, Entrepreneurship, Financial Frictions

JEL Codes: R12, R13, C63, E21, E23

*I am grateful to Andrea Caggese and David Nagy for their constant guidance and support, and to Edouard Schaal, Elisa Giannone, Giacomo Ponzetto, Priit Jeenas and Andrea Sy for substantial comments that have improved this paper. I thank Gianmarco Ruzzier, Alejandro Rábano, Madalena Gaspar, Robert Wojciechowski, Zidong Lin and participants at CREi's International and Macro seminars for all their valuable comments. All errors are my own.

†Universitat Pompeu Fabra and Barcelona School of Economics. Email: marko.irisarri@upf.edu.

1 Introduction

Entrepreneurship is widely recognized as an engine of job creation and growth, and has consequently gained the support of policymakers. [1] Currently, expenditure in entrepreneurial subsidies amounts to 0.74% of GDP in the EU and 0.65% in the US. In the Spanish economy, which is the focus of this paper, specific policies include a flat monthly social security contribution of 80€ for the first year. [2] A primary motivating factor behind these subsidies are the financial frictions faced by entrepreneurs ((Banerjee and Duflo, 2014), (Schmalz et al., 2017), among others).

Despite this support, there is no consensus on how these funds should be allocated. However, entrepreneurship varies considerably across space, as it is shaped by the composition of the local workforce and agglomeration forces. Thus, a natural question arises: should policymakers account for the spatial dimension when designing entrepreneurial policies? Studying this question calls for a framework with occupational and location choice, endogenous capital accumulation, agglomeration forces and financial frictions. Thus far, neither the spatial nor the entrepreneurship literatures have provided an answer, as they have not explored the spatial misallocation of entrepreneurship and their capital due to the dimensionality of the problem.

To this end, we present a dynamic spatial quantitative framework featuring these occupational and location choices, financial frictions, endogenous capital accumulation and agglomeration forces. We take the model to the largest 20 urban areas (UA henceforth) of the Spanish State using rich administrative-level panel data on households and balance-sheet data on firms. A key prediction of the model, which estimates from the data on support, is that there are heterogeneous returns to capital across space, and more productive UA are more capital constrained. Intuitively, even in the presence of symmetric financial frictions across locations, since these are more binding for the more productive entrepreneurs in the more productive locations, this leads to more productive UAs operating disproportionately below their production frontier. We tackle the com-

[1] In the European Union (EU), for example, where growing concerns regarding the widening gap in productivity and innovation between itself and both the USA and China are ubiquitous (the recently published Draghi report embodies precisely this perception) fostering entrepreneurship has been identified as a key policy instrument to ameliorate said concerns.

[2] The figures for the EU and USA are as reported in the Draghi report for R&D expenditure. Other specific examples in the Spanish economy include: (i) the ability to capitalize unemployment benefits in one payment (ii) in several regions, a lump-sum transfer averaging 3,000€ (varies by age, gender, urbanization, and other factors). To provide context on the monthly 80€ flat contribution, the standard minimum one is 300€ and on average entrepreneurs need to pay 28.3% of their net income as social security contributions.

putational challenge of solving this class of models by proposing a solution method that exploits the parallelizable nature of the algorithms employed in macroeconomics along with a hardware component specifically optimized around throughput and parallel tasks: GPUs (Graphics Processing Units). The attained speed-ups from employing GPUs are up to 20,000, higher than the previously reported two to three times faster in the literature. [3] Lastly, we use our framework to explore the long-run policy implications of place-based entrepreneurial policies. Our results suggest that compared to a spatially neutral policy, targeting entrepreneurial subsidies towards a subset of the most productive Urban Areas leads to higher welfare and production gains. However, these policies pose a trade-off to policy-makers between aggregate gains and increased regional disparities.

The first section of the paper briefly presents three stylized facts that motivate the model ingredients. The first one provides empirical evidence of the substantial variation in incorporated entrepreneurial rates across space, thus highlighting the spatial nature of entrepreneurship. The second one shows the heterogeneity in the correlation between entrepreneurial rates and agglomeration force proxies (density and skill composition) across sectors of the economy, which motivates the inclusion of heterogeneous agglomeration forces by sector. Lastly, the third provides suggestive evidence on the correlation between individual resources and entry into entrepreneurship, which along with the vast evidence on financial frictions in the literature motivates the financial constraints in the model.

The second section presents a dynamic spatial model that combines key elements from the spatial and entrepreneurship literatures. It thus features heterogeneous workers and entrepreneurs along the assets, entrepreneurial productivity, skill, sector and location dimensions who are mobile across space and endogenously accumulate capital. Since we are interested in entrepreneurship, the occupational choice is a central component. Workers provide labour in the local labour markets and entrepreneurs produce intermediate sectoral goods which are combined to form the final good subject to agglomeration forces and financial frictions [4] . The latter stem in the model

[3] This is with respect to the single-core C++ implementation in Fernández-Villaverde and Valencia (2018). We take this paper as the main reference as it has the most comprehensive set of languages and has publicly available code for replication.

[4] We corroborate empirically key findings in the literature that motivate these elements in the model. To this end, we leverage rich administrative-level data on the employment histories of individuals across time and space from the MCVL (Muestra Continua de Vidas Laborales - Continous Sample of Employment Histories) for the period 2013-2018 in the Spanish Economy.

from limited enforceability in the capital rental market, which leads to a borrowing constraint as a function of the assets of the entrepreneur. Sectors are heterogeneous in their sensitivity to the agglomeration forces, degree of financial frictions, and their intensity of the use of the inputs in the economy. Locations differ in their productivity, housing supply elasticity and amenities [5]. A key challenge to solve this class of models, as in Giannone et al. (2020), is that the rich individual heterogeneity, combined with multiple locations, generates a high dimensional state space. Moreover, the spatial dimension imposes additional structure by requiring agents to make forward-looking dynamic decisions while keeping track of the distribution of prices within and across locations. An additional consideration is that the presence of non-convexities (occupational choice) and the highly non-linear policy functions stemming from the borrowing constraints of the entrepreneurs prevent us from employing other methods such as that in Kleinman et al. (2023), which relies on linearization. [6]

The third section discusses the employed numerical methodology. To solve and calibrate the model, plus conduct policy experiments efficiently, we make extensive use of GPUs (Graphics Processing Units), whose distinguishing feature is the availability of thousands (even tens of thousands) of cores. The intuition is simple: as an example, to sum $c[i] = x[i] + y[i]$ for $i = 0, ..., N$, rather than iterating sequentially on $i = 0$, $i = 1$, $i = 2$ ..., each index of the vector i is assigned to a core c on the GPU. This allows the entire vector sum to be computed in parallel, leading to substantial speed-ups. In particular, we are able to compute the model up to 20,000 times faster than a standard fully-vectorized MATLAB implementation would do [7], and the proposed custom implementations in native CUDA (the language employed to program on GPUs), attain higher speed-ups than previously reported in the literature (Fernández-Villaverde and Valencia, 2018) (x20,000 against x27-28). We are careful to explain and decompose the sources of these speed-ups in sections 4 and Appendix D.

In the fourth section, the model is taken to the largest 20 UAs of the Spanish State which account for roughly 50% of the population and above 70% of GDP by relying on rich micro data

[5] A summary of the main blocks of the model is provided in table 1

[6] Note that in this model the heterogeneity in returns to capital is critical in order to study the effect of policies on aggregate outcomes.

[7] To put this in perspective, it takes around 4 minutes to compute a place-based policy in general equilibrium so that it results in a given % expenditure out of the country-wide GDP. This would require 56 days in MATLAB. At an electricity cost of 0.25$/kWh, that would result in 0.00396$ in CUDA and 31.92$ in MATLAB.

from both the household side and the firm side. On the households side, administrative-level panel data on social security affiliation episodes is employed, the MCVL (Continuous Sample of Employment Histories). One key advantage from this dataset is the ability to identify who, when and where becomes an incorporated entrepreneur since 2013. On the firm side, given that the MCVL provides no data on the firms of the entrepreneurs, balance sheet data from SABI (Iberian Balance Analysis System) is employed, which offers the required geographical coverage. The population share and skill composition of each location are closely matched, as they regulate the agglomeration forces and hence productivity. On the untargeted dimensions, the calibrated model can properly replicate the heterogeneous slopes across sectors in the correlation between the stock of entrepreneurs or entry rates and both the size and skill composition of UAs. In addition, the mobility patterns in the model match those of the data: agents are more likely to out-migrate from smaller, less productive urban areas to their larger and more productive counterparts.

The fifth section discusses the main predictions of the model regarding the interaction between financial frictions, endogenous capital accumulation and heterogeneous locations. At the individual level, entrepreneurs face collateral constraints and must therefore accumulate wealth in order to reach the optimal production scale. Spatially, they sort across locations in order to maximize their profits, with higher productivity locations providing the fastest path for wealth accumulation. At the aggregate level, there are heterogeneous returns to capital across space, and more productive UAs have higher MPK Premiums [8]. The main driver behind this result is the intensive margin. Given the symmetric borrowing constraints across locations, a higher productivity of the urban area widens the gap between the attainable capital and the unconstrained one, thus increasing the MPK Premiums for given idiosyncratic asset holdings and productivities across the entire distribution. [9] In other words, the gap between the attained production and the production frontier, both at the individual and aggregate level, is wider in more productive UAs given the

[8] MPK Premiums are defined as the difference between the MPK and the user cost of capital. Intuitively, a positive value implies that entrepreneurs would be willing to pay a higher user cost of capital to have increased access to capital.

[9] The second force is compositional. On the extensive margin, the assumption on tradable goods leads to a negative selection channel between the idiosyncratic productivity of the entrepreneur and the productivity of the UA. Given the log-normal distribution assumption on the idiosyncratic productivity of the entrepreneurs, the marginal entrepreneur is notably less productive idiosyncratically, thus muting the benefits obtained from the higher productivity of the UA. Thus, we find no clear evidence that the marginal entrepreneur in a more productive UA is more constrained than its counterpart in a less productive UA. The bulk in the variation in MPK Premiums across UAs is therefore explained by the intensive margin, leading to a positive relationship between average UA productivity and average MPK Premiums at the UA level.

more pronounced impact of financial frictions.

Lastly, motivated by the obtained heterogeneous returns to capital across locations, the seventh section studies place-based entrepreneurial transfers. We begin with a place-based lump-sum transfers to entrepreneurs policy that mimicks the current EU's 0.1% of GDP expenditure level. [10] This policy is subsidized by an increase in the labour tax. Both in terms of efficiency and welfare, at country-wide level, targeting the most constrained and productive UAs leads to the largest positive gains. Intuitively, these are the locations where the additional resources spent on capital have the higher returns for both entrepreneurs and workers, as wages increase as a result of the increased productivity. However, the increase in labour taxes to subsidize a single location as well as the increased house prices due to congestion forces lead to a more muted response of welfare compared to that of output. These policies targeted at a single UA, while they can achieve higher efficiency and welfare gains than a country-wide untargeted policy, result in a notable increase in regional economic disparities.

Therefore, the targeting of a set of UAs is considered. Since this is an expensive combinatorial problem at 2^{20} possible sets, a heuristic rule consisting on progressively targeting the UAs by size is followed. The results suggest that targeting a subset of the largest and most productive UAs leads to higher welfare and efficiency gains than an untargeted policy at little cost in terms of regional disparities. Intuitively, expanding the set of covered UAs expands the mass of agents directly benefiting from the policy and mitigates the congestion forces by spreading the influx of people across multiple locations.

Does the design of the policy matter? To address this question, an alternative proportional-to-collateral policy is considered which consists on handing out entrepreneur-specific transfers to alleviate the collateral borrowing constraints in equal proportion. This is relevant since it speaks to size-dependent policies such as investment subsidies, loans or guarantees. We find that such policy leads to a small and positive efficiency response. The intuition stems from the fact that this policy concentrates most of the funds on larger, less constrained entrepreneurs. In terms of welfare, the regressive nature of taxing workers to subsidize these on average larger and less constrained entrepreneurs (compared to lump-sum transfers) at limited wage growth and extensive

[10] This is the expenditure level on the Horizon Europe programme and EIC (European Innovation Council) institution.

margin occupation switching leads to negative welfare gains across the board. This holds regardless of what set of locations is targeted or what the expenditure level of the policy is.

Lastly, to underline the role of financial frictions, we conduct the same policy experiments under no borrowing constraints. In this scenario, firms operate at optimal scale everywhere, entry is determined by productivity only and there are no heterogeneous returns to capital. The results indicate that increasing the labour tax in order to subsidize entrepreneurs leads in this case to a negative response of welfare. Intuitively, workers pay higher taxes to subsidize wasteful expenditure [11] targeted towards on average richer entrepreneurs. In terms of production, there is increased inefficient entry into entrepreneurship as a consequence of the subsidies. Consequently, the marginal entrepreneur is now less productive, and the average entrepreneur operates at a lower scale. At the aggregate, the extensive margin dominates and this results in limited but positive gains in total production. Therefore, financial frictions are key in generating a positive response of both efficiency and welfare and quantitatively drive most of the efficiency gains (>85%) that are attainable through place-based entrepreneurial subsidies. [12]

Relationship to the literature

This work relates to several strands of the literature. First, it is related to the literature that studies misallocation across space either due to zoning, taxation, or local externalities Desmet and Rossi-Hansberg (2013), Ferrari and Ossa (2023), Fajgelbaum et al. (2019), Herkenhoff et al. (2018), Niebuhr et al. (2020), Fajgelbaum and Gaubert (2020). We contribute to this literature by studying the implications of the misallocation of capital across space induced by financial frictions in an spatial setting.

Related to this, we contribute to the growing body of dynamic spatial models Kleinman et al. (2023), Giannone et al. (2020), Desmet et al. (2018) by providing an efficient solution method and studying capital accumulation subject to frictions in an occupational choice setting. We also show that dynamics and space interact in a non-trivial manner for the wealth accumulation of entrepreneurs: more productive entrepreneurs sort across space to the locations that allow them to accumulate capital faster.

[11] Note that in this setting place-based entrepreneurial policies could still address a potentially suboptimal allocation of agents due to the agglomeration forces.

[12] In the absence of financial frictions, optimal policy would shift to place and skill based transfers to agents (both workers and entrepreneurs) to optimally exploit the agglomeration forces, as in Rossi-Hansberg et al. (2019).

We also relate to the literatures on the macroeconomics of entrepreneurship and its policy implications Quadrini (2000), Cagetti and De Nardi (2006), Boar and Midrigan (2023), Morazzoni and Sy (2022) and financial frictions Evans and Jovanovic (1989), Holtz-Eakin et al. (1994), Banerjee and Duflo (2014), Schmalz et al. (2017), Banerjee and Blickle (2021), by exploring the spatial dimension of the interaction between financial frictions and entrepreneurship. We find that heterogeneous returns are not only created across individuals, but also across locations, leading to a joint distribution of returns to capital over individuals and space. This has clear implications for policy: there are heterogeneous returns to capital across space and public spending in more constrained urban areas attains higher marginal returns.

Lastly, we contribute to the literature on computational economics arguing for the use of GPUs Fernández-Villaverde and Valencia (2018), Aldrich (2014), Aldrich et al. (2011). Here, we provide a full-fledged custom implementation in native CUDA/C++ that allows the model to be solved, from calibration to policy exercises, within this CPU-GPU heterogeneous model. We also discuss the benefits of the newly introduced features in CUDA, like the Cooperative Groups API for within-kernel thread synchronization. In addition, we find higher relative gains with respect to either a MATLAB or serial C++ implementation than previously reported in Fernández-Villaverde and Valencia (2018) (which offers the most comprehensive set of languages and open source code). In particular, we find a speed-up of x20,000 relative to MATLAB compared to x27 and a speed-up relative to C++ of 2,000 compared to x2-3 . Lastly, we discuss two algorithms to understand when the parallelization works more satisfactorily and we break down the speed-up gains by idiosyncrasy of the CUDA programming model.

This paper is organized as follows: Section 2 provides suggestive evidence for the assumed model ingredients. Section 3 presents a Dynamic Spatial model with Occupational Choice. Section 4 discusses the solution methodology. Section 5 discusses the calibration procedure and performance of the model on targeted and untargeted moments. Section 6 builds the intuition for the policy exercises by discussing key results from the steady-state. Section 7 presents the results of the policy exercises. Lastly, 8 concludes.

2 Stylized facts

The present section discusses the three main stylized facts that motivate the modelling choices in the quantitative model. The data employed for this section is the MCVL (Continous Sample of Employment Histories). This is an administrative-level panel dataset on social security affiliation episodes which allows us to identify who, when and where becomes an incorporated entrepreneur since 2013. The empirical appendix A provides the details on the methodology, data source and variable construction.

2.1 Stylized Fact #1: the substantial Spatial Variation of Incorporated Entrepreneurial Rates in the Spanish State

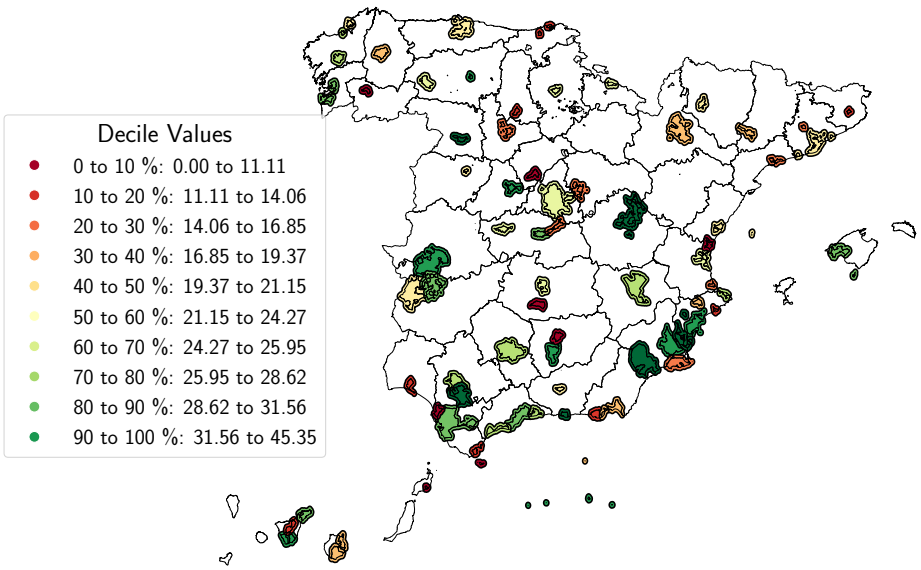
A consistent finding in the empirical literature studying the relationship between local economic characteristics (human capital, social norms towards entrepreneurship, industrial composition, ...) and entrepreneurship ((Eriksson and Rataj, 2019), (Hundt and Sternberg, 2016), (Ghani et al., 2017), among others) is the differential rates in entrepreneurship across space. This first stylized fact thus aims to provide evidence on the spatial variation of incorporated entrepreneurship in the Spanish State. Figure 1 displays the obtained results. We focus on incorporated entrepreneurship rather than any type of self-employment. This stems from evidence in the literature pointing towards this group showing higher employment, income and growth potential than their unincorporated counterparts ((Levine and Rubinstein, 2017), (Åstebro and Tåg, 2017), ...).

Two major comments are in place regarding the map. First, it is worthwhile to note that becoming an incorporated entrepreneur is a rather unlikely event. On average (throughout the 2013-2018 period), about 25 people out of 10,000 became one at the median urban area (UA) by year. [13] [14] The main point to notice is the substantial spatial variation of these entrepreneurial rates. The obtained $\frac{p_{90}}{p_{10}}$ ratio is about three, which implies that the top 10% UAs experience triple the incorporated entrepreneurship than their bottom 10% counterparts. Moreover, this spatial distribution is greatly amplified when looking at the conditional distributions by sector of the economy (see

[13] While comparisons across countries in the rates of incorporated entrepreneurs are not widely available, this figure is in line with the 10 to 15 out of 10,000 rate reported in the Kauffman New Employer Businesses series `https://indicators.kaufn.org/indicator/rate-of-new-employer-businesses`

[14] Appendix A.1.2 provides the details on identifying these entrepreneurs in the data and well as their individual characteristics, and Appendix A.1.3 provides details on the Urban Areas under study.

Figure 1: Spatial Distribution of Incorporated Entrepreneurship



Notes: Spatial distribution of incorporated entrepreneurship. The rate is defined as those transitioning from workers in year $t-1$ to those starting an incorporated business in year t over 10,000 working-age inhabitants per Urban Area over the period 2013-2018. The colours correspond to a given decile, whose values are provided on the legend to the left.

Appendix B.1). For example, in high skilled sectors (such as the “Telecommunications" and “Professional and Scientific"), this ratio explodes as some UAs (small ones in particular) experience no entrepreneurship into these sectors at all during this period.

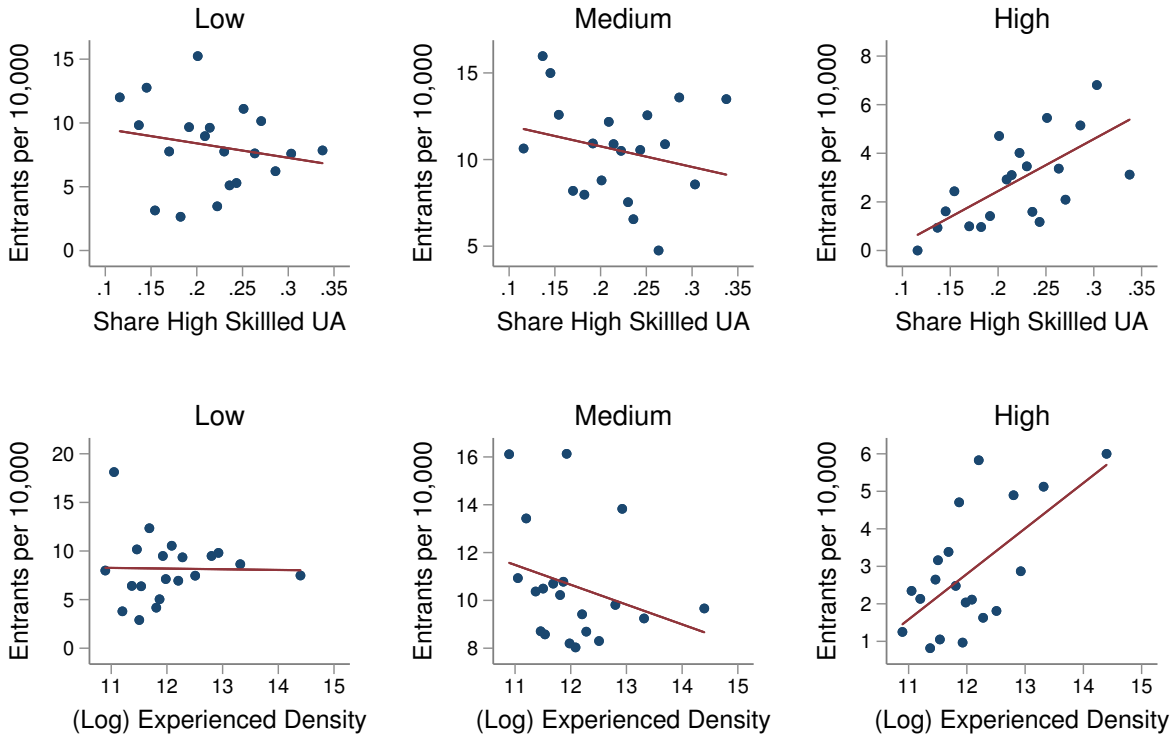
2.2 Stylized Fact #2: Sectoral Heterogeneity in the Correlation between Incorporated Entrepreneurship and the Local Agglomeration Forces

The empirical literature has also has long documented the relationship between the density of a UA and its rate of entrepreneurship ((Carlino et al., 2007), (Eriksson and Rataj, 2019), (Glaeser and Gottlieb, 2009)). Overall, agglomeration economies bring with them the benefits of infrastructure and input sharing, knowledge spillover effects and matching, among others, which contribute towards the UA’s productivity [15]. The main goal of this second point is to verify that this cor-

[15]Input sharing allows firms to access specialized suppliers and services at lower costs, while labor pooling enhances the efficiency of matching between firms and workers (Duranton and Puga, 2004). Knowledge spillovers—the informal exchange of ideas and innovations facilitated by proximity—are particularly crucial for fostering creativity and new

relation holds in the context of the Spanish UAs. Figure 2 displays the obtained results by sectors of the economy. The construction of these sectors is detailed in Appendix A.1.4. Following Rossi-Hansberg et al. (2019) and Giannone (2017), our two proxies of agglomeration forces are the experienced density [16] (Duranton and Puga, 2020), see Appendix A.1) and skill composition of the UAs.

Figure 2: Correlation between Incorporated Entrepreneurship and the Proxies of the Agglomeration Forces by UA and sector



Notes: Correlation between incorporated entrepreneurship and the proxies for the agglomeration forces, share of high skilled and (log) experienced density, at the UA-sector level (averaged over 2013-2018). The top row presents the results for the different sectors (columns) as a function of the share of high skilled by UA. The bottom row does similarly for the density of the UA. The estimated coefficients are available at B.7.

Upon inspection, one feature is salient: the heterogeneous response by sector. While the correlation between the proxies for the agglomeration forces and entry into the Low or Medium

business formation (Rosenthal and Strange, 2004).

[16] This is a measure of density that aims to control for the artificial administrative boundaries that delimit a given territory. Rather than computing the usual population to terrain ratio, this definition measures density by computing how many people there are on average in a 10km radius from the average person in an UA. We estimate this measure of the UAs of the Spanish State by relying on grid cell population data.

Skilled sectors is either flat or downward sloping, the results for the High Skilled sectors are non-equivocal: there exists a positive and significant relationship between incorporated entry and both the size and skill level of the UA for the highly skilled 1st digit NACE sectors. These findings motivate the heterogeneous agglomeration forces by sector in the model. The estimated coefficients are reported in Appendix B.2.

2.3 Stylized Fact #3: Positive Correlation between Individual Resources and Entry into Entrepreneurship

The literature on entrepreneurship has long emphasized the role of financial frictions in shaping entrepreneurial dynamics. A substantial body of work documents both the presence and consequences of these frictions ((Evans and Jovanovic, 1989), (Holtz-Eakin et al., 1994), (Banerjee and Duflo, 2014), (Schmalz et al., 2017), among others). Key findings highlight that financial constraints adversely affect entrepreneurship by (i) reducing the rate of entry into entrepreneurial activities, (ii) establishing a positive relationship between individual wealth and the likelihood of starting a business, (iii) limiting the growth potential of firms on the intensive margin, and (iv) generating a mismatch between entrepreneurial ability and probability of entry, among others.

The MCVL (see Appendix A.1 for the data description) provides the opportunity to find suggestive evidence on the first and second results in the context of the Spanish economy [17]. The proposed approach is to study whether the accumulated income during one's working history is correlated with entry into entrepreneurship [18]. A positive and significant coefficient would suggest that individuals need to wait to have accumulated enough wealth before pursuing their entrepreneurial endeavours.

Overall, we find that that a 10% increase in cumulative income leads to a 3.9% increase in the relative probability of becoming an entrepreneur. We are careful to control for observable characteristics both at the individual (sex, age, household composition, skill, ...) and UA (average skill, average income, experienced density, age composition, unemployment rate, ...) level. This suggestive evidence hence complements the findings in the literature in motivating the incorporation

[17] Note that since neither fiscal data on entrepreneurial episodes nor information on the entrepreneurial firm is observed in the MCVL the last result can not be tested.

[18] Note that the MCVL provides no wealth data but only labor income. Consequently, past accumulated income is treated as a proxy for an individual's own funds

of financial frictions in the model. Further details are provided in Appendix B.2.2.

3 Model

We propose a quantitative dynamic spatial equilibrium model with occupational choice that aims to explain the geographical variation in entrepreneurship. Primarily, it allows to study the interaction between financial frictions, dynamic capital accumulation and heterogeneous locations. Table (1) provides a summary of the environment of the model. The model borrows insights from the literature in dynamic models of economic geography ((Giannone et al., 2020), (Kleinman et al., 2023)) as well as from those in entrepreneurship ((Cagetti and De Nardi, 2006), (Quadrini, 2000)).

Table 1: Summary of the Model

Block	Key Points
Occupational Choice	Agents choose between being workers or entrepreneurs every period
Technology	J sectors Tradable Intermediate Sectoral Goods Y_j, produced by price-taker Entrepreneurs at prices p_j Intermediate Good Production subject to Agglomeration Forces (size and skill), Financial Frictions and DRS These are aggregated into the Final Good Y
Heterogeneity	Ex-ante Heterogeneity in Sectors and Skill Level Ex-ante Homogeneity in Entrepreneurial Ability z
Financial Markets	Financial Frictions: Entrepreneurs can borrow up to $k \leq \lambda_j a$ Single Country-Wide market for Capital at rate r
Mobility	Agents are able to move across the L locations of the Economy subject to Utility Costs $\tau_{l,l'}$
Labour Markets	Labour Markets are at the location-sector-skill level and Supply and Demand clear at wages $w_{s,j,l}$
Trade	Costless trade across locations, prices of intermediates and final good equalize
Housing	Assume simple supply equation $H_l = p_{h,l}^{\eta_l} \bar{H}_l$ (Supply) HH have CB preferences over the final good and housing (Demand) Housing sector profits get rebated to households on a location basis
Government	The Government levies taxes on labour τ_L and on profits τ_K to pay for an exogenous level of expenditure G and may employ place-based policies $\tau_{L,l}, \tau_{K,l}, T_l$ to promote entrepreneurship

Notes: The table presents the main ingredients of the model. The first column indicates the block of the model under consideration. The second column provides the details on the assumptions made for each block.

3.1 Setting

3.1.1 Heterogeneous Households

The model features infinitely-lived heterogeneous households. These households have permanent and fixed sectoral j and skill types s (either low or high skilled). They face the period by period problem of choosing their occupation. They may choose to either become workers and supply a unit of labour in the local labour markets or become entrepreneurs and produce the intermediate good in their sector j. In addition, these agents are mobile across locations in an endogenous manner, subject to utility moving costs. Lastly, every period they receive innovations to their entrepreneurial productivity z in their given sector j. The entrepreneurial ability z is assumed to follow an AR(1) process:

$$z_t = \rho_z z_{t-1} + \epsilon_t \text{ s.t. } \epsilon_t \sim \mathcal{N}(0, \sigma_y)$$

Where z denotes productivity, ρ is the auto-correlation coefficient and σ_z is the standard deviation of the error term.

3.1.2 Sectors

In order to account for the observed heterogeneous distribution of sectors across space in the data and their differential sensitivities to agglomeration forces, the model features J sectors. These sectors are heterogeneous along three key dimensions. First, each sector has an exogenous amount of workforce of each skill s category, given the assumption that agents in the economy have fixed sectoral j and skill s types. Second, each sector employs the three inputs in production (unskilled labour, skilled labour and capital) in varying intensities $\alpha_{i,j}$, where i denotes the input and j the sector. This allows for some sectors to be more capital or labour intensive than others. Lastly, sectors have varying degrees of sensitivities to agglomeration forces. [19]

[19] Thus, the model allows for some sectors to benefit more from knowledge spillovers, and others from urbanization economics (or from neither).

3.1.3 Space

The economy consists of L locations indexed by $l \in \{1, 2, ..., L\}$. Locations are heterogeneous along several dimensions. First, each location has an exogenous sector-specific productivity $A_{j,l}$. Second, each location has a housing supply elasticity η_l that determines how costly it is to build in that location. Third, there are location-skill specific amenities $a_{s,l}$. The discussion of these parameters shall be postponed to their respective sections.

3.2 Model Environment

Technology

Production in the economy has two layers. First, each sector of the economy produces an intermediate good Y_j. Second, these are aggregated in CES fashion into the final good of the economy Y. We shall discuss each layer in detail.

Intermediate Goods

The first layer of production is carried out by the entrepreneurs. An entrepreneur in location l and sector j, with idiosyncratic productivity z, skill level s and assets a produces an amount $Y_{e,j,l}$ of the intermediate good Y_j, subject to the following profit function:

$$\pi_{a,z,s,j,l} = \underbrace{p_j}_{\text{Price-takers}} \Big(\underbrace{z}_{\text{Idiosyncratic Productivity}} + \underbrace{\phi_j}_{\text{Productivity Scaler}} \Big) \underbrace{\Phi_{s,j}}_{\text{Multiplier}} \underbrace{A_{j,l}}_{\text{High-Skilled Exogenous Productivity}}$$
$$\left(1 + \underbrace{\left(\frac{L_l}{L}\right)^{\Omega_{\text{SIZE,J}}}}_{\text{Size Agg. Forces}} \underbrace{\left(\frac{H_l}{U_l}\right)^{\Omega_{skill,j}}}_{\text{Skill Agg. Forces}} \right) \left(\alpha_{H,j} h^{\frac{(\xi-1)}{\xi}} + \alpha_{U,j} u^{\frac{(\xi-1)}{\xi}} + \alpha_{K,j} k^{\frac{(\xi-1)}{\xi}} \right)^{\underbrace{\mu_j}_{\text{DRS,}\mu_j<1} \frac{\xi}{\xi-1}}$$
$$- \underbrace{(W_{H,j,l} h + W_{U,j,l} u_e)}_{\text{Labour Costs}} - \underbrace{(r_t + \delta) k}_{\text{Capital Costs}} - \underbrace{F_j}_{\text{Fixed Costs}}$$

S.T.

$$\underbrace{k \leq \lambda_j a}_{\text{Borrowing Constraint}}$$

A few comments are in place regarding the structure thereof. First, entrepreneurs are assumed to be price-takers at price p_j on the single variety of the intermediate good Y_j. Second, each entrepreneur e has its own idiosyncratic productivity z_e, which is assumed to follow a standard AR(1) process. Third, the productivity of the entrepreneur is further amplified by the terms ϕ_j and $\Phi_{s,j}$. These are sector and sector-skill specific parameters whose role will be discussed in length in the calibration section [20]. Lastly, there is the exogenous productivity term $A_{j,l}$ which is location-sector specific. [21]

The production function in this economy is a CES with three inputs: high-skilled labor, low-skilled labor, and capital. Each input has its corresponding weight $\alpha_{I,j}$, which are input and sector specific. The elasticity of substitution between inputs in production is ξ. Lastly, production takes place at decreasing returns to scale (DRS). Thus, we would obtain a well-defined distribution of entrepreneurs over the state space even in the absence of financial frictions, as this assumption guarantees that the most productive entrepreneurs can not service the entire market by themselves. [22]

Sector-location specific agglomeration forces enter in production through the terms $\Omega_{sizes,j}$ and $\Omega_{skill,j}$. Thus, larger locations $\uparrow \frac{L_l}{L}$ and/or more skilled $\uparrow \frac{H_l}{U_l}$ locations are more productive. Importantly, these agglomeration forces enter in a neutral manner such that the productivity gains affect all inputs equally [23]. The term on the agglomeration forces includes a $1+$ to capture the notion that agglomeration forces increase the productivity of a given location above the base level.

Fixed costs F_j play the role of regulating the overall share of entrepreneurs in the economy, are paid in terms of the final good, and need to be paid every period in order to produce.

Lastly, entrepreneurs are subject to financial frictions. We assume limited enforceability of contracts in the in the capital rental market. Therefore, each entrepreneur will only be able to rent

[20] Intuitively, the sector-specific parameter ϕ_j regulates the share of fixed costs over total profits, and $\Phi_{s,j}$ determines the fraction of high-skilled that become entrepreneurs in each sector.

[21] Intuitively, the exogenous UA-sector productivities allow us to target the spatial distribution of production. The intuition for this is that the GDP per capita differentials across space might differ by more than implied by the agglomeration forces, for example.

[22] Note that this would not be the case in a frictionless world under constant returns to scale.

[23] Giannone (2017) estimate separate skill and size agglomeration forces per worker type. Rossi-Hansberg et al. (2019) find that skill agglomeration forces are more prominent than their sizes counterparts, and more so for the high-skilled. Here, inasmuch as agglomeration forces are estimated for all three sectors and the skill composition varies systematically across sectors, as well as there being a correlation between city size and skill composition, neutral and non-neutral agglomeration forces result in very similar results. The prominence of the skill agglomeration forces is also corroborated, in line with the cited authors.

capital at user cost $r_t + \delta$ up to $k_e \leq \lambda_j a_e$ [24]. This will result in constrained entrepreneurs having a high private return of capital, which will incentivize them to accumulate wealth in order to grow out of the constraint. This is the main mechanism in the model that embodies the stylized fact #3, namely that would-be entrepreneurs accumulate assets prior to entry.

Final Good

Once the intermediate goods are produced, they are combined in CES fashion into the final good of the economy Y:

$$Y = \left(\sum_{j}^{J} \gamma_j^{\frac{1}{\rho}} Y_j^{\frac{\rho-1}{\rho}} \right)^{\frac{\rho}{\rho-1}}$$

Where J is the number of sectors in the economy, ρ is the elasticity of substitution and γ_j are the weight parameters for each intermediate. The final good is taken as the numeraire and its price normalized to unity.

Trade

Trade in this economy is costless across locations and therefore both intermediate goods prices $p_{j,l}$ and final good prices P_l equalize across locations $p_{j,l} = p_j \; \forall l$ and $P_l = P \; \forall l$.

Financial Markets

Households in this economy supply their assets a to an intermediary for a return thereon of r_t. Entrepreneurs rent capital k from these intermediaries, for which they are required to pay the user cost, consisting of the return r_t given to the households and a cost of δ per unit of capital to cover for depreciation.

Several other points are in place. First, borrowing is only allowed for entrepreneurial endeavours and household asset positions are imposed to be non-negative $a \geq 0$ [25]. Second, production takes place after the idiosyncratic entrepreneurial ability realization z has been observed (the exact timing is explained below). Third, there exists a single country-wide market for capital, and therefore all the funds supplied by households are put into the same intermediary which can be

[24] This constraint is the one proposed in Moll (2014). Note that despite the reduced form structure of the constraint, it can be micro-founded from a limited enforcement angle. Assume that an entrepreneur could run away with a fraction $\frac{1}{\lambda}$ of the borrowed funds, then in equilibrium the lender would only lend up to the point of indifference, $k < \lambda a$ (Buera and Shin, 2013).

[25] This ensures that the collateral borrowing constraint of entrepreneurs is always well specified. It also generates precautionary behaviour by households in the event of receiving a bad entrepreneurial productivity draw or a bad location preference shock.

accessed by entrepreneurs in all locations, thus there is a unique interest rate in the economy [26].

Mobility

Agents are free to move across locations l in this economy subject to the presence of utility $\tau_{l,l',o}$ costs of moving from location l to l', and may vary by occupation $\{\text{WORKER}, \text{ENTREPRENEUR}\}$. Every period, agents receive an *i.i.d* vector of L Extreme Type 1 location preference shocks with scale parameter v. This will induce agents to move in probabilistic terms and endogenously based on their position on the state space. The exact timing is explained below.

Labour Markets

The economy is made up of multiple local labour markets that operate at the skill-sector-location level. Agents, both workers and entrepreneurs, are assumed to work or operate in the same location where they reside. [27] Thus, the wage $w_{s,j,l}$ equalizes the supply of labour by workers and demand by entrepreneurs for labour of characteristics s and j in location l, for a total of $S \times J \times L$ labour markets.

Housing

Following the structure in Ganong and Shoag (2017), there exists a construction sector in the economy which can build housing in location l, H_l. The supply is an upward function of the price for housing in that location $p_{h,l}$ with slope given by the elasticity of the housing supply in location η_l. Each location has a pre-existing stock of land $\bar{H}_l$ upon which the construction sector can build. Households are assumed to rent their housing units and home ownership is not allowed.

Therefore, the supply side of the housing sector has the following curve $H_l^s = p_{h,l}^{\eta_l} \bar{H}_l$. Hence, *ceteris paribus*, it will be more expensive to construct in those locations with a lower housing supply elasticity η_l and likewise in those locations with a lower initial level of housing stock $\bar{H}_l$.

Turning to the demand side, households enjoy Cobb-Douglas (CB) utility over the final good and housing, and are able to consume any amount. This implies that the demand for housing in a given location H_l^d is equal to the product of the share of people living in a given location times their housing consumption given their location on the state space.

[26] A concern regarding this assumption might be the heterogeneity in funding opportunities depending on the location, among which Venture Capital stands out. However, note that VC in Europe and Spain in particular is not as prominent as in the US. According to Bellucci et al. (2021), only about 10% of the global VC market is located in the EU, with the Spanish State accounting for only 3% of that figure. This results in the country accounting for 0.3% of global VC volume, significantly less than its 1.36% GDP contribution `https://www.imf.org/external/datamapper/profile/ESP`.

[27] We allow for no commuting in the model but note that since we are going to take the model to the urban areas of the Spanish State these are effectively the local labour markets.

Lastly, the profits made by the housing sector at each location get rebated to households in in the form of reduced income taxes $\tau_{h,l}$. This is done by local councils that have the following budget constraint: $\Pi_{h,l} = \tau_{h,l} I_l$, where $\Pi_{h,l}$ are the profits from the housing sector and L_l is the share of inhabitants in location l. Lastly, I_l is the income being subsidized.

Government

The government in this model economy levies flat tax rates on labour τ_L and profits τ_K in order to finance an exogenous amount G as a fraction of total output. Therefore, workers pay taxes $\tau_L w_{s,j,l}$ on their unitary supply of labour and entrepreneurs pay $\tau_K \pi_{a,z,s,j,l}$. Note that since labour supply is fixed and the tax on profits does not directly interact with the input choices, these taxes are not distortionary inasmuch as they do not distort labour supply or input choices in production.

The budget constraint of the government therefore reads as follows:

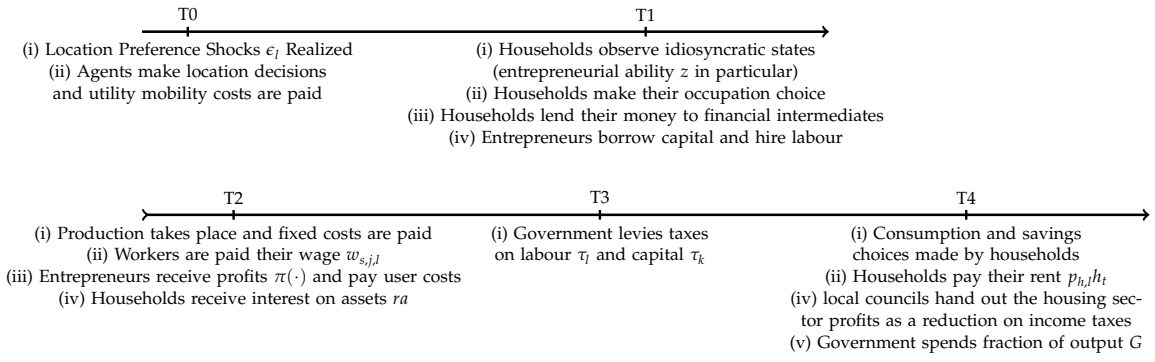
$$\underbrace{\sum_{l=1}^{L}\sum_{j=1}^{J}\sum_{s=1}^{S}\tau_L w_{l,j,s} LS_{l,j,s}}_{\text{Revenues from Labour}} + \underbrace{\int \tau_K \pi(a,z,s,j,l) occ(a,z,s,j,l) d\lambda^*(a,z,s,j,l)}_{\text{Revenues from Profits}} = GY$$

Where $LS_{s,j,l}$ denotes the share of workers of skill s and sector j in location l, $occ(a,z,s,j,l)$ is the occupational choice policy function and $\lambda^*(\cdot)$ stands for the invariant measure in the stationary equilibrium.

Timing

The sequence of events in the economy within a given time period t is outlined in figure (3).

Figure 3: Timeline of the Model



First, at time 0, the location preference vector shock ϵ_t is realized and moving decisions are

made. If households switch locations, $l \neq l'$, moving costs are paid in terms of utility, $\tau_{l,l',o}$.

Second, once households are installed in their new location, idiosyncratic entrepreneurial shocks z are realized and agents observe their idiosyncratic states $\{a, z, s, j, l\}$. Households then make their occupational choice, whether to become workers or entrepreneurs, based on these. Once occupation choices are made, households lend their financial resources a to the intermediary and entreprenerus borrow the capital they require for production up to the borrowing limit $k \leq \lambda_j a$ and hire labour.

Third, production takes place. Intermediate goods Y_j are produced across locations and are combined in CES fashion to produce the final good Y. The inputs of production are paid their renumerations. Workers receive wages $w_{s,j,l}$ and entrepreneurs pay the user cost of capital $r_t + \delta$ for renting it. After the inputs of production are paid, entrepreneurs receive the remainder of the revenues obtained net of the fixed costs F_j. Finally, households are paid their returns on assets.

Fourth, after households receive their income, the government taxes labour income at rate τ_l and profits at τ_k.

Lastly, households simultaneously decide their consumption, housing and assets to carry to the next period, mindful of the uncertainty stemming from the location preference shocks and entrepreneurial productivity process. Local councils then redistribute the obtained profits from the housing sector in the form of reduced income taxes $\tau_{h,l}$. The period ends with the government exogenously spending an amount equivalent to a fraction G of output. After the period is over, the economy transitions to period $t + 1$ which is structured in the same manner.

3.3 Households' Problem

Agents in the economy are characterized by their idiosyncratic states $\{a, z, s, j, l\}$, where a denotes assets, z is the realization of the entrepreneurial ability, s and j are the fixed skill and sectoral types and l is the current location of the individual. After locations choices are made, households decide whether to become workers or entrepreneurs in the given period by evaluating the value yielded by each alternative:

$$V_t(a_t, z_t, s_t, j_t, l_t) = \max\{V_t^w(a_t, z_t, s_t, j_t, l_t), V_t^e(a_t, z_t, s_t, j_t, l_t)\}$$

Where $V_t(a_t, z_t, s_t, j_t, l_t)$ denotes the beginning of period value, $V_t^w(a_t, z_t, s_t, j_t, l_t)$ is the value of becoming a worker and $V_t^e(a_t, z_t, s_t, j_t, l_t)$ is the value of becoming an entrepreneur. Note that households make the occupation choice taking the prices $w_{t,s,j,l}$, r_t, $p_{t,h,l}$, $p_{t,j}$, taxes $\tau_{t,L}$, $\tau_{t,K}$, fixed costs $F_{t,j}$ and housing sector income reductions $\tau_{t,h,l}$ as given.

Worker's Problem

A worker in period t with assets a_t, entrepreneurial ability z_t, skill s_t, sector j_t and in location l_t faces the following problem:

$$V_t^w(a_t, z_t, s, j, l_t) = \max_{a_{t+1}, h_{r,t}, c_t} \frac{\left(c_t^{\alpha_c} h_{r,t}^{1-\alpha_c}\right)^{1-\sigma}}{1-\sigma} + a_{s,l_t} + \beta \, \mathbb{E}_t \left(\max_{l_{t+1}} V_{t+1}(a_{t+1}, z_{t+1}, s, j, l_{t+1}) - \tau_{l,l_{t+1},o} + v\epsilon_{t+1}^{l_{t+1}} \right)$$

subject to the budget constraint:

$$c_t = (1 + r_t)a_t + (1 - \tau_{t,L} + \tau_{t,h,l})w_{t,s,j,l} - p_{t,h,l}h_{r,t} - a_{t+1},$$

and the borrowing constraint:

$$a_{t+1} \geq 0.$$

Where the expectation operator $\mathbb{E}$ is taken with respect to the idiosyncratic entrepreneurial and location preference shocks.

The worker's current utility flow is given by the consumption of the bundle comprised of the final good and housing plus the the skill-dependent local amenities $a_{s,l}$, that enter additively and therefore do not interact with the consumption level. The worker's budget constraint states that consumption is equal to the assets of the worker plus the returns on it $(1 + r_t)a_t$, plus the wage received $w_{t,s,j,l}$ net of taxes $(1 - \tau_{t,L})$, plus the housing sector income tax reductions $\tau_{t,h,l}$ minus the paid housing rental costs $p_{t,h,l}h_{r,t}$ and the assets that are chosen to be carried to the next period a_{t+1}.

The remainder of the value is given by the expected continuation value. It consists of three components. First, there is the value provided by each potential future location l_{t+1}. Second, there is the utility moving costs, which enter directly into the value function in terms of utils. Lastly, there is the term involving the location preference shocks $v\epsilon_{t+1}^{l_{t+1}}$. Given the assumptions

that these follow an *i.i.d* Extreme Error Type I process, note that the expected continuation can be re-expressed as (McFadden, 1972):

$$v \log \left(\sum_{l_{t+1}=0}^{L-1} \exp \left(\beta \mathbb{E} V_{t+1}(a_{t+1}, z_{t+1}, s_{t+1}, j_{t+1}, l_{t+1}) - \beta \tau_{l_t, l_{t+1}, o} \right)^{\frac{1}{v}} \right) \tag{1}$$

This expression greatly helps in the computation in the model by replacing the previous maximization over locations given the expectation over the location preference and entrepreneurial shocks by a sum of the expected continuation values with respect to the entrepreneurial process. Here the role of the scale parameter v also becomes more apparent. A higher $\uparrow v$ implies that fundamental differences across locations $V(a_{t+1}, z_{t+1}, s_{t+1}, j_{t+1}, l_{t+1}) - V(a_{t+1}, z_{t+1}, s_{t+1}, j_{t+1}, q_{t+1})$ for different future locations $l_{t+1} \neq q_{t+1}$ will be muted, while $v \to 0$ will result in moving decisions being based solely on the fundamental value provided by each location. [28]

In addition, the assumption on the distribution of the location preference shocks also allows to express the (occupation-dependent) migration matrices $\mu^{l_t, l_{t+1}, o}(a_t, z_t, s_t, j_t, l_t)$ in closed-form:

$$\mu^{l_t, l_{t+1}, o}(a_t, z_t, s, j, l_t) = \frac{\log \left(\sum_{l_{t+1}=0}^{L-1} \exp \left(\beta \mathbb{E} V_{t+1}(a_{t+1}, z_{t+1}, s_{t+1}, j_{t+1}, l_{t+1}) - \beta \tau_{l_t, l_{t+1}, o} \right)^{\frac{1}{v}} \right)}{\sum_{l_{t+1}=0}^{L-1} \log \left(\sum_{l_{t+1}=0}^{L-1} \exp \left(\beta \mathbb{E} V_{t+1}(a_{t+1}, z_{t+1}, s_{t+1}, j_{t+1}, l_{t+1}) - \beta \tau_{l_t, l_{t+1}, o} \right)^{\frac{1}{v}} \right)}$$

Where $\mu^{l_t, l_{t+1}, o}(a_t, z_t, s_t, j_t, l_t)$ denotes the share of people in occupation o in location l with states $\{a_t, z_t, s_t, j_t\}$ in period t moving to location l_{t+1} in $t+1$.

Therefore, the expression for the expected continuation value in the model can be thought of as a weighted average over locations that is governed by the fundamental value provided by each location as well as the strength of the location preference shocks. The resulting probabilities will embody the spatial equilibrium notion that agents move until utilities are equalized across space.

Entrepreneurs' Problem

An entrepreneur in period t, with assets a_t, entrepreneurial ability z_t, skill s_t, sector j_t and in location l_t faces the following problem:

[28] To put it differently, we can think of v as regulating the strength of the location preference shocks. The higher is v, the more migration decisions are guided by the location preference shocks. The lower is v, the more migration decisions are guided by the fundamental value provided by each location.

$$V_t^e(a_t, z_t, s, j, l_t) = \max_{a_{t+1}, c_t, u_t, h_{r,t}, k_t} \frac{\left(c_t^{\alpha_c} h_{r,t}^{(1-\alpha_c)}\right)^{1-\sigma}}{1-\sigma} + a_{s,l_t} + \beta \mathbb{E}\left(\max_{\{l_{t+1}\}} V_{t+1}(a_{t+1}, z_{t+1}, s, j, l_{t+1}) - \tau_{l_t, l_{t+1}, o} + v\epsilon_{t+1}^{l_{t+1}}\right)$$

Subject to the budget constraint:

$$c_t = (1 + r_t)a_t + (1 - \tau_{t,K} + \tau_{t,h,l})\pi_t(z_t, s, j, a_t, l_t) - h_{r,t}p_{t,h,l} - a_{t+1}$$

Where profits are given by (we omit time subscripts here in the interest of clarity):

$$\pi(\cdot) = p_j(z_t + \phi_j)\Phi_{s,j}A_{j,l}\left(1 + \left(\frac{L_l}{L}\right)^{\Omega_{\text{SIZE}}}\left(\frac{H_l}{U_l}\right)^{\Omega_{skill}}\right)\left(\alpha_{H,j}h_{f,j}^{\frac{(\xi-1)}{\xi}} + \alpha_{U,j}u_{f,j}^{\frac{(\xi-1)}{\xi}} + \alpha_{K,j}k_f^{\frac{(\xi-1)}{\xi}}\right)^{\mu\frac{\xi}{\xi-1}}$$

$$- W_{H,j,l}h_{f,j} - W_{U,j,l}u_{f,j} - (r_t + \delta)k_f - F_j$$

And subject to the collateral and borrowing constraints:

$$k_t \leq \lambda a_t \qquad a_{t+1} \geq 0$$

The problem of the entrepreneur shares many features with that of the worker. However, note that there are three major differences. First, the revenues of the entrepreneur emerge from profits $\pi_t(\cdot)$ net of the capital tax $(1 - \tau_{t,K})$ rather than wages. Second, the entrepreneur needs to choose optimal inputs of production u_t, h_t, k_t. Third, the entrepreneur faces a borrowing constraint for financing the firm. Note that even if the production input choice is static every period for a given level of assets a_t, the associated Lagrange multiplier results in the intertemporal asset accumulation motive being essential in driving the savings policy function.

3.4 Equilibrium

The Stationary Recursive Competitive Equilibrium (SCRE) in this economy consists of prices $\{w_{s,j,l}\}_{s=1,j=1,l=1}^{S,J,L}$, $\{p_{h,l}\}_{l=1}^{L}$, $\{p_j\}_{j=1}^{J}$, r, allocations of people across space $\{L_l\}_{l=1}^{L}$, labour $\{LS_{s,j,l}\}_{s=1,j=1,l=1}^{S,J,L}$, entrepreneurs $\{E_{j,l}\}_{j=1,l=1}^{J,L}$, housing supply stocks $\{H_l\}_{l=1}^{L}$ and income tax reductions $\{\tau_{h,l}\}_{l=1}^{L}$, supply of intermediates $\{Y_j\}_{j=1}^{J}$, exogenous government expenditure G and shares of entrepreneurs

by sector $\{E_j\}_{j=1}^{J}$ that are consistent with households' optimization and that clear the respective markets. The definition of the equilibrium is provided in the appendix (C).

4 Solution Method

Solving dynamic spatial economic models involves two key challenges. The first one is the large state space, which arises from the richness in the heterogeneity that this class of models allows for. In this case, the state space is five dimensional and the number of elements given by $A \times Z \times S \times J \times L$. The second is the structure imposed by the spatial structure of the model. Contrary to a standard DSGE model with a single location, where agents would face common prices and the expected continuation value would involve a single term (the expected value in that location), this class of models requires agents to keep track of prices within and across all locations, as well as evaluating all value functions associated to different locations. [29]

The literature has proposed a number of methods to alleviate these difficulties. Caliendo et al. (2019) develop a hat-algebra methodology to solve for counterfactuals in a dynamic spatial economy with log utility. Kleinman et al. (2023) build upon this methodology to incorporate capital accumulation, and are able to solve for the endogenous transition path by relying on spectral analysis. A key assumption of these methodologies however is the ability to linearize the model. In the present setting, given the endogenous capital accumulation, borrowing constraints, location choice and occupational choice, the model is non-convex and the policy functions are non-linear. Therefore, the approach proposed in Giannone et al. (2020) is followed, which relies on employing global solution methods.

Following the authors, insights from the literature are borrowed in order to solve the model. First, the assumption that location preference shocks follow a Extreme Type I distribution greatly simplify the computation of the model, since it allows to rewrite the discrete maximization over lo-

[29] Note that the spatial dimension imposes more structure. As an example, let's compare the dynamic problem of a firm that needs to choose optimal tangible k and intangible I capital in a single location l versus a firm that needs to solve for optimal tangible capital in an spatial setting. In the former case, there is a single continuation value (that of the single location evaluated at some values for tangible and intangible capital), and the prices are those at that single location (say price of labour and the user cost of capital). In clear contrast, in the latter case the firm needs to evaluate each potential value of future tangible capital at each potential future location. Moreover, since each location has its own prices (at least on some inputs), we need to solve for many additional market clearing conditions in GE. Lastly, we need to compute the migration policy function $\mu(\cdot)$, which increases the state space by an additional dimension, as we need to compute the probability that a firm in a given point on the state space today (k, l) ends up in each potential future location l' tomorrow, $\mu(k, l, l')$.

cations in the expected continuation value $\mathbb{E}\left(\max_{\{k\}_{k=1}^{L}} V_{t+1}(a_{t+1}, z_{t+1}, s, j, k) - \tau_{l,k,o} + v\epsilon_{t+1}^{k}\right)$ as a sum of the value function in the different locations $v \log\left(\sum_{k=0}^{L-1} \exp\left(\beta \mathbb{E}V(a', z', s', j', k') - \beta\tau_{l,k,o}\right)^{\frac{1}{v}}\right)$, thus effectively integrating out the location preference shock.

Second, there exists a literature in computational economics that aims to exploit the parallelizable nature of Graphics Processing Units (GPUs) to solve economics models such as Fernández-Villaverde and Valencia (2018). Here, compared to this literature, we first provide a full-fledged native CUDA/C++ implementation, with all required components, that allows to solve the model, from calibration to policy exercises, within this heterogeneous CPU-GPU model. Second, compared to these authors, we find higher relative gains from employing GPUs. In particular, we find a speed-up of over 20,000 compared to MATLAB (against x27) and of over 2,000 compared to C++ (compared to x2-3). In the computational appendix D we show how speed-ups close to ours can be achieved in the code by Fernández-Villaverde and Valencia (2018) by employing several optimizations. Third, we discuss the benefits for macroeconomists of the latest added features in CUDA, such as the Cooperative Groups API for within-kernel thread synchronization. Lastly, we discuss two different algorithms to understand when parallelization works more satisfactorily and we break down the attained speed-ups by idiosyncrasy of the CUDA programming model. In order to achieve these speed-ups, special care is placed in following the guidelines and recommendations by Nvidia engineers ((Guide, 2020), (Guide, 2020)). A thorough discussion of these optimization techniques is provided in the appendix D.

4.1 Solving for the Steady State

Solving for the stationary equilibrium involves several steps. First, an initial guess on the wage triplet $\{w_{s,j,l}\}_{s=1,j=1,l=1}^{S,J,L}$, population distribution across locations $\{L_l\}_{l=1}^{L}$, share of high skilled across locations $\{\frac{H_l}{L_l}\}_{l=1}^{L}$, rent prices $\{p_{h,l}\}_{l=1}^{L}$, housing sector profits income tax reductions $\{\tau_{h,l}\}_{l=1}^{L}$ interest rate r and intermediate goods prices $\{p_j\}_{j=1}^{J}$ is required. Second, the agglomeration forces $\left(\frac{L_l}{L}\right)^{\Omega_{\text{SIZE,J}}}$ and $\left(\frac{H_l}{U_l}\right)^{\Omega_{\text{SKILL,J}}}$ are computed. Third, given the prices and agglomeration forces, the problem of the entrepreneur can be solved independently of the Bellman equation for the optimal input choices $k^*(\cdot), u^*(\cdot), h^*(\cdot)$, since the problem is static. Fourth, given the prices and optimal input choices, the dynamic problem is solved via Value Function Iteration and the associated poli-

cies $occ(\cdot), a'_w(\cdot), a'_e(\cdot), h_e(\cdot), h_w(\cdot)$, $c_w(\cdot), c_e(\cdot), \mu^{l,l',o}(\cdot)$ are obtained. [30] After that, the invariant distribution over the state space $\lambda^*(\cdot)$ is computed by relying on Young (2010)'s method. The next step involves computing the aggregates of interest by integrating the policy functions with respect to the obtained invariant measure. [31] With the updated guesses on the allocations in the economy A', $\{L'_l\}_{l=1}^{L}$, $\{LS'_{s,j,l}\}_{s=1,j=1,l=1}^{S,J,L}$, $\{LD'_{s,j,l}\}_{s=1,j=1,l=1}^{S,J,L}$, $\{Y'_j\}_{j=1}^{J}$ and housing sector profits $\{\Pi_{h,l}\}_{l=1}^{L}$ an updated guess for the prices r', $\{p'_j\}_{j=1}^{J}$, $\{w'_{s,j,l}\}_{s=1,j=1,l=1}^{J}$, $\{p'_{h,l}\}_{l=1}^{L}$ and income tax reductions $\{\tau_{h,l}\}_{l=1}^{L}$ or transfers $\{T_{h,l}\}_{l=1}^{L}$ is attained. If the distance between both set of sequences of prices meets the convergence criteria an equilibrium is found. Otherwise, new guesses are obtained from the realized errors in the equilibrium conditions and further iterations are performed starting from the second step until the convergence criteria are met.

4.2 Proposed Methodology and Attained Speed-up Gains

As argued above, exploiting the parallelizable nature of the algorithms employed to solve macroeconomic models by employing GPUs which are precisely designed to handle parallel tasks is the main idea behind the computational method. Architecturally, CPUs are optimized around latency, that is, performing a particular task in the least amount of time possible. In clear contrast, GPUs are optimized around throughput, which translates into the ability of performing many tasks at once. This distinction becomes clear when one analyzes the composition of each chip 4.

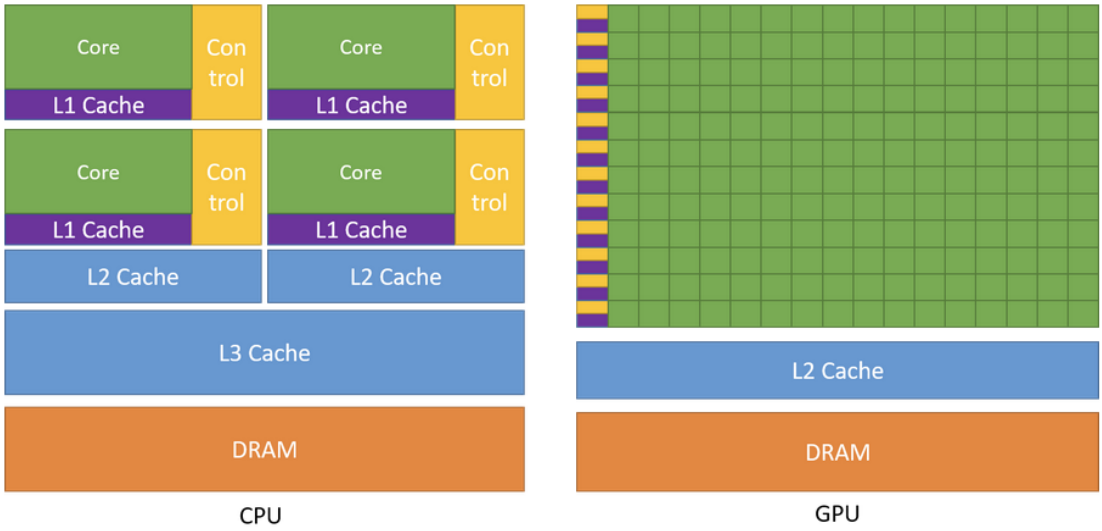
As a motivating example, suppose the economist would like to solve a model that requires summing two vectors, $\mathbf{x}$ and $\mathbf{y}$, with cardinality $\mathcal{N}$ at some step on the algorithm. Clearly, the end result would be $\mathbf{c}[i] = \mathbf{x}[i] + \mathbf{y}[i]$ where i is the index for each of the elements of the vector. How would the typical MATLAB sequential implementation handle this? [32] One of the four cores visible on the left panel would be tasked with performing $\mathbf{c}[i] = \mathbf{x}[i] + \mathbf{y}[i]$ for every $i = 0, ..., \mathcal{N}$ sequentially, performing i first, then $i + 1$ and so on. How would the GPU do it in turn? As can

[30] These refer, respectively, to the occupational choice policy function $occ(\cdot)$, asset policy for workers $a'_w(\cdot)$, asset policy for entrepreneurs $a'_e(\cdot)$, housing policy of entrepreneurs $h_e(\cdot)$, housing policy of workers $h_w(\cdot)$ consumption policy for workers $c_w(\cdot)$, consumption policy for entrepreneurs $c_e(\cdot)$ and migration probabilities $\mu^{l,l',o}(\cdot)$.

[31] This is performed by relying on custom implementations of reduction algorithms, which are the parallel counterparts to sequential cumulative sums that have the added advantage of providing higher accuracy. For more information on reduction algorithms, `https://developer.download.nvidia.com/assets/cuda/files/reduction.pdf`.

[32] For the sake of the discussion we are going to assume a fully sequential *for* loop for summing these vectors. MATLAB does allow the use of vectorized SIMD instructions and multithreading along the different cores of the CPU.

Figure 4: Architectural design of a typical CPU compared to a typical GPU. Source: CUDA documentation, Nvidia Corporation `https://docs.nvidia.com/cuda/cuda-c-programming-guide/`



Notes: The architectural design of typical CPUs and GPUs differs significantly. In CPUs, most transistors are dedicated to fast memory banks, known as L3, L2, L1, and L0 caches, as well as versatile control and logic units that optimize performance across a wide range of tasks. In contrast, GPUs allocate the majority of their transistors to data processing units, with multiple pipelines sharing L2 and L1 caches, as well as a common control and logic unit, to maximize throughput by parallel processing.

be observed in the figure, the GPU contains multiple such cores. [33] CUDA, the platform on which the code that is deployed to the Nvidia GPUs is programmed, provides a clear software-hardware mapping. This means that the code written in CUDA is inherently parallel and ensures that each element i is mapped to a distinct core c on the GPU. Therefore, rather than having a single core performing $\mathcal{N}$ elements, the GPU assigns $\mathcal{N}$ cores c to perform $\mathcal{N}$ elements i simultaneously [34]. This is precisely what is meant by GPUs being designed to maximizes throughput, they allow to execute hundreds (and thousands, even tens of thousands) of points on the state space (VFI) or observations (Panel Simulation) simultaneously.

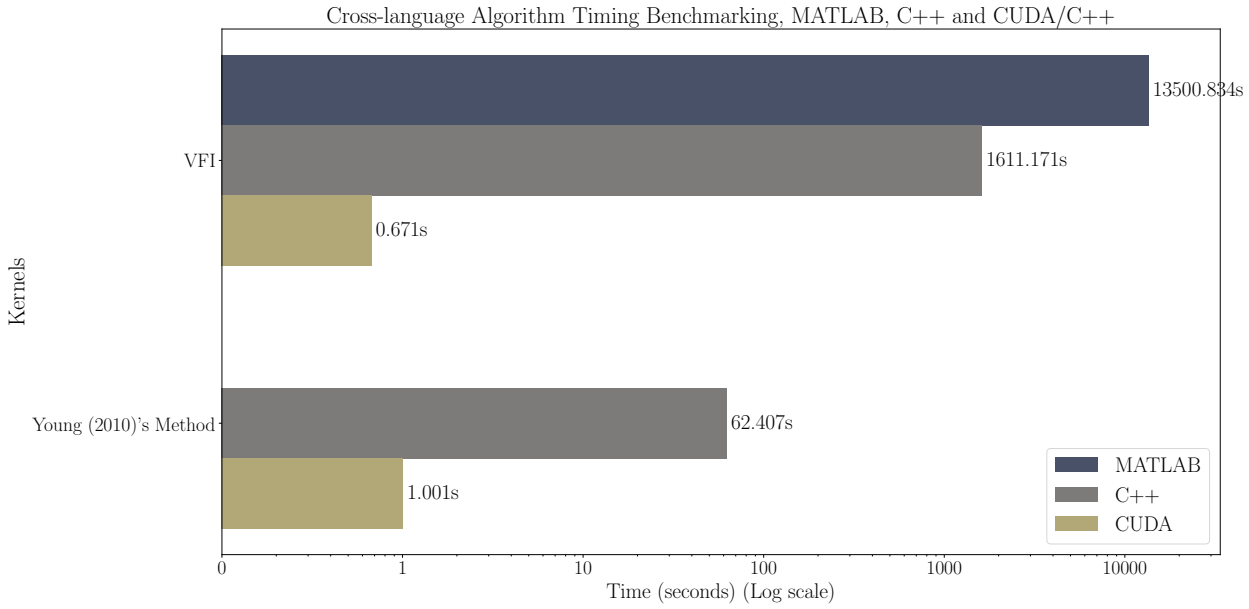
At this point, the reader might wonder whether these theoretical computational gains actually translate into the ability of solving macroeconomic models faster. Figure (5) aims to provide an answer to this question in the context of the present model.

The figure shows the execution time of the main two algorithms employed to solve this model:

[33] Pipelines of a specific data type to be more specific.

[34] Note that each GPU has different specifications on the available pipelines of execution per data type. Here we assume that $\mathcal{N}$ is below that limit. If it were to fall above, the GPU would execute sequential *layers* of threads.

Figure 5: Comparison of the Execution Time for the Main Kernels of the Model (CPU vs GPU)



Notes: Execution time of the main kernels of the model, VFI and Young (2010)'s Non-Stochastic Simulation, in MATLAB, C++ and CUDA/C++. The VFI employs linear interpolation and Howard's Improvement. Young (2010)'s method is performed with SPMV COO multiplication. Each bar represents the execution time of each algorithm in each language in seconds. The MATLAB timing in an estimate based on previous benchmarks of fully-vectorized implementations. The C++ implementation is a single-core sequential implementation. The CUDA implementation is SIMT (Single Instruction Multiple Threads). All computations are performed in fp32 and int32 data types. VFI is an example of a computation parallel, memory quasi-parallel algorithm. Young (2010)'s method is an example of a computation parallel, memory bottle-necked algorithm (given specific custom implementation). The employed CPU is a 10600K and the GPU is an RTX 3070 (5888 fp32 pipelines).

the VFI for solving for the policy functions and Young (2010)'s method for computing the aggregates by employed programming language (MATLAB, C++ and CUDA). In the case of algorithms such as the VFI that are parallelizable both at the computation and memory levels, substantial speed-ups of around 20,000 can be achieved compared to a standard fully-vectorized MATLAB implementation and of upwards of 2,000 compared to the same implementation in C++. In this case, the benefit to the macroeconomist is clear: going from MATLAB to CUDA/C++ would reduce the time required to compute a VFI for a single GE price evaluation round from 3.75 hours to less than one second. Are the results always so favourable to GPUs? If we look at the results for the non-stochastic simulation, we observe that now the speedup relative to C++ falls from around 2,000 to just over 60. This is mainly due to the fact that while the algorithm is parallel in the sense that each point of the state space performs its own independent operations, the memory reads

and writes performed by multiple threads that need to access the same memory address cause traffic jams in the memory pipeline. In plain words, there are many points on the state space today that transition with some probability to the same point on the state space tomorrow. Each of these points adding their respective probability while ensuring safe memory reads and writes effectively creates a sequential request queue for memory. Further details on this are provided in appendix D. A github repository with sample codes for solving the Aiyagari (1994) model using CUDA/C++ can be found at: `https://github.com/markoirisarri/AiyagariModelCUDA`.

5 Calibration and Model Evaluation

This section discusses the calibration of the model, followed by an evaluation of the performance of the model in terms of the targeted and untargeted moments.

5.1 Calibration

The model is calibrated to the 20 largest Urban Areas (UA) of the Spanish state. A mixed strategy is followed whereby a set of parameters are fixed to standard values in the literature, another set is estimated externally by relying on the available datasets, MCVL (household side) and SABI (firm side), and the last set is calibrated internally within the model. Table (E.6) provides a summary of the main calibration elements. The main parameters of interest are discussed in the following paragraphs. Further details are provided in Appendix B.

Sectors J in the Model Economy

As discussed previously in the empirical section, in order to alleviate the dimensionality of the problem the number of sectors is compressed by merging them based on the skill distribution of both the workforce and entrepreneurs. Following this approach, the CNAE 2009 1st digit sectors are assigned a category, either Low, Medium or High. Table $(E.7)$ presents the obtained classification.

Skill S Categories in the Model Economy

To discipline the skill dimension in the model, the approach outlined in Roca and Puga (2017) is followed. Individuals are assigned a skill category, either Low or High, based on the occupational category data in the MCVL during the period 2013-2018. Thus, a "Technical Manager" is classified

as a high skilled individual, whilst a "Labourer" is classified as a low skilled individual. Table $(E.8)$ provides a detailed classification, along with the distributional composition.

Exogenous Distribution over Skills and Sectors $\{\Lambda_{s,j}\}_{s=1,j=1}^{S,J}$

Given the assumption that households are fixed in their skill s and sector j dimension, a set of initial conditions is required for the share of households in each (s, j) pair. This moments are calibrated externally and are taken directly from the MCVL over the period of interest, 2013-2018. The obtained estimates are reported in table $E.11$ in Appendix E.2.1.

Location-Skill specific Amenities $\{a_{s,l}\}_{s=1,l=1}^{S,L}$

Location and skill specific amenities are calibrated internally in order to match the share of people of skill s in location l across all (s, l) pairs. The targeted distribution is taken from the 2013-2018 average of the MCVL dataset, the period of interest. Total population in the economy is normalized to unity.

Targeting each skill group separately is pivotal since given the nature of the location preference shocks not doing so would result in smaller cities having a higher share of high skilled, which goes against the relationship observed in the data. In addition, targeting the average skill level by location is crucial since it governs the strength of the skill agglomeration forces $\Omega_{skill,j}$.

The obtained amenities $\{a_{s,l}\}_{s=1,l=1}^{S,L}$ are reported in Appendix E.2.2, figure (E.13).

Housing Supply Elasticities $\{\eta_l\}_{l=1}^{L}$

Housing Supply Elasticities, which govern the degree to which the housing stock in one location responds to changes in prices, are estimated in this paper for the Spanish State's UAs by following Saiz (2010)'s methodology. The key insight from the paper is that exogenous geographical variation can be exploited to obtain estimates of these elasticities. Intuitively, in cities that are located close to the sea or are sorrounded by rugged terrain (Barcelona, San Francisco) it would be expected that the housing supply stock would be less elastic to price changes relative to their geographically less constrained counterparts (Madrid, Houston). An in-depth discussion of the methodology applied to the case of the UAs of the Spanish state and the obtained estimates is provided in Appendix E.3.1 and figure E.18.

CES weights of Inputs in Production by Sector $\{\alpha_{h,j}, \alpha_{u,j}, \alpha_{k,j}\}_{j=1}^{J}$

The model features a CES production function with three inputs of production (high skilled labour, low skilled labour and capital), and the weights $\alpha_{i,j}$ thereof vary by sector of the economy.

In order to obtain estimates for these parameters, the CES FOCs are employed in conjuction with SABI data on wagebill, employment, capital and capital costs. This procedure allows us to estimate these weights as the relative importance that these inputs have on output as a function of the observed input prices and quantities, subject to the substitutability. The externally obtained estimates as well as a more in-depth discussion of the approach are available in Appendix E.4.6.

Size and Skill Agglomeration Forces by Sector $\{\Omega_{sizes,j}\}_{j=1}^{J}$ and $\{\Omega_{skill,j}\}_{j=1}^{J}$

An important force in the model are the skill and size agglomeration forces, which determine the extent to which an entrepreneur is effectively more productive in more skilled and denser UAs. To estimate them, a procedure based on Giannone (2017) is followed. The methodology relies on employing, as in the case of the weights of the inputs in production, the FOCs from the model directly. Conceptually, we estimate the effect of agglomeration forces on wages after accounting for observed labour supply and production levels. To account for heterogeneity on the size and skill composition of UAs, we employ the housing supply elasticity and a Bartik instruments exploiting differential trends in migrant composition as instruments. The externally obtained estimates along with a more in-depth discussion can be found in Appendix E.4.7.

Exogenous UA-Sector specific Productivities $\{A_{j,l}\}_{j=1,l=1}^{J,L}$

A key set of parameters in the model are the exogenous productivities by each UA-Sector $A_{j,l}$, which determine how productive a location is before taking into account the agglomeration forces. We estimate them internally by targeting the spatial distribution of production across UA-Sectors. This is required as the per-capita differences in production might vary by more than implied by the estimated agglomeration forces. The estimated productivities and as well as further details are provided in Appendix E.4.5.

Fixed Costs by Sector $\{F_j\}_{j=1}^{J}$

Since the model features occupational choice, the measures of entrepreneurship are fully endogenous and therefore an element is required in the model which regulates the share of entrepreneurs. This role is primarily played by the fixed costs F_j. A higher F_j implies that a lower proportion of potential entrepreneurs find it optimal to become entrepreneurs as opposed to becoming a worker.

The fixed costs are calibrated to target the share of entrepreneurs by sector in the MCVL, which are 1.1%, 1.3% and 0.7% for the low, medium and high sectors, respectively. Appendix E.4.1

provides the specific obtained values.

Maximum allowed Borrowing by Sector $\{\lambda_j\}_{j=1}^{J}$

The leverage ratio parameters by sector λ_j are calibrated to match the revenue-weighted average debt-to-total assets ratio. The targeted values are 56.8%, 55.2% and 51.2% for the Low, Medium and High sectors, respectively. These values are the 2013-2018 averages by sector of Small firms in the SABI dataset. By targeting the debt-to-assets ratio at the sector level, we seek to capture that some technologies might be more subject to financial frictions.

Intuitively, $\uparrow \lambda_j$ relaxes the borrowing constraint of the entrepreneurs $k \leq \lambda_j a$. This implies that the share of borrowed capital over total capital in the model $\frac{\max\{k-a,0\}}{k}$ increases the higher the leverage ratio is. See Appendix E.4.2 for further details.

Entrepreneurial Process parameters σ_z **and** ρ_z

The parameters governing the entrepreneurial productivity AR(1) are calibrated internally in order to match moments of interest.

Regarding the autocorrelation parameter ρ_z, it is calibrated to match the entry rate of entrepreneurs in the MCVL dataset, which is on average 6.6% of the existing stock of entrepreneurs. Since the model is stationary, the entry rate equals the entry rate so that the invariant measures can be reproduced. Intuitively, $\uparrow \rho_z$ implies that the difference between the current period's z and next period's z' is smaller. Therefore, those entrepreneurs that are active today can expect to get a similarly high entrepreneurial ability and those that are not can equally expect to get a relatively low future entrepreneurial ability [35].

As for the standard deviation of the process, σ_z, it is calibrated to generate an average MPK Premium [36] in the economy of 1.5%. This in line with estimates from the data such as Gilchrist et al. (2013), who report a mean credit spread of 1.7%, and generates about 42% of constrained entrepreneurs in the model. Intuitively, for a given ρ_z and set of λ_j, a $\uparrow \sigma_z$ implies that in the cross-section entrepreneurs demand more capital and are more constrained, which pushes $\uparrow MPK$.

Government flat Tax Rates τ_L **and** τ_K

In the model economy the government levies flat taxes on labour τ_L and capital τ_K in order

[35] Note that the assumed AR(1) process is mean-reverting. Therefore, individuals with a low realization in the current period will enjoy the highest realization eventually.

[36] Recall that we define this as the difference between the private returns of capital to entrepreneurs (MPK) and the user cost of capital ($r + \delta$).

to finance an exogenous amount of government expenditure as a fraction of GDP G. The tax on labour τ_L is calibrated externally to match the marginal tax rate at the average salary in the Spanish state, which is 30%. The tax on capital τ_K is again externally calibrated, this time to match the effective corporate tax in the Spanish state, which is reported to be 15%. This two sources of income generate a revenue equivalent to around 10% of GDP in the model economy, which is in line with the evidence (López-Rodríguez and García Ciria, 2018).

Mobility Costs $\{\tau_{l,k,o}\}_{o=1}^{2}$

A critical parameter that governs the extent to which agents move across space in the model are the utility moving costs $\tau_{l,k,o}$. To emphasize the fact that the model is able to generate heterogeneous moving patterns based on the position on the state space, these are calibrated symmetrically $\tau_{l,k,o} = \tau_{k,l,o}\ \forall\ k,l\ s.t.\ k \neq l$ for each o. It is assumed that remaining in the same location $l' = l$ entails no utility costs. Occupation specific economy-wide (outgoing) moving rates of 0.4% (workers) and 0.29% (entrepreneurs) are targeted, which is obtained from the MCVL dataset over the periods (2013-2018). The obtained values and migration matrices are reported in Appendix E.2.3.

5.2 Model Evaluation

This section presents the performance of the model along key targeted and untargeted moments of interest. On the targeted moments, we shall discuss the population and skill composition distribution across UAs. On the untargeted front, the primary focus will be on the entrepreneurial stock and entry across UAs.

Population Distribution in the Model Economy

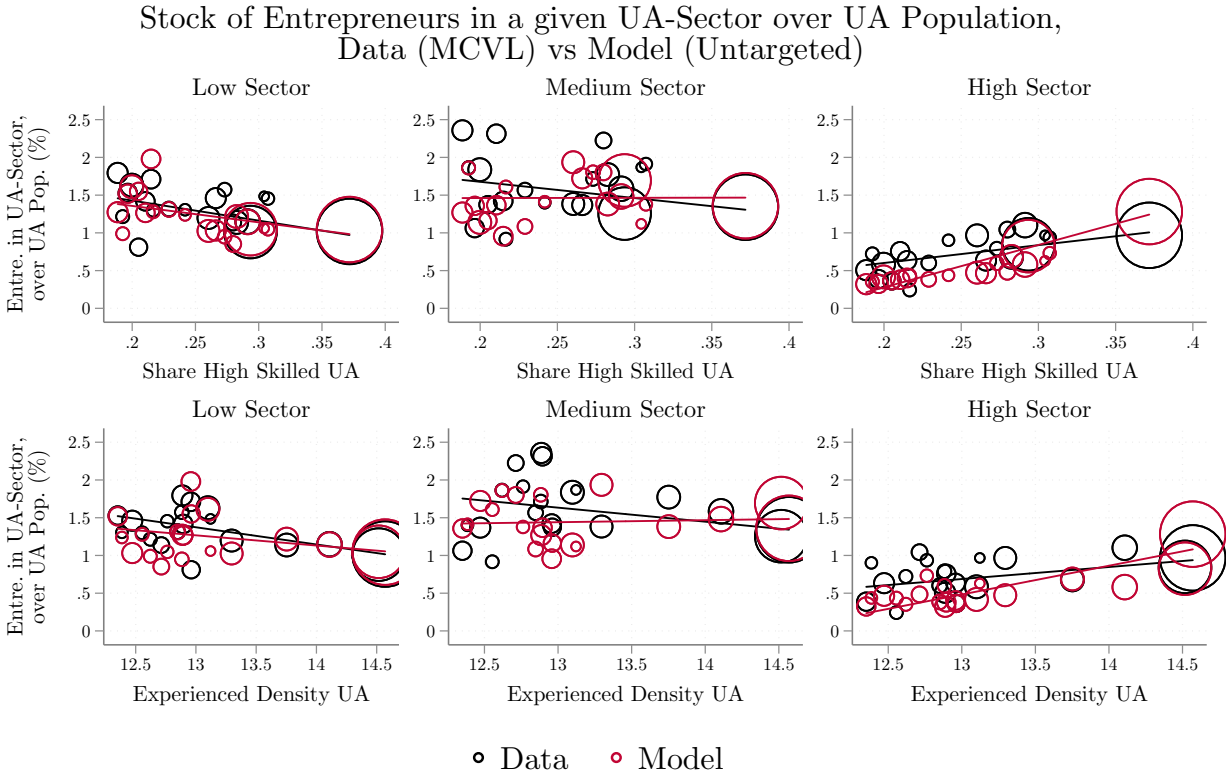
The population and skill distributions are of relevance as they regulate the agglomeration forces in the model economy, which in turn regulate the productivity of the UAs. Figure $(E.23)$ presents the distributions both in the model and the data. The model is able to match these moments to the desired precision.

Entrepreneurial Stock

The entrepreneurial stock is defined as the share of people within a given UA-Sector that choose to become entrepreneurs over the population of the UA. Figure 6 presents the results obtained for the share of entrepreneurs in the model by each UA-Sector and compares them with the

data:

Figure 6: Entrepreneurial Stock: Model and Data



Notes: Entrepreneurial stock over UA population by UA-sector. Each column corresponds to one sector of the economy. The rows show the relationship between the entrepreneurial stock and a given variable of interest (x-axis). The y-axis is in absolute terms, and denotes the percentage share of people in a given UA-sector that are entrepreneurs over the population of the UA. The black lines and dots refer to the actual data and the red are the model counterparts. Fit lines are weighted by the population of each UA. The overall share in the economy of entrepreneurs by sector is targeted, while the slopes are untargeted. Dots are scaled by the population of the UA.

The entrepreneurial rates are plotted against the two proxies for the agglomeration forces (share of high skilled and density [37]). The rationale behind this lies in the fact that these are observable in the data and are the main drivers of productivity differences across locations.

For this exercise, the share of overall entrepreneurs by sector in the economy is targeted. In contrast, the slopes, weighted by population, are untargeted. This allows us to evaluate how the model performs in explaining the variation in entrepreneurial stocks and rates. Overall, the model does a good job at matching the data. First, it captures the heterogeneity in the slopes across

[37] Recall that we employ Duranton and Puga (2020)'s measure.

sectors (negative for the low and medium sectors and positive for the high sector). Second, the results are broadly consistent across the two proxies. The main mechanisms that generate these relationships in the model shall be discussed in the next section. Results for the entry rates are in Appendix E.6.

6 Model Predictions: Heterogenous Returns to Capital across Space

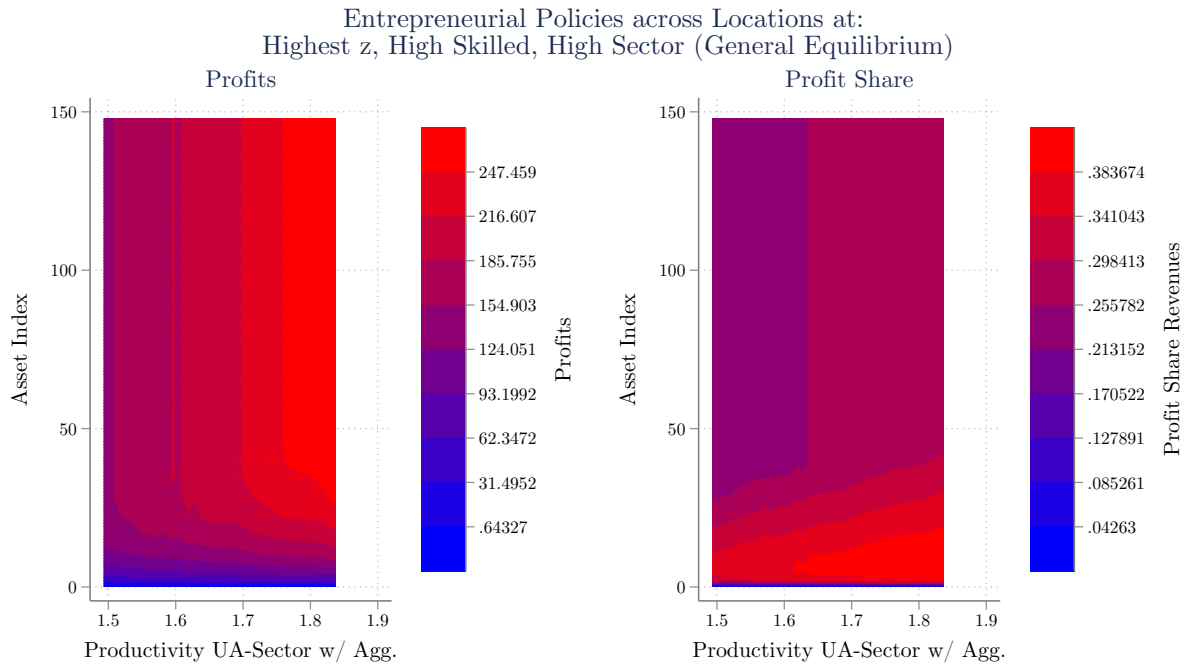
As discussed in the introduction, a salient feature of financial frictions in a spatial economy is the heterogeneity it generates on the returns to capital across locations. This section aims to provide the insight behind this result, before studying its policy implications in the next section.

Recall that financial frictions entered in the form of collateral borrowing constraints with sector-specific maximum borrowing limits over one's own assets λ_j, $k_c \leq \lambda_j a$. As long as optimal unconstrained capital k^* is such that $k^* > k_c$, the entrepreneur is constrained and this leads to a differential between the economy-wide user of cost of renting capital $(r + \delta)$ and the specific MPK of the entrepreneur, MPK PREMIUM $\equiv MPK - (r + \delta)$. What are the implications of these financial frictions for entrepreneurs across space? Figure (7) sheds light on this.

First, as shown by the leftmost panel, absolute profits are increasing in city productivity for a given level of assets. This is expected, since a higher productivity level increases the production scale and consequently absolute profits. Second, one may observe that for a given UA productivity level the profit share is first increasing and then decreasing. The intuition behind this result is that for small production scales, the fixed costs of production make up a significant share of profits. As the assets of the entrepreneur increase, the financial frictions become less severe and the profit share slowly converges towards the frictionless $1 - \mu_j$. Note that in the frictionless world there would be no interaction between asset holdings and neither profits nor profit shares, as the ability to finance any amount of capital would result in these policies being determined completely by the individual productivity.

The main point to be taken from figure (7) is therefore that (for a given level of asset holdings and individual productivity) entrepreneurs in more productive UAs will be more constrained than their counterparts living in less productive UAs. This is because for every idiosyncratic productivity z and asset level a touple (a, z), the productivity boost in more productive UAs stemming from

Figure 7: Profits and Profit Shares by UA-Sector in the Model

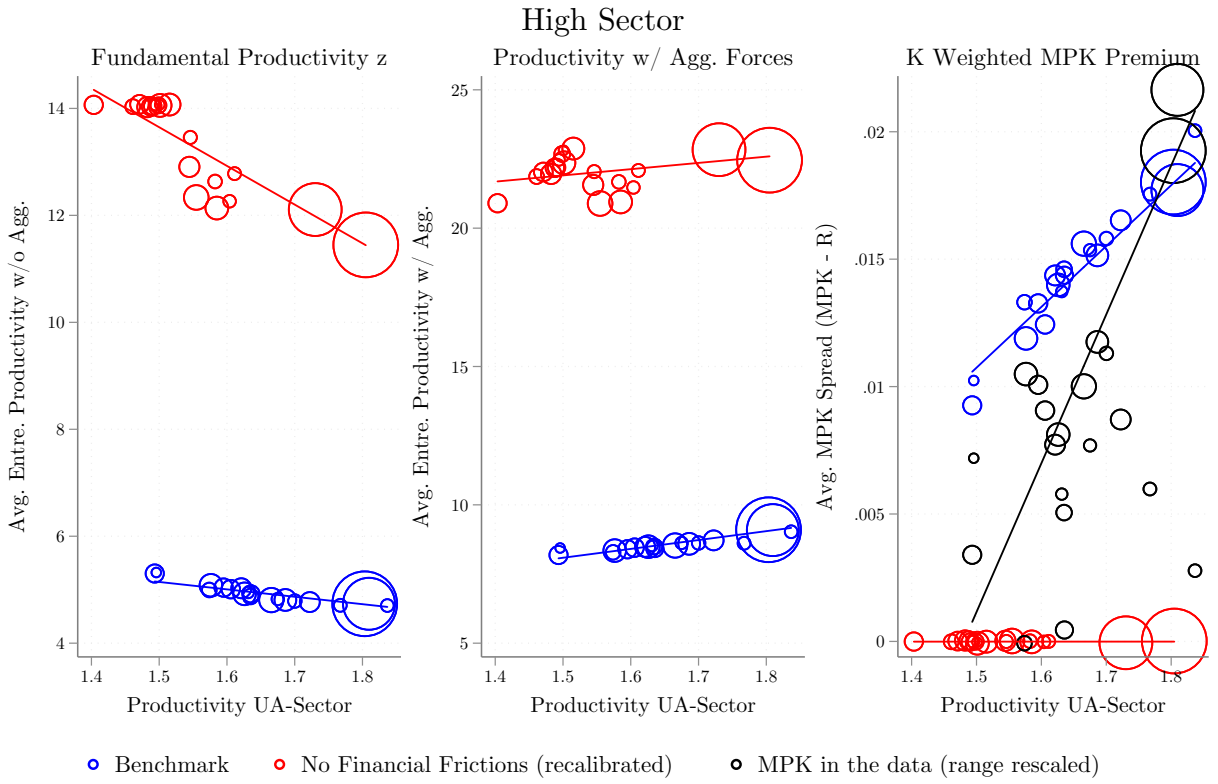


Notes: Profit and profit share policy functions in the Benchmark Economy in GE. The individual policy functions are fixed at the highest idiosyncratic productivity draw, highest skill level and high sector. The x-axis shows the variation of the policies across the location dimension and the y-axis shows the variation across the asset dimension. The colorbars display the values of the z-axis (profits and profit shares, respectively).

the agglomeration forces widens the gap between the attainable capital k_{con} and desired capital $k*$, which amplifies the MPK Premium. The key question at this point is whether this intuition from the individual policy functions carries over to the aggregates. Figure (8) aims to provide the intuition on this. The results are consistent across sectors, see Appendix F.1.

The figure shows steady-state relationships between aggregates of interest and the UA-Sector productivity for the high skilled sectors. If we focus on the no financial frictions counterfactual first, one may observe that the MPK Premiums across space are equalized and equal to zero (this is natural since every entrepreneur can now finance their optimal scale of production everywhere). In this frictionless world, if one shifts the focus to the leftmost panel, it can be observed that the average fundamental productivity z of entrepreneurs is decreasing in the UA-Sector's productivity. Albeit this might surprising, it stems from the assumption of costless trade in tradables and the presence of fixed costs in the economy. Since entrepreneurs earn the same price on their goods

Figure 8: Steady-State Results: Composition of Productivity by UAs and MPK Premiums.



Notes: Steady-state results. Each panel displays one variable of interest and its relationship to UA-sector productivity. The size of the circles represents the size of the UAs. The first panel shows the average fundamental productivity (z) of entrepreneurs (unweighted) by UA-sector. The second shows the average productivity of entrepreneurs once agglomeration forces are included (unweighted). The last shows the MPK Premium by location. All results are for the high skilled sector. The results are displayed for the benchmark economy and the recalibrated no financial frictions counterfactual.

regardless of location, it has to be that those entrepreneurs that are able to operate in less productive UAs that do not enjoy the benefits of higher agglomeration forces are indeed very productive. Especially, when considered that the per-period fixed costs that need to be paid every period in order to operate are location independent. This negative selection channel on idiosyncratic productivity is overturned when accounting for the agglomeration forces in the given UA (second panel).

Introducing financial frictions leads to a positive relationship between the productivity of the UAs and the average capital-weighted MPK Premium. The main force behind this result is the intensive margin. As we discussed above, the productivity boost stemming from the agglomeration forces widens the gap between the desired capital $k*$ and attainable capital k_{con} for every en-

trepreneur. The more productive is the UA, $Z_{UA,1} > Z_{UA_2}$, the higher the MPK Premiums will be along the entire distribution over (a, z), $MPK(a, z)_{UA,1} - (r + \delta) \geq MPK(a, z)_{UA,2} - (r + \delta) \; \forall \; (a, z)$ (with equality in the case of being unconstrained in both). Conceptually, entrepreneurs want to scale up projects in more productive UAs yet face the same borrowing constraints, leading to financial frictions becoming tighter. Then there is the compositional effect. The negative selection channel between the productivity of the UA and the idiosyncratic productivity of the entrepreneur leads to more entry. Are these entrants more constrained? We don't find evidence that the marginal entrepreneurs in terms of productivity are more constrained in more productive UAs. Even if the average entrepreneur is effectively more productive once accounting for agglomeration forces in more productive UAs, the marginal entrepreneurs are not. For the marginal entrepreneur the increased productivity from the agglomeration forces just compensates for the fall in the idiosyncratic productivity (note that since these are log-normally distributed, the fall is along the exponentiated values). For this reason, they still require a substantial amount of assets in order to find it profitable to operate a firm rather than working. This implies that only a small fraction of the entrepreneurs in this lower marginal productivity value enter, [38] so that on average the increased productivity from the agglomeration forces dominates and the average entrepreneur is more productive. Therefore, we find no significant compositional effects (in terms of productivity) and the intensive margin effect is the main force leading to this positive relationship. Further details on this are provided in appendix F.2.

An alternative approach is to view this through long-run wealth accumulation dynamics. Note that given the concavity of the MPK, a relative difference in location productivities requires a proportionally larger difference in accumulated assets in order to attain the same MPK Premium. However, first, the underlying persistence of the entrepreneurial productivity process is the same across locations. Second, given that entrepreneurs in more productive UAs are more constrained, in that they obtain a lower fraction of the unconstrained profits, they require a longer horizon of accumulating the flow of realized profits in order to reach optimal scale. Third, dynamically,

[38] Recall that our assumption that the productivity distribution follows an AR(1) process implies that along the ergodic distribution, a higher mass is concentrated towards the central values. Since entrepreneurs are a subset of the upper realizations, if the cutoff were to decrease and all agents in that realization were to enter they would represent the mode of the active entrepreneurs. Here, this is not the case because for the marginal entrepreneur the increase in productivity from the agglomeration forces just compensates for the decrease in the idiosyncratic productivity, and only a subset of the ones that have enough assets to operate at a large enough scale can enter.

the incentives to accumulate more wealth wane as the marginal utility of consumption decreases. Therefore, we find that MPK premiums are larger in more productive cities for a given UA-sector-specific asset decile. In particular, the differences across locations in MPK Premiums are mainly driven by the bottom 50% of less wealthy entrepreneurs by location, which points to entrants and smaller firms accounting for the bulk of this geographical variation. [39] This is reinforced by the fact that in the more productive UAs, given that the share of profits is inefficiently higher given the financial frictions, wages are relatively more depressed (the share of labour is lower) and entrepreneurs find it optimal to operate even at a lower wealth level.

Overall, the main point is that there are heterogeneous returns to capital across locations. Looking at the values from the third panel, MPK Premiums in Barcelona and Madrid would be roughly 1.0% higher than in Granada (the lowest productivity UA in the high sector). Importantly, this is consistent with MPK estimates from the data, please refer to Appendix A.2. Lastly, one may also wonder about the productivity distribution across locations in the spirit of Combes et al. (2012). These results are provided in Appendix F.3.

7 Policy Experiments

The previous section provided a clear message: the model features heterogeneous returns to capital across locations, and more productive locations are, on average, more constrained. The present section will discuss its policy implications for entrepreneurial place-based subsidies.

7.1 Place-based Lump-sum Transfers T_l

One common policy instrument in subsidizing entrepreneurship consists on handing out grants to entrepreneurs, which usually take the form of fixed-amount transfers. Here, the current expenditure level of the EU of 0.1% of GDP will be replicated in the form of place-based lump-sum transfers.

To be more specific, the exercise at hand is to provide T_l level of lump-sum transfers to all active entrepreneurs in location l. How are these lump-sum transfers financed? We assume that

[39] Note that this is different from our claim before that marginal entrepreneurs were not more constrained on average. Before, the statement was about the productivity of the entrepreneurs. Here, the statement is about the location specific life-cycle and wealth-accumulation processes of firms across the entire distribution of productivities.

the additional funds needed are financed through and increase δ_P in the labour tax. The rationale for choosing the labour tax is twofold. First, since the vast majority of the agents in the economy receive their income from wages, this allows for a more conservative assessment of the welfare-efficiency trade-off. Second, since an increased labour tax leads, ceteris paribus, to increased incentives to become an entrepreneur, we aim to also be conservative in capturing the full extent of misallocation along the extensive margin. Therefore, the government's budget constraint would read:

$$\underbrace{\sum_{l=1}^{L}\sum_{j=1}^{J}\sum_{s=1}^{S}(\tau_L+\delta_P)w_{l,j,s}LS_{l,j,s}}_{\text{REVENUES FROM LABOUR}}+\underbrace{\sum_{l=1}^{L}\sum_{j=1}^{J}\tau_K TBK_{l,j}}_{\text{REVENUES FROM PROFITS}}=\underbrace{G^{SS}Y}_{\text{EXOGENOUS EXPENDITURE}}+\underbrace{\sum_{l\in L_P}T_l S_{e,l}}_{\text{TRANSFERS PROGRAMME}}$$

Where $S_{e,l}$ denotes the share of entrepreneurs in location l, $LS_{l,j,s}$ is the labour supply of skill s in sector j at location l, $TBK_{l,j}$ is the tax base of the profit tax in sector j and location l, and G^{SS} is the same share of expenditure out of GDP as in the steady-state of the economy. Note that this policy experiment is a long-run exercise where two different steady-states will be compared.

The key question we will be asking ourselves is that related to the spatial dimension: does the economy respond heterogeneously to the chosen location of the transfers?

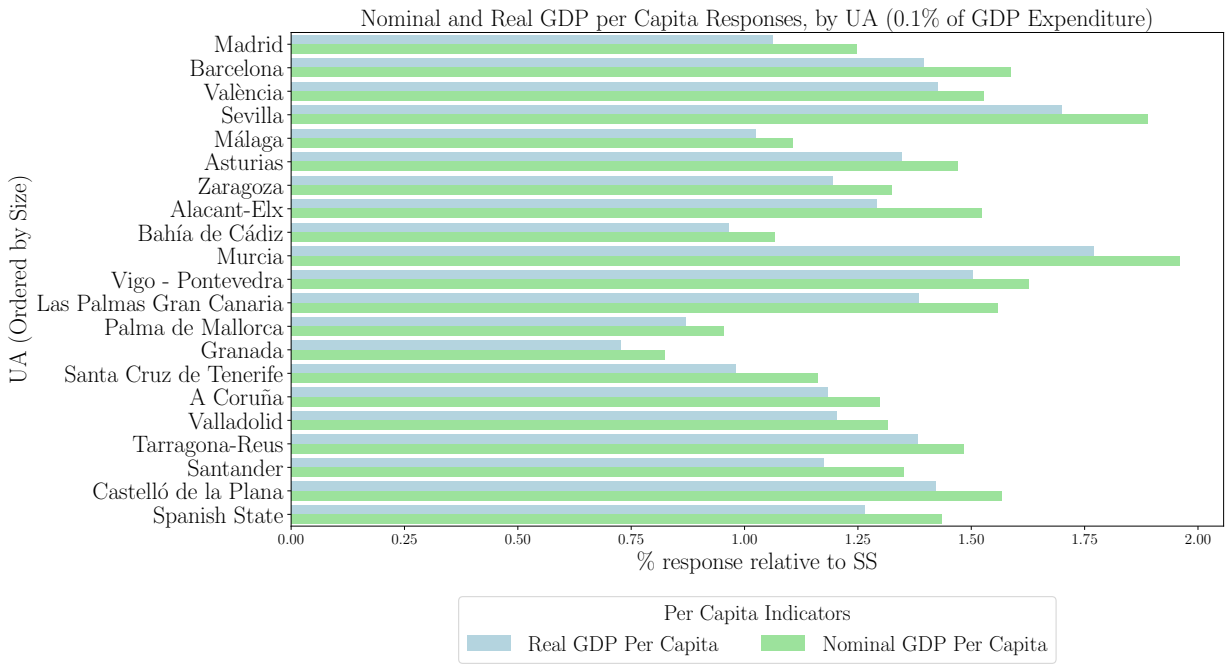
7.1.1 Benchmark: A Country-wide Subsidy to Entrepreneurship $T_l = T \; \forall \; l$

Before exploring the place-based entrepreneurial subsidies, it is worthwhile to have the untargeted version as a benchmark. This may be understood as what a policy-maker constrained to apply the same subsidy everywhere would be required to do. Figure 9 presents the response of output per capita by Urban Area to this policy.

A few comments are in place regarding the figure. First, there is a positive response of both real and nominal GDP per capita across the board. The intuition for this stems from our previous discussion: since entrepreneurs are constrained at the SS in their access to capital, handing out lump-sum transfers allows entrepreneurs to operate at a larger scale of production. This results in a country-wide (last row) real GDP per capita increase of 1.25%.

Second, overall the response of per capita real GDP is less pronounced than that of the nominal.

Figure 9: Response of per Capita Output (Nominal and Real) by Urban Area to an Untargeted Country-wide Policy (0.1% of Country-wide GDP Expenditure)



Notes: Response of Nominal and Real (deflated by the local price index $p_{h,l}^{(1-\alpha_c)}$) GDP per Capita by Urban Area to a country-wide untargeted lump-sum transfer T such that total programme expenditure is 0.1% of GDP. Urban Areas are ordered from largest in population size (Madrid) to smallest (Castelló de la Plana). The last row indicates the country-wide response. Blue bars indicate the response of the per capita real GDP by UA while green bars do likewise for the nominal GDP per capita.

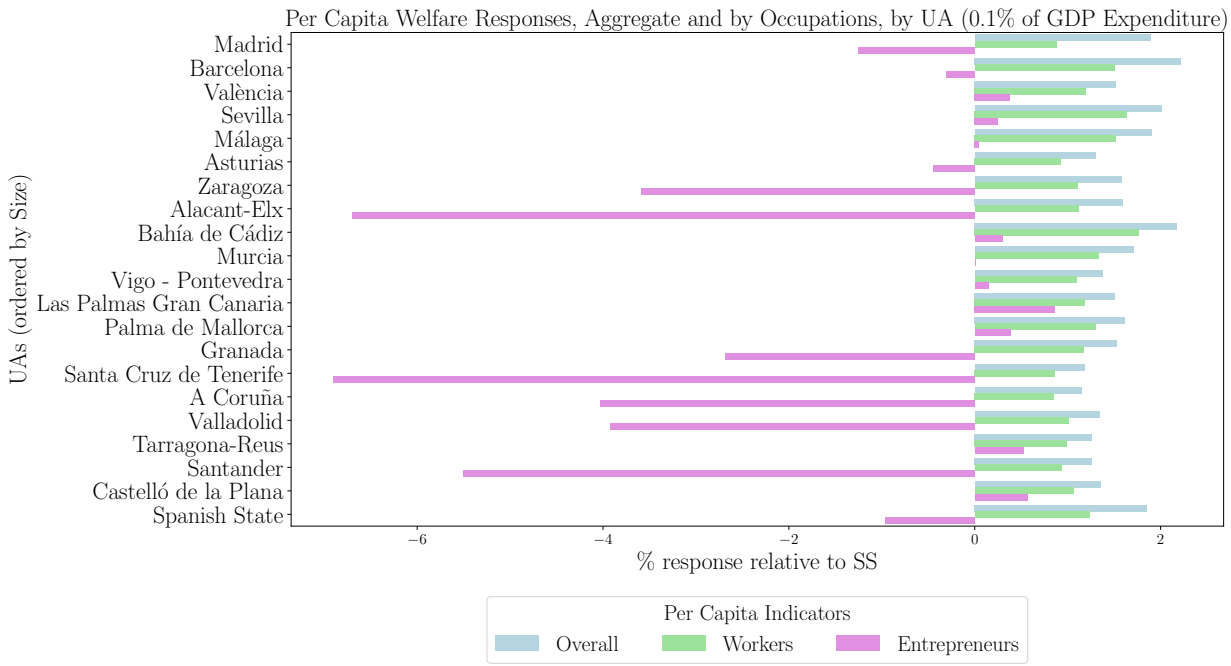
The rationale for this stems from the increases in the price of housing: a higher housing price in a location implies that less units of the final good can be consumed for the same level of expenditure.

Third, one may inquire about the implications for regional disparities that such a policy creates relative to the SS. This is motivated by the observation that the figure suggests a negative correlation between output gains and size of the UA. In fact, while overall real GDP per capita increases by 1.25%, the weighted standard deviation (as a measure of disparities across locations) *increases* by 0.92%. Those locations that presented a larger MPK Premium at the SS are the ones that benefit from this policy the most.

A natural question that arises at this point is how this increase in efficiency translates into welfare. Figure 10 presents the response of the per capita utilitarian welfare by UA to this country-wide untargeted programme.

Overall, the increases in efficiency translate into welfare gains, and all locations experience

Figure 10: Response of Per Capita Welfare, Aggregate and by Occupation, to an Untargeted Policy (0.1% of Country-wide GDP Expenditure)



Notes: Response of per capita welfare, aggregate and by occupations, a Country-wide untargeted lump-sum transfer T such that total programme expenditure is 0.1% of GDP, by Urban Area. Urban Areas are ordered from largest in population size (Madrid) to smallest (Castelló de la Plana). The last row indicates the Country-wide response. Blue bars indicate the response of the per capita welfare by UA. Green bars do likewise for the per capita welfare response of workers. Lastly, the violet bars indicate the per capita response of entrepreneurs. The welfare function under consideration is the utilitarian one, SWF = $\int V(\cdot)d\lambda(\cdot)$. The results for the response of the aggregate stock of welfare rather than the per capita one can be found at: G.30.

an aggregate per capita welfare increase. By occupations, the per capita response for workers is positive in all locations, a result primarily driven by the increase in wages. The per capita response for entrepreneurs is less obvious; should not a policy targeted at entrepreneurs increase their welfare? In fact, a welfare decomposition in the compositional and intensive margins shows that the welfare of continuing entrepreneurs does increase. However, given the movement of people across the extensive margin (those that transition from being workers to entrepreneurs, the compositional effect), now the average entrepreneur is smaller and obtains less profits than at the original steady-state. We employ an alternative definition of the response of welfare that accounts for this extensive margin. The results are shown in Appendix G.1 and figure G.30.

In terms of regional disparities, one may notice upon inspection that the previous negative

correlation between size and gains now decreases (Palma de Mallorca for example experiments one of the largest welfare gains while experiencing one of the smallest output gains). There are two main forces behind this. The first is congestion: those locations that experience a population decrease also see their house prices decrease, while those locations with a large influx of people need to pay higher housing prices due to the pressure on the supply side of the housing sector. The second is compositional: overall, lower skilled individuals, who are at a lower welfare level are the more likely to move and UAs that experiment a population decrease tend to retain higher than average skilled individuals. In fact, in the case of welfare, while the average country-wide increase is 1.82%, regional disparities in terms of per capita welfare *decrease* by 0.68%, in clear contrast to the increase in terms of economic output.

7.1.2 Place-based Transfers, T_l at some l

The benchmark with untargeted transfers conveyed a clear message: even a country-wide subsidy is able to raise aggregate output and welfare per capita. The question then becomes whether the policymaker could do better by targeting specific UAs rather than all of them, and what the implications would be in terms of regional disparities.

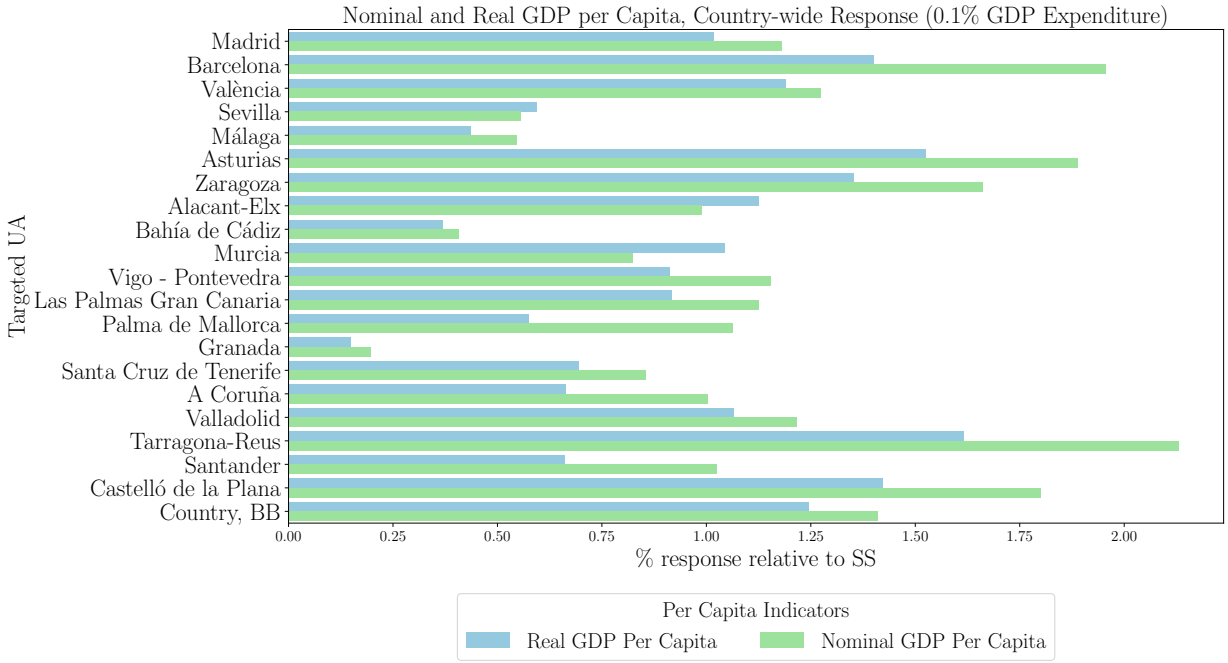
We will first discuss the **Country-wide response** of nominal and real GDP per capita **as a function of which UA is targeted** (see figure 11). [40]

The first point that deserves attention is the heterogeneity in the response of country-wide GDP per capita (nominal or real) to the location where these place-based entrepreneurial subsidies are handed out. In line with the discussion in figure 8, targeting those locations that are more capital constrained at the benchmark economy (Barcelona, Asturias, Tarragona-Reus, ...) results in the largest efficiency increases at the country level. Space does therefore matter for policy.

Secondly, the efficiency gains attainable by targeting certain locations surpass those that were registered with the untargeted policy. While an untargeted policy provided a 1.25% response of country-wide per capita GDP, there are five locations that provide a higher response in term of real GDP per capita: Castelló, Tarragona-Reus, Zaragoza, Asturias and Barcelona. The benefits from these policies have an intensive and extensive dimension. In locations that are smaller at the

[40] Note that in this section we are presenting the results differently to the previous section. Before, the entire country was targeted and the responses by UA were presented. In this section, specific UAs are targeted and the country-wide response is shown.

Figure 11: Response of Country-wide nominal and real GDP per capita to place-based entrepreneurial transfers (0.1% of Country-wide GDP Expenditure)



Notes: Response of the country-wide nominal and real (deflated by the local price index $p_{h,l}^{(1-\alpha_c)}$) GDP per capita to place-based entrepreneurial transfers T_l such that total programme expenditure is 0.1% of GDP. The y-axis denotes which UA is being targeted by the policy. The last two rows indicate the country-wide response when the entire country is targeted. The "Country BB" case is the case where the entire country is targeted and budget balance holds. Blue bars indicate the response of the country-wide per capita real GDP while green bars do likewise for the nominal GDP per capita.

original SS, a 0.1% country-wide GDP expenditure being targeted only towards these locations results in entrepreneurs receiving larger transfers (intensive margin). [41] In contrast, targeting larger UAs with the same level of expenditure results in a larger mass of the population benefiting from these policies (extensive margin). [42]

What are the implications for the spatial disparities? Taking the case extreme case of Tarragona-Reus, while country-wide real GDP per capita increases by 1.61%, this results in disparities across locations increasing by 8.38%, [43] which is over *eight times* that of the untargeted policy (0.92%). Targeting Barcelona results in a lesser 2.1% increase in disparities across locations.

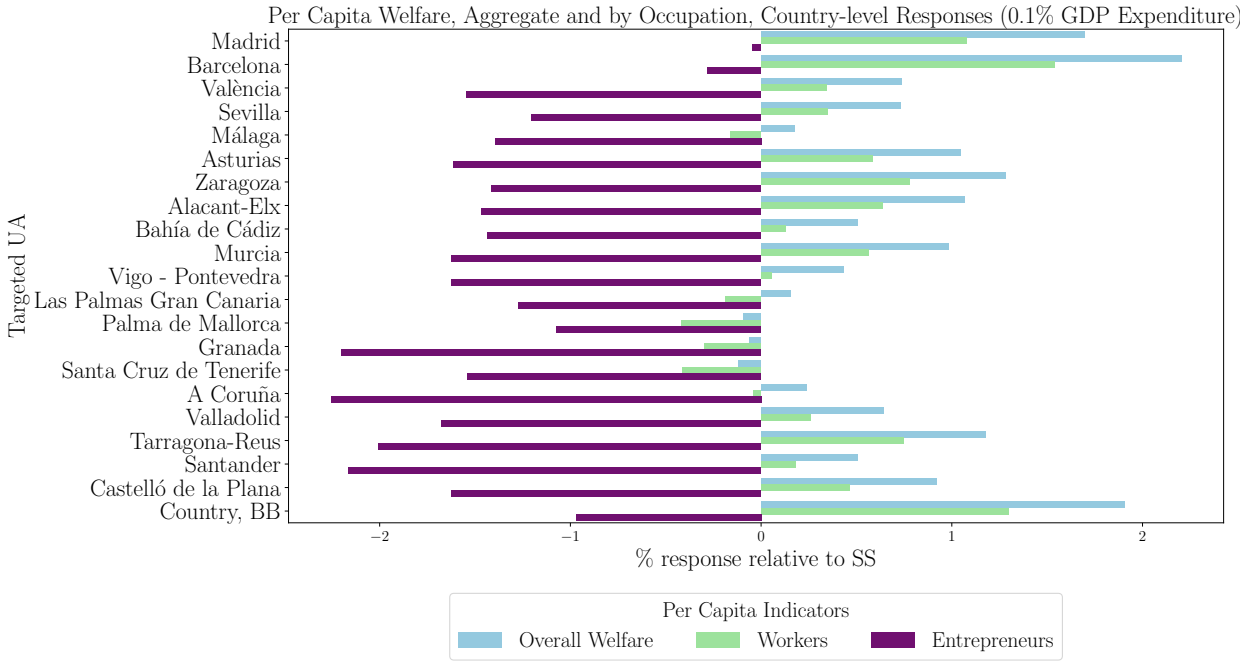
[41] The model, through the endogenous occupational choice margin, is able to capture misallocation along the extensive margin, that is, agents who become entrepreneurs not because they are productive but because the transfers are large.

[42] Note that the 0.1% GDP expenditure is that of the new policy steady-state, therefore the amounts of the transfers respond endogenously to population movements and occupational choices.

[43] As measured by the relative change in the weighted standard deviation of real GDP per capita.

The place-based policy experiments therefore suggest that targeting certain locations might be more efficiency enhancing at the country-level, albeit to the cost of aggravating regional disparities. To verify whether these results carry through to welfare, figure 12 presents the results of its response:

Figure 12: Response of Country-wide Per Capita Welfare, Aggregate and by occupation, to Place-based Entrepreneurial Transfers (0.1% of Country-wide GDP Expenditure)



Notes: Response of the country-wide per capita welfare, aggregate and by occupations, to place-based entrepreneurial transfers T_l such that total programme expenditure is 0.1% of GDP. The y-axis denotes which UA is being targeted by the policy. The last two rows indicate the country-wide response when the entire country is being targeted. The "Country BB" case is the case where the entire country is targeted and budget balance holds. Blue bars indicate the response of the country-wide per capita welfare. Green bars show the response of the per capita response of workers. Violet bars display the per capita response of entrepreneurs. The results for the aggregate response of the stock of welfare rather than the per capita one can be found at: G.31.

In the case of welfare, only targeting Barcelona results in a larger response than targeting the entire country. The main reason for this result are the higher amenities provided by Barcelona. That notwithstanding, targeting a single rather than all locations introduces additional considerations. First, the spatial coverage of the policy: while in the untargeted experiment all entrepreneurs regardless of location had access to the subsidy, targeting a single location reduces the

spatial coverage and therefore the mass of agents directly benefiting from the policy. [44] Second, the congestion forces get aggravated. While the untargeted policy saw the most productive cities increase their population and house prices, targeting a single location accentuates this congestion force by increasing the incentives of agents to move to the same location.

Interestingly, the implications for regional disparities on welfare differ from those in output. In the case of targeting Barcelona, for example, a country-wide increase in per capita welfare of 2.17% is attained. Nonetheless, this results in a *negative* -0.95% lower dispersion across regions, as opposed to the -0.68% of the untargeted policy.

7.1.3 Targeting a Set of UAs, T_l for $l \in L_p$

The previous two sections pointed towards a middle ground in the targeting criteria: while an untargeted policy increases both welfare and efficiency at low regional disparity cost, some locations have higher marginal returns. At the same time, increased regional disparities make targeting a single location less desirable. Therefore, a natural exercise arises: could targeting a set of locations provide higher efficiency and welfare gains than the untargeted policy at comparable regional disparities cost?

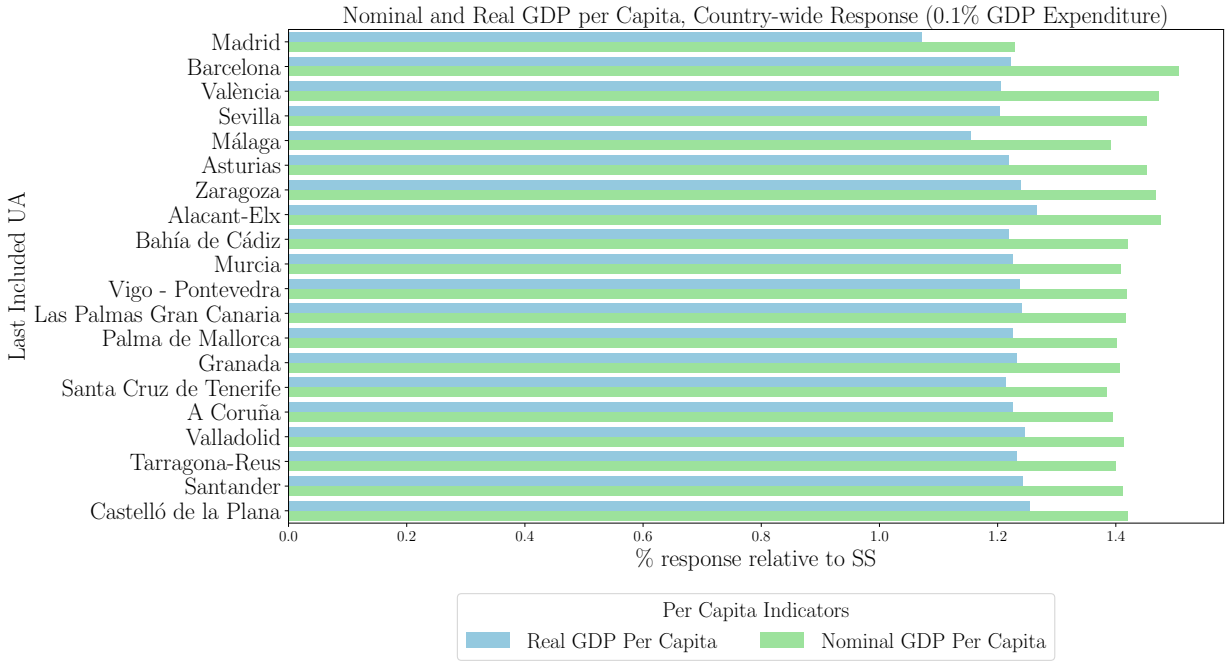
While solving for the global combinatorial problem would be expensive, a reasonable first approximation would be to perform a cumulative targeting criteria where the largest, which are on average more skilled and productive (and constrained), are targeted first, and the set is gradually expanded to cover the remaining UAs.

The results in terms of output are provided in figure 13. Note that now the y-axis indicates the last included UA in the set of targeted UAs. Upon inspection, a notable first result is the hump-shape as a function of the included UAs. Targeting only Madrid results in an efficiency gain, yet further expanding the set to include Barcelona and València attains the largest gains in (nominal) productivity, and including up to Alacant-Elx results in the largest real GDP per capita gains.

There are three main forces at play. One is purely mechanical: including UAs that had large individual responses increases the response of the set. However, the two other are endogenous: enlarging the set of included UAs allows for a broader coverage of the policy and at the same time

[44] Note that all workers regardless of location still need to pay for the subsidy in the form of increased labour taxes $\tau_{L,p} > \tau_{L,ss}$.

Figure 13: Response of Country-wide Nominal and Real GDP per Capita to **Cumulative** Place-based Entrepreneurial Transfers (0.1% of Country-wide GDP Expenditure)



Notes: Response of the country-wide nominal and real (deflated by the local price index $p_{h,l}^{(1-\alpha_c)}$) GDP per capita to **cumulative** place-based entrepreneurial transfers T_l such that total programme expenditure is 0.1% of GDP. The y-axis denotes which UA is being included last in the set of UAs being targeted. Blue bars indicate the response of the country-wide per capita real GDP while green bars do likewise for the nominal GDP per capita. The last row is equivalent to targeting the entire country, since all UAs are included.

splits the congestion forces among multiple UAs. As more funds are diverted towards less constrained locations, the response becomes less positive as the marginal spent resource is generating lower MPK gains.

Overall, the highest real GDP per capita gains are attained when targeting the eight largest UAs of the State, leading to a 1.257% efficiency gain compared to the untargeted policy's 1.25%. While non-negligible, it is worthwhile to consider the spatial disparities this policy would entail. Compared to the untargeted policy's 0.92% disparity increase, the increase under the "optimal" is 1.19%. This presents a challenging situation for policymakers, as they face a trade-off between achieving efficiency gains and addressing regional disparities.

Lastly on the lump-sum transfers programme, we will explore the implications of the cumulative targeting for welfare. Figure 14 presents the results.

A few comments are in place. First, contrary to the single-targeting criteria, there is consis-

Figure 14: Response of Country-wide per Capita Welfare, Aggregate and by Occupation, to **Cumulative** Place-based Entrepreneurial Transfers (0.1% of Country-wide GDP Expenditure)

Notes: Response of the country-wide per capita welfare, aggregate and by occupations, to **cumulative** place-based entrepreneurial transfers T_l such that total programme expenditure is 0.1% of GDP. The y-axis denotes which UA is being included last in the set of UAs being targeted. Blue bars indicate the response of the country-wide per capita welfare. Green bars do likewise for the workers. Violet bars denote that of entrepreneurs. The last row is equivalent to targeting the entire country, since all UAs are included. The results for the response of the total stock of welfare rather than the per capita one can be found at: G.32.

tency: those locations that provide the largest increases in (nominal) efficiency (Madrid, Barcelona and València), also provide the largest increases in welfare. We interpret this as the relaxation of the congestion forces: in the single targeting case welfare in València responded less strongly than efficiency and resulted in a lower response than that of the untargeted policy. In contrast, adding València to the set of targeted locations in this exercise *increases* the aggregate welfare response. Second, the hump-shaped pattern also holds in this the case of welfare. Intuitively, targeting a single location leads to the highest house price increases and a lower mass of the population is able to enjoy the benefits of the policy. As more locations are targeted, this reduces the congestion forces while expanding the policy to entrepreneurs and switchers to a larger mass of the population. However, as more unconstrained locations are targeted the benefits slowly decay as the additional resource spend on the policy results in a more muted response of wages and entry into

entrepreneurship.

Overall, the "optimal" cumulative policy is able to attain a 2.07% increase country-wide welfare as opposed to the 1.82% attainable with an untargeted policy. In terms of regional disparities, the optimal policy actually *decreases* the dispersion in regional welfare per capita by -0.74%, as opposed to the -0.68% of the untargeted policy.

To sum up, these policy experiments have shown that space does matter for policy, and that targeting a subset of the most productive and constrained UAs leads to higher welfare and efficiency gains than an untargeted policy at a small regional disparity cost.

7.2 The role of financial frictions

We have studied the welfare, efficiency and regional disparity implications of place-based entrepreneurial subsidies. Nonetheless, a pivotal question remains: what is the role played by financial frictions in driving these results?

In figure 8 we discussed how the absence of financial frictions would result in the equalization of marginal returns to capital at the user cost level across space and entrepreneurs, as each entrepreneur would be able to finance their unconstrained level of capital everywhere. Under these conditions, would place-based entrepreneurial subsidies attain the welfare and efficiency gains that we documented before?

Figures G.35 and G.36 in Appendix G.3 present the obtained results. First, in terms of production, we obtain a positive marginal response that slowly increases as the set of targeted UAs expands. This might strike as surprising, given the absence of financial frictions in the capital rental market and the efficient selection into entrepreneurship. [45] However, there is an extensive margin effect. The presence of the subsidies induces marginally less productive entrepreneurs to enter, which drives up the total number of entrepreneurs and hence total production. [46] Nevertheless, the marginal entrepreneur is markedly less productive, which drives down the average scale of production per entrepreneur. Thus, the increased aggregate production is driven by the inefficient extensive margin of entry into entrepreneurship.

[45] Conditional on the calibrated spatial allocation of agents, which is not guaranteed to be efficient.

[46] In the case of only targeting Madrid, the extensive margin of additional entry is limited and the larger subsidies result in the average entrepreneurial productivity declining substantially.

In terms of welfare, the responses are either indistinguishable from zero or negative. Intuitively, increasing the labour tax τ_l to subsidize entrepreneurial subsidies that attain a very limited gain in production and hence wages results in a *de facto* redistribution from on average poorer workers to on average richer entrepreneurs, which, coupled with the concavity of the value function leads to the attained negative responses.

Therefore, the main conclusion from this short discussion on the role of financial frictions is that they are the main drivers of the obtained simultaneous welfare and efficiency gains and that the bulk of the efficiency gains (>85%) documented in the previous policy experiments is explained by these.

7.3 Does the design of the policy matter?

So far, we have discussed the long-run effects of place-based entrepreneurial lump-sum subsidies. However, at this point a natural question arises: does the design of the policy matter?

To tackle this question, we perform an additional set of exercises which consist on handing out proportional-to-collateral transfers (see appendix G.2). This is relevant in practice since it is reminiscent of the commonplace size-dependent policies such as investment subsidies, loans or guarantees. Compared to the lump-sum transfers, this policy has the distinguishing feature that now the majority of the funds are concentrated on larger entrepreneurs who are less constrained on average.

Still, the results point to a positive efficiency response. However, given that the MPK Premiums for this set of entrepreneurs are smaller, the responses are muted in comparison to the lump-sum transfers, and the largest response in real GDP per capita is 0.18%, in clear contrast with the 1.257% obtained through the lump-sum transfers. In fact, the efficiency response is no larger than in the frictionless case under the lump-sum programme. In terms of welfare, increasing the labour tax in order to subsidize this less constrained set of entrepreneurs leads to smaller wage increases and entry into entrepreneurship, which translates into a negative welfare response across the board, regardless of what UA is targeted.

Therefore, this exercise highlights the importance of the specific implementation of place-based policies. Given that both within and across locations the larger MPK Premiums are found

in the lower 50% of the wealth distribution, policies that target this set of entrepreneurs are expected to attain higher returns to investment, particularly those located in more productive UAs. [47] In particular, we have discussed how a simple place-based lump-sum transfer leads to more desirable outcomes than an informationally more intensive size-dependent proportional transfer programme.

8 Conclusion

The aim of this paper has been to study the policy implications of the interaction between entrepreneurship, financial frictions and space. We posited that this interplay could lead to misallocation of capital across Urban Areas and that would point directly to space as being a potential targeting channel for entrepreneurial subsidies.

We have developed a quantitative dynamic spatial framework with occupational choice that has allowed us to study this question. Taken to the largest 20 Urban Areas of the Spanish State, this model has provided a clear prediction: there is spatial misallocation of capital as a result of the interaction between financial frictions and heterogeneous productivities across space, and more productive Urban Areas, which are on average larger and more educated, are more constrained.

Solving this model efficiently has been possible by exploiting the parallelizable nature of the commonly used algorithms in Macroeconomics (VFI, PFI, Young (2010)'s method, ...) by employing dedicated hardware components whose architecture is optimized around throughput and parallel tasks: GPUs (Graphics Processing Units). We have shown how employing these for this class of models leads to substantial speedups, with the range varying from around 2 to 5 orders of magnitude (x60 to x20,000) depending on the specific algorithm.

We have then studied the long-run effects of place-based entrepreneurial policies. First, we have concluded that while single-targeted transfers to the most productive UAs led to the highest country-wide welfare and efficiency gains, their effects on regional disparities would be too strong to be ignored. Second, we have considered targeting a set of UAs rather than individual locations and have found that targeting a subset of the largest and most productive UAs led to higher welfare and efficiency gains than a country-wide untargeted policy at little regional disparity cost.

[47] Subject to the consideration of decreasing marginal returns to investment in a given location.

The current priority is to compute the actual optimal policy in this framework.

While this paper has made progress along several fronts, there remain several potential avenues for future research. First, we have assumed a symmetric rental market across space. However, there exists literature pointing to the effects of local bank branch closures ((Jiménez et al., 2022), (Amberg and Becker, 2024)), particularly on the survivability and credit access of SMEs. The model, augmented with a banking sector, could be employed as a laboratory to study the effects of such local credit shocks. Second, future research could focus on solving the dynamic social planner's problem, extending the static frameworks of Rossi-Hansberg et al. (2019), Fajgelbaum and Gaubert (2020) or Fajgelbaum and Schaal (2020) into dynamic settings. This could yield new insights into how the optimal allocation of resources across space is influenced by capital accumulation subject to frictions.

References

AIYAGARI, S. R. (1994): "Uninsured idiosyncratic risk and aggregate saving," *The Quarterly Journal of Economics*, 109, 659–684.

ALDRICH, E. M. (2014): "Gpu computing in economics," in *Handbook of computational economics*, Elsevier, vol. 3, 557–598.

ALDRICH, E. M., J. FERNÁNDEZ-VILLAVERDE, A. R. GALLANT, AND J. F. RUBIO-RAMÍREZ (2011): "Tapping the supercomputer under your desk: Solving dynamic equilibrium models with graphics processors," *Journal of Economic Dynamics and Control*, 35, 386–393.

AMBERG, N. AND B. BECKER (2024): "Banking without branches," *Available at SSRN 4723120*.

ÅSTEBRO, T. AND J. TÅG (2017): "Gross, net, and new job creation by entrepreneurs," *Journal of Business Venturing Insights*, 8, 64–70.

AUTOR, D. H. (2019): "Work of the Past, Work of the Future," in *AEA Papers and Proceedings*, American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203, vol. 109, 1–32.

BANERJEE, A. V. AND E. DUFLO (2014): "Do firms want to borrow more? Testing credit constraints using a directed lending program," *Review of Economic Studies*, 81, 572–607.

BANERJEE, R. AND K. BLICKLE (2021): "Financial frictions, real estate collateral and small firm activity in Europe," *European Economic Review*, 138, 103823.

BELLUCCI, A., G. GUCCIARDI, D. NEPELSKI, ET AL. (2021): *Venture Capital in Europe. Evidence-based insights about Venture Capitalists and venture capital-backed firms*, Publications Office of the European Union.

BOAR, C. AND V. MIDRIGAN (2023): "Should we tax capital income or wealth?" *American Economic Review: Insights*, 5, 259–274.

BUERA, F. J. AND Y. SHIN (2013): "Financial frictions and the persistence of history: A quantitative exploration," *Journal of Political Economy*, 121, 221–272.

CAGETTI, M. AND M. DE NARDI (2006): "Entrepreneurship, frictions, and wealth," *Journal of political Economy*, 114, 835–870.

CALIENDO, L., M. DVORKIN, AND F. PARRO (2019): "Trade and labor market dynamics: General equilibrium analysis of the china trade shock," *Econometrica*, 87, 741–835.

CAN, E. (2022): "Income taxation, entrepreneurship, and incorporation status of self-employment," *International Tax and Public Finance*, 29, 1260–1293.

CARLINO, G. A., S. CHATTERJEE, AND R. M. HUNT (2007): "Urban density and the rate of invention," *Journal of urban economics*, 61, 389–419.

COMBES, P.-P., G. DURANTON, L. GOBILLON, D. PUGA, AND S. ROUX (2012): "The productivity advantages of large cities: Distinguishing agglomeration from firm selection," *Econometrica*, 80, 2543–2594.

DESMET, K., D. K. NAGY, AND E. ROSSI-HANSBERG (2018): "The geography of development," *Journal of Political Economy*, 126, 903–983.

DESMET, K. AND E. ROSSI-HANSBERG (2013): "Urban accounting and welfare," *American Economic Review*, 103, 2296–2327.

——— (2014): "Spatial development," *American Economic Review*, 104, 1211–1243.

DURANTON, G. AND D. PUGA (2004): "Micro-foundations of urban agglomeration economies," in *Handbook of regional and urban economics*, Elsevier, vol. 4, 2063–2117.

——— (2020): "The economics of urban density," *Journal of economic perspectives*, 34, 3–26.

EECKHOUT, J. (2004): "Gibrat's law for (all) cities," *American Economic Review*, 94, 1429–1451.

EECKHOUT, J., C. HEDTRICH, AND R. PINHEIRO (2021): "IT and urban polarization," .

ERIKSSON, R. AND M. RATAJ (2019): "The geography of starts-ups in Sweden. The role of human capital, social capital and agglomeration," *Entrepreneurship & Regional Development*, 31, 735–754.

EVANS, D. S. AND B. JOVANOVIC (1989): "An estimated model of entrepreneurial choice under liquidity constraints," *Journal of political economy*, 97, 808–827.

FAJGELBAUM, P. D. AND C. GAUBERT (2020): "Optimal spatial policies, geography, and sorting," *The Quarterly Journal of Economics*, 135, 959–1036.

FAJGELBAUM, P. D., E. MORALES, J. C. SUÁREZ SERRATO, AND O. ZIDAR (2019): "State taxes and spatial misallocation," *The Review of Economic Studies*, 86, 333–376.

FAJGELBAUM, P. D. AND E. SCHAAL (2020): "Optimal transport networks in spatial equilibrium," *Econometrica*, 88, 1411–1452.

FERNÁNDEZ-VILLAVERDE, J. AND D. Z. VALENCIA (2018): "A practical guide to parallelization in economics," Tech. rep., National Bureau of Economic Research.

FERRARI, A. AND R. OSSA (2023): "A quantitative analysis of subsidy competition in the US," *Journal of Public Economics*, 224, 104919.

GANONG, P. AND D. SHOAG (2017): "Why has regional income convergence in the US declined?" *Journal of Urban Economics*, 102, 76–90.

GHANI, E., W. R. KERR, AND S. O'CONNELL (2017): "Spatial determinants of entrepreneurship in India," in *Entrepreneurship in a Regional Context*, Routledge, 133–151.

GIANNONE, E. (2017): "Skill-biased technical change and regional convergence," in *2017 Meeting Papers*, Society for Economic Dynamics, 190.

GIANNONE, E., Q. LI, N. PAIXAO, AND X. PANG (2020): "Unpacking moving," *Unpublished manuscript*.

GILCHRIST, S., J. W. SIM, AND E. ZAKRAJŠEK (2013): "Misallocation and financial market frictions: Some direct evidence from the dispersion in borrowing costs," *Review of Economic Dynamics*, 16, 159–176.

GLAESER, E. L. AND J. D. GOTTLIEB (2009): "The wealth of cities: Agglomeration economies and spatial equilibrium in the United States," *Journal of economic literature*, 47, 983–1028.

GLAESER, E. L., W. R. KERR, AND G. A. PONZETTO (2010): "Clusters of entrepreneurship," *Journal of urban economics*, 67, 150–168.

GLAESER, E. L. AND M. G. RESSEGER (2010): "The complementarity between cities and skills," *Journal of regional science*, 50, 221–244.

GUIDE, D. (2013): "Cuda c programming guide," *NVIDIA, July*, 29, 31.

——— (2020): "Cuda c++ best practices guide," .

HALTIWANGER, J., R. S. JARMIN, AND J. MIRANDA (2013): "Who creates jobs? Small versus large versus young," *Review of Economics and Statistics*, 95, 347–361.

HERKENHOFF, K. F., L. E. OHANIAN, AND E. C. PRESCOTT (2018): "Tarnishing the golden and empire states: Land-use restrictions and the US economic slowdown," *Journal of Monetary Economics*, 93, 89–109.

HOLTZ-EAKIN, D., D. JOULFAIAN, AND H. S. ROSEN (1994): "Sticking it out: Entrepreneurial survival and liquidity constraints," *Journal of Political economy*, 102, 53–75.

HUNDT, C. AND R. STERNBERG (2016): "Explaining new firm creation in Europe from a spatial and time perspective: A multilevel analysis based upon data of individuals, regions and countries," *Papers in Regional Science*, 95, 223–258.

JIMÉNEZ, G., A. MARTIN-OLIVER, J.-L. PEYDRÓ, A. TOLDRA-SIMATS, AND S. VICENTE (2022): "Do bank branches matter? Evidence from mandatory branch closings," *Unpublished manuscript*.

KLEINMAN, B., E. LIU, AND S. J. REDDING (2023): "Dynamic spatial general equilibrium," *Econometrica*, 91, 385–424.

LEVINE, R. AND Y. RUBINSTEIN (2017): "Smart and illicit: who becomes an entrepreneur and do they earn more?" *The Quarterly Journal of Economics*, 132, 963–1018.

LÓPEZ-RODRÍGUEZ, D. AND C. GARCÍA CIRIA (2018): "Spain´ s tax structure in the context of the European Union," *Documentos Ocasionales/Banco de España, 1810*.

MCFADDEN, D. (1972): "Conditional logit analysis of qualitative choice behavior," .

MOLL, B. (2014): "Productivity losses from financial frictions: Can self-financing undo capital misallocation?" *American Economic Review*, 104, 3186–3221.

MORAZZONI, M. AND A. SY (2022): "Female entrepreneurship, financial frictions and capital misallocation in the US," *Journal of Monetary Economics*, 129, 93–118.

NIEBUHR, A., J. C. PETERS, AND A. SCHMIDKE (2020): "Spatial sorting of innovative firms and heterogeneous effects of agglomeration on innovation in Germany," *The Journal of Technology Transfer*, 45, 1343–1375.

QUADRINI, V. (2000): "Entrepreneurship, saving, and social mobility," *Review of economic dynamics*, 3, 1–40.

ROCA, J. D. L. AND D. PUGA (2017): "Learning by working in big cities," *The Review of Economic Studies*, 84, 106–142.

ROSENTHAL, S. S. AND W. C. STRANGE (2004): "Evidence on the nature and sources of agglomeration economies," in *Handbook of regional and urban economics*, Elsevier, vol. 4, 2119–2171.

ROSSI-HANSBERG, E., P.-D. SARTE, AND F. SCHWARTZMAN (2019): "Cognitive hubs and spatial redistribution," Tech. rep., National Bureau of Economic Research.

SAIZ, A. (2010): "The geographic determinants of housing supply," *The Quarterly Journal of Economics*, 125, 1253–1296.

SCHMALZ, M. C., D. A. SRAER, AND D. THESMAR (2017): "Housing collateral and entrepreneurship," *The Journal of Finance*, 72, 99–132.

SHEN, J. (2023): "Capital misallocation and financial market frictions: Empirical evidence from equity cost of capital," *International Review of Economics & Finance*, 87, 486–504.

YOUNG, E. R. (2010): "Solving the incomplete markets model with aggregate uncertainty using the Krusell–Smith algorithm and non-stochastic simulations," *Journal of Economic Dynamics and Control*, 34, 36–41.

A Empirical Appendix

This section presents the data employed for the construction of the three main stylized facts in the main paper. Further empirical work aimed at calibrating the model is relegated to the calibration section (5.1).

A.1 MCVL

The main dataset employed to conduct the empirical research is the MCVL (Muestra Continua de Vidas Laborales, or Continous Sample of Employment Histories in english). This is an administrative panel data covering a non-stratified random 4% of those Spanish citizens that have had any relationship with the Social Security system during the previous year. This includes workers, contributive pension recipients, recipients of unemployment subsidies, and, crucially for our analysis, entrepreneurs. The three key advantages of the MCVL for this analysis are that it allows us to identify entrepreneurs, and that is has both a spatial and temporal dimension. First, an overall introduction to the dataset is provided, while the sample construction and restrictions, the identification of entrepreneurs in the dataset and the definition of our geographical units merit their separate discussion.

A.1.1 Data Description

This dataset is a collection of Social Security, Tax records and Census data. Therefore, a wide range of individual characteristics are available. The Social Security component provides valuable information such as the employment status of the individual (retired, full-time worker, part-time worker...), the sector of the firm at up to 3 NACE digits, the occupational category of the worker (labourer, analyst, director...) and the contribution regime, from which entrepreneurs can be identified. Since it is also merged with the data from the AEAT (the Spanish state's tax office, Agencia Tributaria) at the yearly level, the uncensored gross labour income [48] of the individuals for the year can be observed, with the notable exception for our purposes of the self-employed [49]. Lastly,

[48] More precisely, every income category in form 190 can be observed, which consists mainly of labour income. Therefore, capital gains, wealth and other financial income are not observable.

[49] Note that although the income for the self-employed can not be observed (including incorporated entrepreneurial episodes), one can still recover the path of incomes prior to this event.

data from the Census or the Municipality's Padrón provide insights on the socio-demographic composition of the labour force, such as age, gender and nationality.

The unit of observation is at the affiliation episode level. Thus, every change in status of the relationship with the SS is recorded, up to a precision of 1 day. Therefore, one can construct a panel data by combining all these affiliation episodes [50], effectively backing out the employment history of the individual. This is paramount to our analysis, as it will enable us to identify the moment in time when the individual decides to become an entrepreneur, as well as observing the potential subsequent exit.

Crucially, the MCVL provides geographical data both on the firm and the individual. Data is disaggregated up to municipalities of 40,000 inhabitants, which account for roughly 70% of the Spanish population. The municipalities below this threshold are aggregated at the province level. The availability of this geographical dimension [51] enables us to explore the spatial variation of entrepreneurship across Urban Areas (UA henceforth).

A.1.2 Identifying the Entrepreneurs in the Data

As argued previously, a major advantage of the MCVL is the ability single out the entrepreneurs. Every individual (in fact, every affiliation episode) that is present in the MCVL has a "Contribution Regime" value that determines the nature of the relationship with the Social Security in terms of the specific contribution requirements, type of employment and rights. The two main groups are the 111 and 521, which are the general and entrepreneurial regime, respectively [52]. For the sake of our discussion, the general regime 111 shall be understood as that including all types of relationships that are not self-employment. This includes those receiving unemployment benefits, those getting pensions and employed workers, among others.

Starting in 2013, the MCVL includes an additional variable which provides information on the sub-categories among the self-employed within the 521 contribution regime. The following table

[50] Entry into the dataset is dependent on a permanent tax identifier at the individual level and everyone that has had a relationship of at least 1 day with the SS is a potential candidate. Exit happens if the relationship with the SS has been broken for more than one year. This includes mostly people who have died and those that have permanently moved abroad (note that pensioners are also part of the sample). The representativeness of the sample is checked at every release.

[51] Compared to another competing dataset for entrepreneurial research, the GEM, this is a major advantage of the MCVL for our purposes.

[52] The entrepreneurial regime is composed of those that work without any ties to an employer, among which there are different classifications which shall be discussed shortly

presents the sub-categories that are available within this contribution regime:

Table A.1: Entrepreneurial Types in the MCVL and in Official Statistics

Type	%
Self-Employed	60.75%
Relative	7.4%
Coop. Member	17.2%
Incorporated	13.2%

(a) Spanish Ministry of Social Security (2018 Q3)

Type	%
Self-Employed	60.84%
Relative	7.46%
Coop. Member	17.02%
Incorporated	13.05%

(b) Constructed from the MCVL (2018)

Notes: Sub-types of self-employed in the Spanish state's economy, as reported by the Spanish Ministry of Social Security (left) and as constructed from the MCVL (right)

Table $(A.1)$ provides data on the distribution of the sub-types of the self-employed for the year 2018 [53]. Although there are other minor categories, the main included four account for roughly 98.5% of the entire pool of self employed [54]. The self-employed are those individuals that work on their own without ties to any direct employer and are not incorporated, that is, their firm/endeavour has no separate legal personality. The relatives category refer to those that may be under the umbrella of the self-employed (in the broad sense) but are not themselves the major responsibles of any endeavour. Cooperative members are those that form part of a cooperative or worker-managed firm. The last group is the one that will be the focus of the analysis, the incorporated entrepreneurs. These are the individuals that have set up a formal firm with a distinct legal personality as part of the entrepreneurial endeavour. Note that becoming an incorporated entrepreneur is a rather unlikely event in the context of the Spanish state's economy. At any given year, around 15% of all individuals in the sample has at least one affiliation episode in the 521th contribution regime, and out of these just over 13% are incorporated.

The distinction between the self-employed and incorporated entrepreneurs is non-trivial. Incorporated entreprenerus, as opposed to the unincorporated, need to pay a fixed upfront cost for establishing the firm (a minimum capital of 3,000 euros is required), pay the corporate tax on the profits, and have tighter regulatory requirements (tighter accounting requirements, may be subject to being audited ...). The optimal choice for becoming an incorporated or unincorporated may

[53] The distributional composition is very similar across years. The focus on the year 2018 is arbitrary and its objective is to provide a counterpart to the official statistics.

[54] The remaining sub-types include religious institutions, those under temporary administrative changes and professional associations, none of which are of particular relevance for the present analysis.

depend on a variety of factors, which range from the tax being paid to the government (the unincorporated pay the personal income tax, which may be higher than the corporate tax for higher incomes [55]) as well as the value attached to the limited liability, which the incorporated enjoy but the unincorporated do not. In the context of the Spanish state's legal system, those that are required to contribute as incorporated are the following: (i) those that have over 33% of ownership of a firm (ii) those that have over 25% of ownership of a firm and have active managerial duties (iii) those involved in a firm (be it managers, suppliers, workers) with a family share of over 50%.

Legal requirements and differences aside, this distinction has been found to be relevant in the field. Levine and Rubinstein (2017) argue that (i) the incorporated then to engage in more non-routine cognitive tasks than either workers or unincorporated (ii) on the psychological realm, the incorporated are shown to have shown more self-steem and have engaged in more illicit activities during their youth (iii) the incorporated earn significantly more per hour than either workers or the unincorporated, and work more hours. In addition Åstebro and Tåg (2017) find that the incorporated status is a strong predictor of job creation. While the average self-employed (across all categories) creates on average 0.66 jobs two years after creation, the incorporated create around 2.48.

Who are the Entrepreneurs in the Spanish Economy?

Having outlined the procedure employed to identify the entrepreneurs in the data, in the interest of corroborating that the profile of the thereof in the Spanish economy is consistent with the findings in the literature, table $(A.2)$ displays the differences in individual characteristics between the incorporated entrants and the rest of the population.

A few comments are in place regarding the obtained differences. First, as documented in the literature, incorporated entrepreneurs tend to be more male and more skilled than the broader population. Interestingly, however, incorporated entrants in the dataset are found to have accumulated less income and working experience than the overall population. Regarding mobility, while the differences in migration patterns out and in is not significant at the common significance levels, the data points towards a correlation between the moment when incorporation is chosen and migration towards or from a given UA.

[55] Can (2022) argues that higher personal income tax rates encourage incorporated entrepreneurship, while entry into unincorporated self-employment is disencouraged.

Table A.2: Individual Characteristics of Incorporated Entrants

	Incorporated Entrants	Rest of the Population	
	Mean	Mean	diff.
Sex: Male=1 Female=2	1.292	1.404	-0.112***
Age of the Individual	39.272	39.501	-0.229
Cumulative Income	63002.470	67936.345	-4933.875**
Share of Highly Skilled	0.420	0.297	0.123***
Cumulative Experience as Worker (Days)	4288.018	5511.887	-1223.868***
Foreigner	0.144	0.108	0.036***
Foreigner (non-high GDP countries)	0.109	0.091	0.018*
Foreigner (high GDP countries)	0.035	0.017	0.018***
Internal across UA migrant (incoming)	0.028	0.019	0.009*
Internal across UA migrant (outgoing)	0.018	0.012	0.006*
Observations (Individual $\times$ Year)	1858	816192	818050

Notes: Individual characteristics of incorporated entrants and the rest of the population. Incorporated entrants are defined as the workers in the previous period that transition to becoming an incorporated entrepreneur during the next. The observations are at the individual-year level.

A.1.3 Geographical Areas under study: Urban Areas

The geographical units of interest are the Urban Areas (UA henceforth) of the Spanish State. The main reason for focusing on UAs rather than individual municipalities is that given the extent of commuting for labour purposes, these are the effective labour markets of the Spanish economy. Following Roca and Puga (2017) and the Spanish Ministry of Housing and Urban Agenda, 85 UA areas are identified. This set of UAs accounts for around 68% of the Spanish population yet 10% of the area (Roca and Puga, 2017) [56]. Madrid and Barcelona are the largest UAs, with 6,258,132 and 5,270,691 inhabitants, respectively [57].

An important restriction made on the set of included UAs is that since the MCVL provides no fiscal data for the UAs in the Basque Autonomous Region or Navarre (Bilbo, Donostia, Gasteiz and Iruñea) given their distinct tax offices [58], these four UAs are left out of the analysis. In addition,

[56]Note that although the density of the Spanish state is around 96 people per squared km (far behind Germany's 233), it is home to the densest inhabited squared km in Europe. This partly has to do with the concentration of the Spanish population on coastal areas (around 70% of the total population). For an insightful infographic on the matter check `https://www.elconfidencial.com/mundo/europa/2024-05-21/mapa-poblacion-europa-gente-kilometro-cuadrado_3884367/`.

[57]For context, Madrid and Barcelona are the 2nd and 5th largest UAs in the EU27.

[58]The historical roots of the current Basque Autonomous Region and Navarre's fiscal autonomy date back to the Middle Ages, around the 8-13th centuries, when these regions were part of the Kingdom of Navarre. During this period, it was common for regions to hold local fueros (charters or rights), which included privileges such as the ability to collect taxes independently. Even after the Kingdom of Navarre was largely conquered by Castile in 1512, these regions retained a significant degree of their administrative autonomy.

In the 19th century, efforts at centralization by the liberal supporters of Queen Isabella II (known as the Isabelinos) posed a threat to these local rights. The Carlist Wars (1833–1876), in which the Basque provinces played a central role,

those UAs that have no central municipality above 40,000 inhabitants (Soria, Teruel, Sant Feliu de Guixols, ...) are not observable in the MCVL are are therefore also excluded. This results in a final set of 77 UAs.

As mentioned previously, the MCVL contains geographically dis-aggregated data starting at municipalities above 40,000 inhabitants. This has non-trivial consequences for the construction of the UAs. For example, the UA of Madrid is made up of 52 distinct municipalities, 23 of which have populations above this threshold. In contrast, the Barcelona UA is made up of 165 municipalities, the majority of which (143) fall below this threshold. This results in the UA of Madrid retaining 91.7% of its population in the MCVL compared to Barcelona's 72.53%. To address this issue, the following procedure is employed. Since a precise calibration is desired for the model, the population distribution of the UAs will be the one obtained in the MCVL. This guarantees that mobility rates, earnings, sectoral distributions and entrepreneurial figures are accurate at each UA. That notwithstanding, Duranton and Puga (2020)'s "experienced density" measure will be employed when constructing measures for the agglomeration forces. The rationale for this stems from the fact that firms in these UAs "inhabit" within the actual UA rather than the subset provided by the MCVL.

Following Duranton and Puga (2020), a distinct measure to the standard population density is employed in order to measure a given UA's "thickness". The proposed metric, "experienced density", consists on measuring the number of people within a "x"km radius of the average people in a given UA. The main advantages of this are twofold. First, since administrative boundaries of municipalities are influenced by historical, agricultural and geographical factors, the ratio of actually built space over the total terrain varies substantially across municipalities and UAs. Second, polycentric UAs might be more populated and therefore have a higher population density than a monocentric UA [59].

Data from the WorldPop project is employed (`https://hub.worldpop.org/geodata/listing?id=74`) in order to construct this measure at an annual frequency over the 2013-2018 period. This

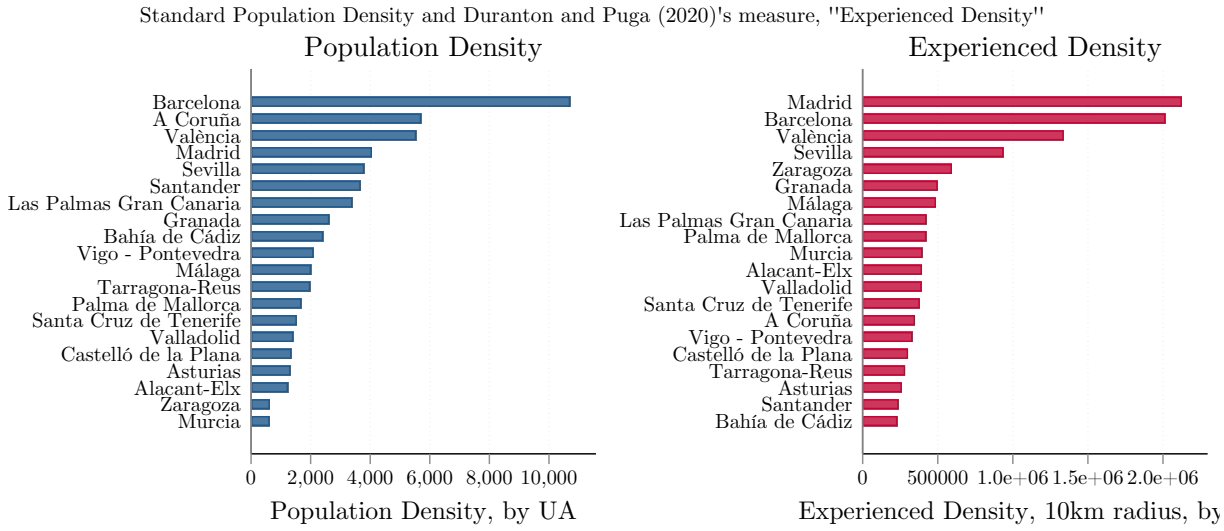
were partly fueled by resistance to these centralizing efforts. The wars culminated in the Convention of Bergara (1839), which secured the preservation of the fueros in these territories, albeit to a more limited extent.

To this day, the Spanish Constitution of 1978 recognizes the historical rights of the Basque Country and Navarre, reflecting their unique legal and fiscal autonomy within the Spanish State.

[59] As Roca and Puga (2017) point out for the case of the Spanish state, while the Asturias UA is made up of three main municipalities and is more populated than Zaragoza, the latter is a monocentric UA whose population is spatially concentrated, whilst that of Asturias is spread across many municipalities, and is hence more dispersed.

dataset contains the population counts of the Spanish state at a 1km $\times$ 1km resolution. The followed procedure is to compute the population in a 10km radius at each grid cell and then to compute the weighted average across all grid cells within a given UA. Figure $(A.1)$ reports the standard population density and experienced density for the biggest 20 UAs.

Figure A.1: Population Density and "Experienced Density" of the Spanish State's UAs



Notes: Standard arithmetic population density (population of the UA divided by terrain in squared kilometers) and Duranton and Puga (2020)'s "experienced density" measure for the largest 20 UA of the Spanish state. The left panel displays the former while the second displays the latter. Experienced density is defined as the amount of people in a 10km radius from the average person in a given UA. The provided data is the average over the 2013-2018 period.

If one was to measure a labour market's thickness by the standard density measure, A Coruña (the 16th in population) would be the second densest UA, and some of the largest and most vibrant UAs of the state would rank in modest positions (Madrid, Sevilla, Zaragoza, Málaga...). Duranton and Puga (2020)'s measure overturns these results in part by weighting the *absolute* rather than *relative* density more heavily. This results in a ranking that puts Madrid, Barcelona, València and Seville in the first positions, which are at the same time the major hubs of economic activity.

A.1.4 Sample Construction and Restrictions

The main challenge in structuring the data is to transition from an affiliation episode level panel dataset including all provinces and municipalities above 40,000 inhabitants, NACE 3digit sectors,

employment categories (workers, entrepreneurs, unemployed, retired...) and occupation categories (labourer, director...) to a dataset that can be employed in order to compute the micro and macroeconomic moments of interest for the quantitative model.

The employed MCVL dataset is that comprising the years 2012 to 2019. However, as mentioned before, given the structure of the data, all past affiliation episodes can be observed for a given individual in the sample. First, the duration of all spells by year are computed and the cumulative experience of a person as both worker and incorporated entrepreneur is registered up to that point in time. Second, following Roca and Puga (2017), affiliation episodes in the public sector, apprenticeships, other entrepreneurial types other than incorporated, non-contributive contribution regimes (widowhood...) are dropped. The same is done for the following 1st NACE sectors: Agriculture, Forestry and Fishing, Extractive Activities, Public Administration and International Organizations. Third, since the interest in this work lies on exploring the entry, exit and stock of entrepreneurship in the 2013-2018 period, due to data limitations on identifying the incorporated entrepreneurs before 2013, all affiliation episodes before 2012 are dropped. This leaves us with 7,330,877 affiliation episodes.

The next key step is to aggregate the potential multiple per year and individual episodes to the yearly level. To this end, the longest affiliation episode per contribution regime (worker or incorporated entrepreneur) per year is identified. If the individual has been an active incorporated entrepreneur during year, then the incorporated episode is retained for this individual in a given year. Otherwise, the longest labour affiliation episode is kept. This pre-eminence of the incorporated episodes is motivated by the fact that the main interest of the present work is studying when people decide to enter and exit incorporated entrepreneurship, and to that end it is crucial to keep all entrepreneurial episodes [60] [61] [62]. In order to be consistent with the model, entry is computed in a year-to-year basis, by focusing on the transitions from those that are only workers in time period

[60] Suppose that an individual had two affiliation episodes during the year, one as a worker and one as an incorporated entrepreneur. Suppose the longest was the labour contract, lasting from January to September, and the entrepreneurial lasted for the remainder of the year. If one was to simply use the longest contract, the entrepreneurial endeavour of this individual would not be recorded.

[61] Although the analysis could be done, potentially, at the monthly or even daily level, the aggregation to the yearly level is of importance for the quantitative analysis given that most of the data available to calibrate the model (SABI balance sheet data, experienced density data, fiscal income data, ...) are available at this level only. In addition, given the granularity of the objects of interest at the location $\times$ sector $\times$ skill level, aggregation is required in order to have a sufficient sample size at each of these tuples.

[62] The accumulated experience as a worker during the discarded episodes is recorded.

t to those that have begun an entrepreneurial endeavour at $t+1$. This results in just above 5,000 entries. Since entries in 2012 and exits in 2019 can not be observed, these years are dropped.

The previous step results in 4,110,153 individual $\times$ year observations. Next, the fiscal, personal and processed affiliation data is merged. At this point, all individuals below 16 and above 65 years are dropped from the sample. Those not residing in the UAs are also dropped.

In order to alleviate the dimensionality in the quantitative exercise, the set of occupations and sectors are compressed. The occupations are classified into "low" and "high" skilled based on Roca and Puga (2017) [63]. The criteria to group the sectors is based on the skill composition of both the incorporated entrepreneurs and the workers. Three main groups are created: the "low skilled" (transportation, hostelry and construction), the "medium skilled" (manufacturing, artistic sector, sales & car repairs), and "high skilled" (information and telecommunications, the professional scientific and technical services sector). Those that do not have either of these dimensions in their yearly aggregated affiliation episode are dropped (notably, the unemployed). This leaves us with 1,179,944 individual $\times$ year observations in the UAs of interest. Figure $(A.2)$ presents the joint distribution between skills and sectors for both incorporated and workers.

The obtained sector classification is fairly consistent across occupations. The low sectors have around 28% of skilled entrepreneurs, the medium ones around 38% and the high jump to around 68% on average. From the workers' side, low sectors have about 80% of low skilled, the medium ones 70% (with the exception of Wholesale and Car Repairs), and the high sectors have more skilled than unskilled workers.

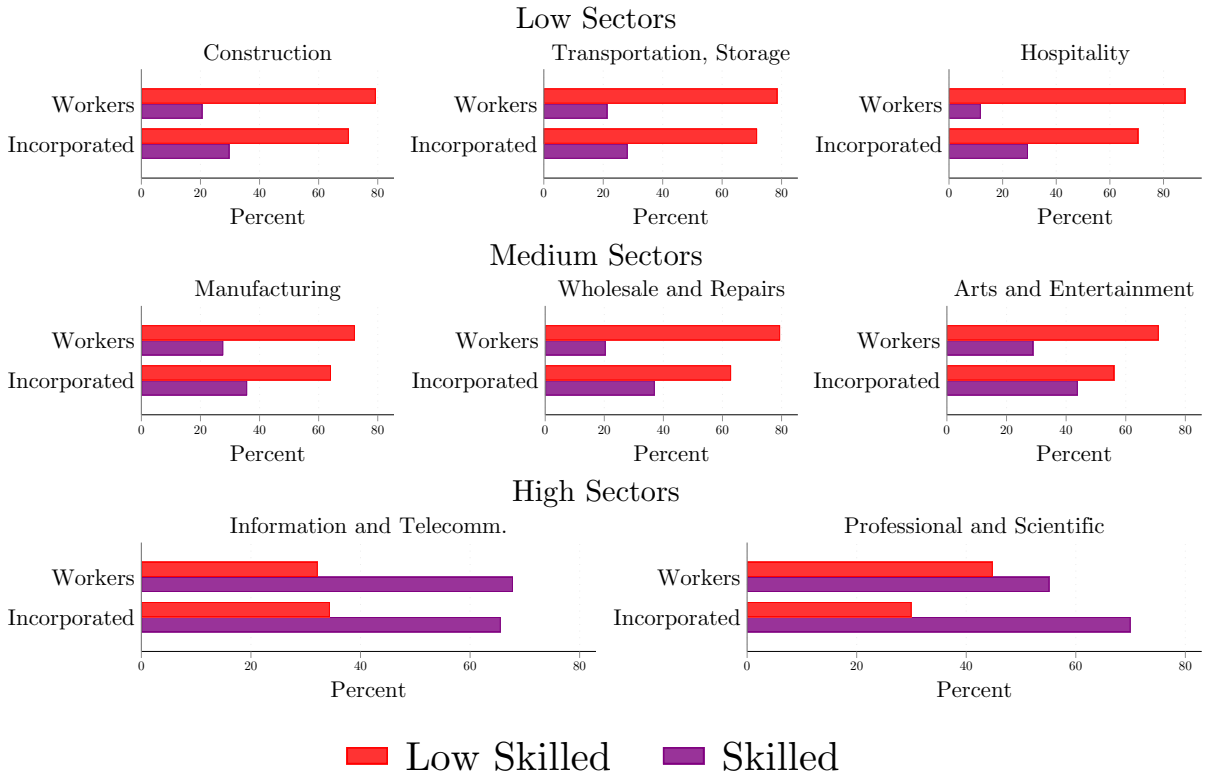
A.2 MPKs by UA-Sector in the Data

A main result from the model is the prediction of heterogeneous returns of capital across space. In particular, the model predicts that those UAs that enjoy higher productivity have, ceteris paribus, higher MPK Premiums.

The key question is whether there is evidence from the data that supports this prediction. Figure 8 pointed toward this direction by plotting the obtained (rescaled) MPK in the data for the high sector. In this section we shall discuss how these estimates are obtained and whether it holds for the rest of the sectors.

[63] More details on this classification will be provided in the discussion of the calibration.

Figure A.2: Skill Heterogeneity by Occupation and Sector

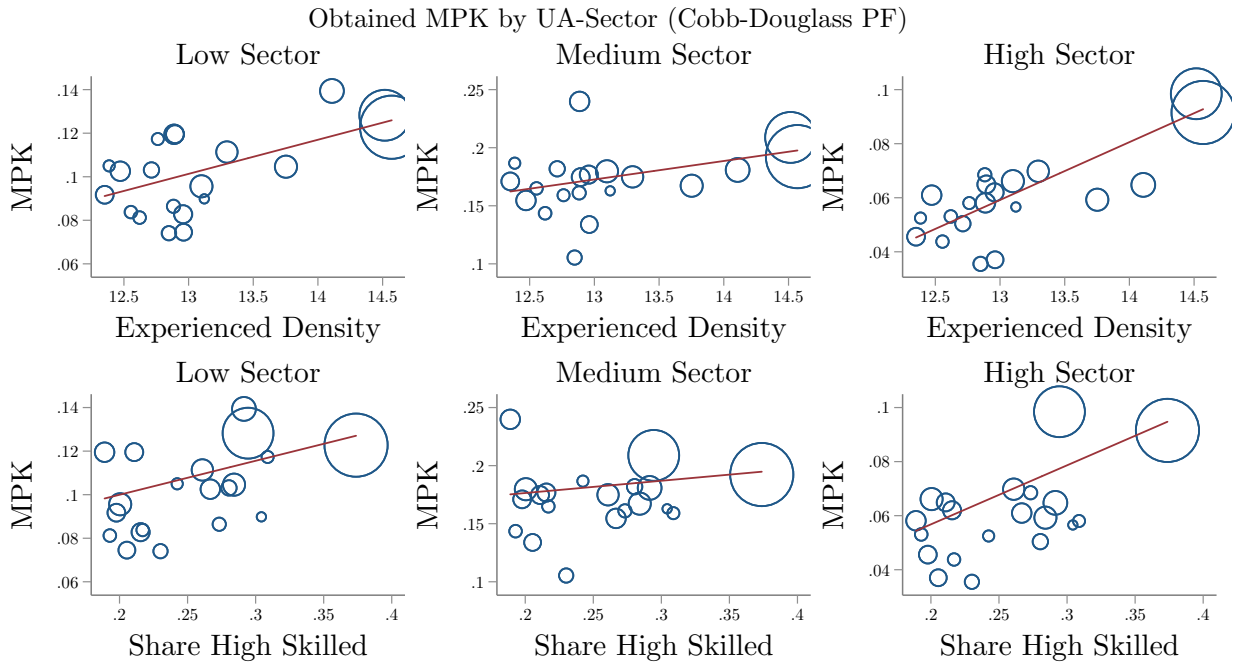


Notes: Joint distribution of skill category (low skilled and skilled) and compressed sectors of the economy by occupation (workers and incorporated). Each row represents one compressed sector of the economy in increasing order of skill (Low, Medium, High). Each panel shows the percentage of workers that are low skilled in the first bar, the percentage of workers that are highly skilled in the second bar, and the percentage of incorporated that are low and highly skilled in the third and fourth bars, respectively. At each sector, bars by occupation sum to 100%.

The spirit of the exercise is simple: a Cobb-Douglass production function is estimated at the sector level and the obtained MPKs are plotted against the size and skill of the UAs. The procedure employs data on the small set of firms from SABI for the period 2013-2018. Capital cost is constructed as $EBITDA - EBIT$ and the labour expenditure is captured by the wagebill. One can then recover $\frac{\alpha_j}{1-\alpha_j} = \frac{rK}{wL}$. Exogenoous productivities are backed out as $A_{l,j} = \frac{Y_{l,j}}{K_j^{\alpha} L^{(1-\alpha_j)}}$. Lastly, the MPK is computed as $A_{l,j} K^{(\alpha_j - 1)} L^{(1-\alpha_j)}$. Figure A.3 presents the obtained results as a function of UA size and skill.

The obtained results are in line with the main prediction of the model: those UAs that are more productive have higher MPKs, regardless of sector or proxy. It is important to note however that

Figure A.3: Estimated MPKs in the Data (SABI Small Firms, 2013-2018)



Notes: Estimated MPKs in the data by UA-Sector by assuming a CB production function at the sector level. Columns indicate the sectors while rows indicate the proxies for the agglomeration forces. The circles indicate the size of the UAs.

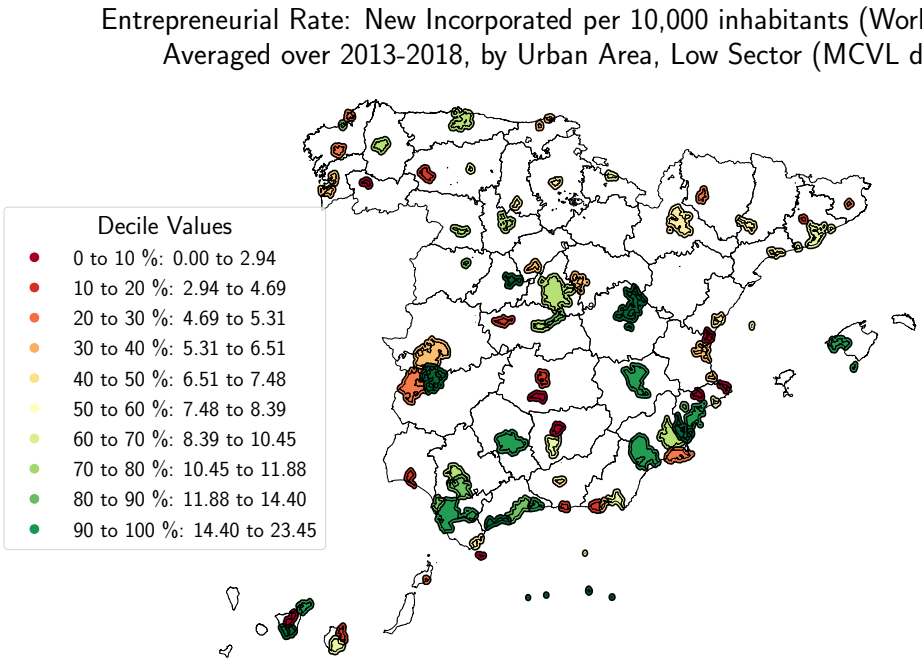
while the model explicitly reports the MPK Premium per UA-Sector, the data counterpart is the estimated CB MPK. This is mainly due to the lack of a proper counterpart to the model's user cost of capital at the required level of granularity. This notwithstanding, the main takeaway from this exercise would be that qualitatively the data suggests a correlation between city productivity and MPKs.

B Appendix Stylized Facts

B.1 Spatial Distribution of Entrepreneurship by Sector

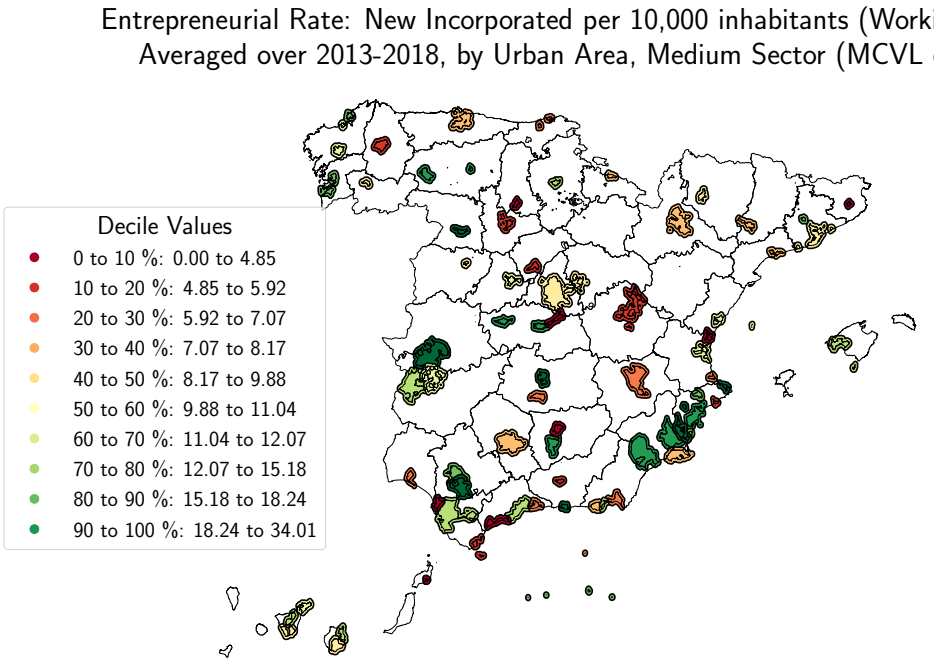
In the introduction we discussed the marked spatial variation of entrepreneurship in the Spanish State. Breaking down the distribution by sector further amplifies the spatial dimension of entrepreneurship. Figures B.4, B.5 and B.6 show precisely this.

Figure B.4: Spatial Distribution of Incorporated Entrepreneurship into the Low Sectors (Construction, Transportation and Hospitality)



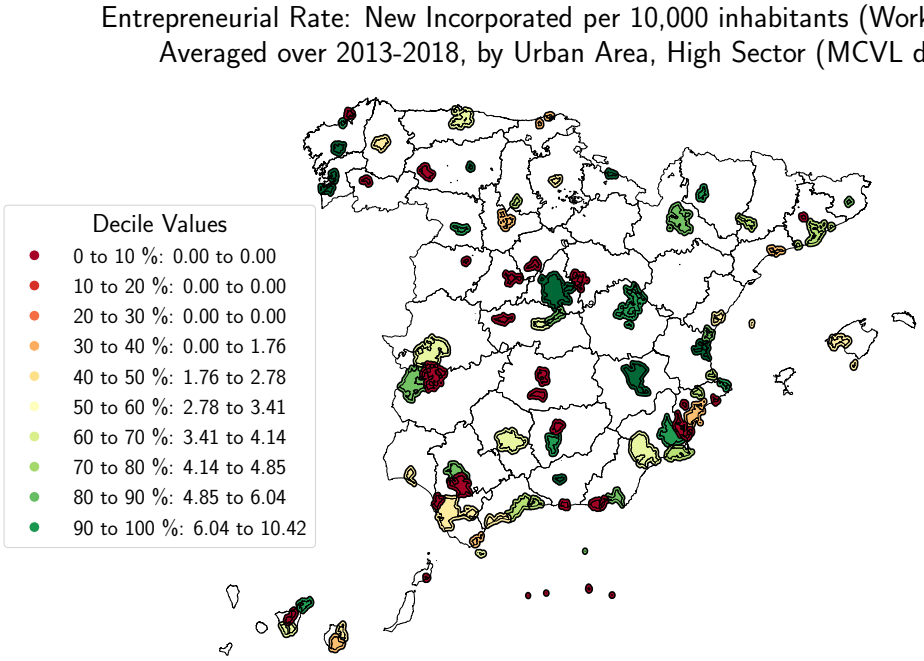
Notes: Spatial distribution of incorporated entrepreneurship into the low skilled sectors of the economy (Construction, Transportation and Hospitality). The rate is defined as those transitioning from workers in year $t-1$ to those starting an incorporated business in year t over 10,000 working-age inhabitants per Urban Area over the period 2013-2018. The colors correspond to a given decile, whose values are provided on the legend to the left.

Figure B.5: Spatial Distribution of Incorporated Entrepreneurship into the Medium Sectors (Manufacturing, Wholesale and Entertainment)



Notes: Spatial distribution of incorporated entrepreneurship into the medium skilled sectors of the economy (Manufacturing, Wholesale and Entertainment). The rate is defined as those transitioning from workers in year $t-1$ to those starting an incorporated business in year t over 10,000 working-age inhabitants per Urban Area over the period 2013-2018. The colors correspond to a given decile, whose values are provided on the legend to the left.

Figure B.6: Spatial Distribution of Incorporated Entrepreneurship into the High Sectors (Telecomm., Professional and Scientific)

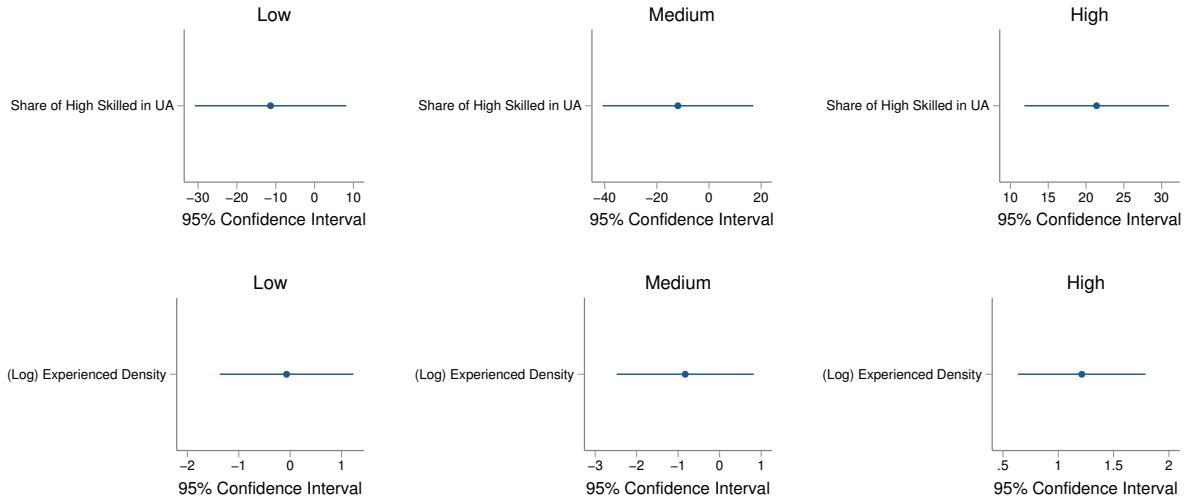


Notes: Spatial distribution of incorporated entrepreneurship into the high skilled sectors of the economy (Telecomm., Professional and Scientific). The rate is defined as those transitioning from workers in year $t-1$ to those starting an incorporated business in year t over 10,000 working-age inhabitants per Urban Area over the period 2013-2018. The colors correspond to a given decile, whose values are provided on the legend to the left.

B.2 Additional Results on the Stylized Facts

B.2.1 Estimated Coefficients in Stylized Fact #2

Figure B.7: Estimated Coefficients for the Correlations in Stylized Fact # 2



Notes: Estimated coefficients for the correlations in stylized fact #2. Each column corresponds to one sector of the economy and each row corresponds to one of the agglomeration forces proxies. The x-axis denotes the 95% CI.

As argued in the main paper, the high sector is the most sensitive to the proxies for the local agglomeration forces and both the size and skill coefficients are positive and significant at the standard significance level.

B.2.2 Empirical Specification, Stylized Fact # 3

The following specification is considered to test the hypothesis:

$$\begin{aligned}\mathbf{1}_{i,ua,s,t} = \alpha_i + \Theta\,&\text{Lagged Cumulative Income}_{i,ua,s,t} \\ &+ C^{I}_{i,ua,s,t}\Xi + C^{UA}_{i,ua,s,t}\Gamma + d_t + S_s + UA_{ua} + u_{i,ua,s,t} \end{aligned} \tag{2}$$

The spirit of (2) is to regress the entry decision on the LHS on the cumulative income and a set of controls and fixed effects at the individual, sector, year and/or UA levels. Thus, $\mathbf{1}_{i,ua,s,t}$ is an indicator variable that takes a value of 1 if individual i, at UA ua, in sector s and time t

became an entrepreneur and 0 otherwise [64]. α_i is a term that accounts for the individual fixed effects. Θ Lagged Cumulative Income$_{i,ua,s,t}$ is the main term of interest, which captures whether the lagged cumulative income of an individual has any effect Θ on the entry decision. $C'^{I}_{i,ua,s,t}$ are controls at the individual level (Skill, Age, Sex, Family Size, Nationality, Internal Migrant). $C'^{UA}_{i,ua,s,t}$ are controls (Average Skill, Average Income, Experienced Density, Age Composition, % Workers in Large Firms, Unemployment Rate, Share of Foreigners, Labour Herfindahl Index) at the UA Level. Lastly, d_t, S_s, UA_{ua} are the time, sector and UA fixed effects, respectively and $u_{i,ua,s,t}$ is the idiosyncratic error term. A more detailed explanation of the variables is provided in Appendix B.2.3

Table B.3: Estimation of the Effect of Cumulative Income on Entrepreneurial Entry

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Dependent Variable	Pr. of Entry	Pr. of Entry	Pr. of Entry	Pr. of Entry	Pr. of Entry	Pr. of Entry	Pr. of Entry	Pr. of Entry
L.(Log) Cumulative Income	-0.000411***	-0.000295***	0.000929***	0.00103***	0.000835***	0.000925***	0.000818***	0.000921***
	(-5.75)	(-5.64)	(3.70)	(4.31)	(3.19)	(3.66)	(3.19)	(3.68)
Controls	Yes	No	Yes	No	Yes	No	Yes	No
Individual FE	No	No	Yes	Yes	Yes	Yes	Yes	Yes
Urban Area FE	Yes	Yes	Yes	Yes	Yes	Yes	No	No
Year Effects	Yes	Yes	Yes	Yes	No	No	No	No
Sector FE	Yes	Yes	Yes	Yes	No	No	No	No
Sector x Year FE	No	No	No	No	Yes	Yes	No	No
UA x Sector x Year FE	No	No	No	No	No	No	Yes	Yes
R2	0.000943	0.000455	0.375	0.373	0.375	0.373	0.381	0.380
Cluster	UA	UA	UA	UA	UA	UA	UA	UA
N	641376	641397	592645	592671	592645	592671	592640	592666

t statistics in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: Regression results of specification (2). The dependent is an indicator variable on whether individual *i* at UA *ua* in sector *s* and time *t* has become an entrepreneur. Columns (1)-(2) show the results (with and without controls, whilst keeping UA, Year and Sector FE) without individual fixed effects. Columns (3)-(8) show a combination of results (with and without controls, for a different set of FE) that include individual fixed effects. Coefficients are reported with standard errors in parentheses, which are clustered by Urban Area in all columns. ***, ** and * indicate significance at the 1%, 5% and 10% levels, respectively. The reported R2 is the overall one of the regression. The Cumulative Income is computed as the sum of the stream of labour income previous to becoming an entrepreneur from yearly available fiscal data. Controls include those at the individual and UA-level.

[64] Given the persistence in the entrepreneurial pursuit (those that have entered at least once are more likely to enter in the future), data is only kept until the first entrepreneurial episode is observed during the period 2013-2018. This is done in order to avoid the higher entry rate probability of those that have already attempted to enter once.

Table $(B.3)$ provides a set of results for specification (2). Columns (1) and (2) show the estimates when individual FE are not included (whilst keeping UA, Year and Sector FE). In this case, both coefficients on cumulative income are negative and significant. Note however that these two first columns exploit the variation across individuals in a given UA, Year and Sector. This result is in line with the observation in table $(A.2)$ which showed that incorporated entrants have accumulated less income on average than the overall population.

The main observation of interest is that when incorporating individual FE (that is, when exploiting the variation across time within an individual) in columns (3)-(8), the coefficient on cumulative income turns positive, significant and stable regardless of which specific is considered (3)-(8). To provide some context on the results, the 0.000921 coefficient in column (8) would indicate that an increase in the lag (log) cumulative income of 1% would increase the probability of becoming an entrepreneur by 0.00000921 absolute points, which given the unconditional mean of the dependent variable of 0.22% would translate to a 0.39% in relative terms.

To sum up, results from table $B.3$ imply that a 10% in cumulative income leads to a 3.9% increase in the relative probability of becoming an entrepreneur. This result hence motivates the incorporation of financial frictions in the model.

B.2.3 Variable Description, Stylized Fact # 3

This section provides the variable description of those employed in stylized fact #3

- Explanatory variable:
  - Lagged Cumulative Income: Lag of the cumulative income from the year 2013 up to time t taken from the AEAT's fiscal data. It includes all income available in form 190. The lag is employed given that the contemporaneous income of the entrepreneurs (incorporated) is not observable in the data.
- Individual Controls:
  - Individual Skill: this is the skill category as defined in Roca and Puga (2017) based on the occupational categories for the contemporaneous year.

  - Age: age is computed as the difference in years between the date of birth provided by the personal data component of the MCVL and a given year.
  - Sex: the sex variable is taken directly as reported in the personal data section of the MCVL.
  - Family Size: family size is the sum of all members of the household as reported by the relatives section of the MCVL.
  - Nationality: nationality is taken directly from the personal data section of the MCVL and represents the legal status with respect to the state.
  - Internal Migrant: a person is an internal migrant in the contemporaneous period if the UA of residence at time t is different from the one in the previous period.

- UA Controls:
  - Average Skill: average skill at the UA-Year level is the (arithmetic) mean over the skill of all residents at time t in a given Urban Area.
  - Average Income: average income at the UA-Year level is the (arithmetic) mean over the income of all residents at time t in a given Urban Area.
  - Experienced Density: experienced density is the proxy for the sizes agglomeration force and is the avearage amount of population within 10km of the average person by Urban Area. Its construction follows the steps outlined in Duranton and Puga (2020). It is constructed by relying on grid cell population data at a 1km $\times$ 1km resolution from `worldpop.org` along with the geographical data on the delimitations of the Spanish municipalities taken from GIS `https://centrodedescargas.cnig.es/CentroDescargas/catalogo.do`.
  - Age Composition: the age composition is a set of variables indicating the proportion of population in the UA-Year of each of the following groups: 16-25,25-35,35-45,45-55,55-65,>65.
  - % of Workers in Large Firms: this is the % of workers at the UA-Year level that work in a firm with over 500 employees. The firm size data is taken from the social security affiliation episode corresponding to each worker and year.

- Unemployment Rate: the unemployment rate at the UA-Year level is computed as the fraction of people whose longest affiliation episode with the social security at that year has been an unemployment spell where they have received unemployment subsidies.
- Share of foreigners: the share of foreigners at a given UA-Year level is computed as the fraction of people born outside the country from the total population of the UA. The data is taken from the personal data section of the MCVL.
- Labour Herfindahl Index: the HHI index is constructed at the UA-Year level by applying the standard definition to the 1st digit NACE sectors. Data on the sector of the employees comes from their social security records.

C Definition of the Equilibrium

A Stationary Recursive Competitive Equilibrium (SRCE) in the economy consists of a set of value functions $V(\cdot), V^w(\cdot), V^e(\cdot)$, policy functions $occ(\cdot), a'_w(\cdot), a'_e(\cdot), c_w(\cdot), c_e(\cdot), h_{r,w}(\cdot), h_{r,e}(\cdot), k(\cdot), u(\cdot), h(\cdot), \mu^{l,l',o}(\cdot)$, allocations $\{L_l\}_{l=1}^{L}$, $\{L_{s,l}\}_{s=1,l=1}^{S,J}$, $\{LS_{s,j,l}\}_{s=1,j=1,l=1}^{S,J,L}$, $\{H_l\}_{l=1}^{L}$, $\{\tau_{h,l}\}_{l=1}^{L}$, $\{Y_j\}_{j=1}^{J}$, G, K, $\{E_j\}_{j=1}^{J}$, $\{E_{j,l}\}_{j=1,l=1}^{J,L}$ prices $\{w_{s,j,l}\}_{s=1,j=,l=1}^{S,J,L}$, r, $\{p_j\}_{j=1}^{J}$, $\{p_{h,l}\}_{l=1}^{L}$ and a stationary measure λ^* such that:

1. Given prices, the policy functions $occ(\cdot), a'_w(\cdot), a'_e(\cdot), c_w(\cdot), c_e(\cdot), h_{r,w}(\cdot), h_{r,e}(\cdot), k(\cdot), u(\cdot), h(\cdot), \mu^{l,l',o}(\cdot)$ solve the households' problem and $V(\cdot), V^w(\cdot), V^e(\cdot)$ are the associated value functions.

2. Given prices, entrepreneurs employ optimal inputs in production $k(\cdot), u(\cdot), h(\cdot)$ and $y(\cdot)$ and $\pi(\cdot)$ are the associated production and profit functions.

3. Population in each location L_l is endogenous and given by all transitions from other locations to location l (L $EQ.$):

$$L_l = \sum_{k=1}^{L} \mu^{k,l,o}(a,z,s,j,k) L_k$$

   Where $\mu^{k,l,o}$ is the migration probability of an agent in location k with occupation o to location l at some point on the state space:

$$\mu^{k,l,o}(a,z,s,j,l) = \frac{\log\left(\sum_{l=0}^{L-1} \exp\left(\beta \mathbb{E}V(a',z',s',j',l') - \beta\tau_{k,l,o}\right)^{\frac{1}{v}}\right)}{\sum_{l=0}^{L-1} \log\left(\sum_{l=0}^{L-1} \exp\left(\beta \mathbb{E}V(a',z',s',j',l') - \beta\tau_{k,l,o}\right)^{\frac{1}{v}}\right)}$$

4. The interest rate r clears the country-wide supply and demand for capital (1 $EQ.$):

$$r \; s.t. \int a'(a,z,s,j,l) d\lambda(a,z,s,j,l) = \int I_{occ(a,z,s,j,l)} k(a,z,s,j,l) d\lambda(a,z,s,j,l)$$

5. The housing sector clears (L $EQ.$): $H_l^d \equiv \int\int\int\int h_r(a,z,s,j,l)\Lambda'(a,z,s,j,l) dadzdsdj = p_{h,l}^{\eta} \bar{H}_l$ for each l

6. Local councils collect the profits made by the housing sector and rebate them back to households as income tax reductions $\tau_{h,l}$ over income I_l (L $EQ.$):

$$\Pi(H_l^d) = \tau_{h,l} I_l \text{ for each } l$$

   Where $\Pi(H_l^d) = \dfrac{1}{1+\eta_l} \dfrac{H_l^{d\,\frac{1+\eta_l}{\eta_l}}}{\bar{H}_l^{\frac{1}{\eta_l}}}$

7. Local labour markets clear ($S \times J \times L$ $EQ.$):

$$w_{l,j,s} \; s.t. \quad \underbrace{LS_{s,j,l}}_{\text{Labour Supply of type \{s,j\} in location l}} = \underbrace{\int_a\int_z\int_s LD_{s,j,l}(a,z,s,j,l) dadzds}_{\text{Labour Demand by Entrepreneurs. in location } l \text{ and sector } j \text{ for a given skill across all } (a,z,s) \text{ triplets}}$$

8. Overall demand for intermediate sectoral goods clears (J $EQ.$): p_j s.t. $Y_j = \left(\frac{\gamma_j P_j^{-\rho} Y}{P^{(1-\rho)}}\right) \forall j \in J$

9. The Government Budget clears (1 $EQ.$):

$$\sum_{l=1}^{L}\sum_{j=1}^{J}\sum_{s=1}^{S} \tau_L w_{l,j,s} LS_{l,j,s} + \int \tau_K \pi(a,z,s,j,l) occ(a,z,s,j,l) d\lambda^*(a,z,s,j,l) = GY$$

10. The aggregate resource constraint of the economy is satisfied and equates production in the economy of the final good to the total expenditure (1 $EQ.$):

$$Y = C + \underbrace{GY}_{\text{Government Expenditure}} + \delta K + \underbrace{\sum_{l=1}^{L} p_{h,l} H_l^d - \Pi(H_l^d)}_{\text{Net Expenditure on Housing}} + \underbrace{F}_{\text{Total Fixed Costs}}$$

11. The invariant probability measure λ^* satisfies:

$\lambda^*(A \times Z \times S \times J \times L) = \int T((a,z,s,j,l), A \times Z \times S \times J \times L) d\lambda^*(a,z,s,j,l) \; \forall A \subset \mathcal{A}, Z \subset \mathcal{Z}, S \subset \mathcal{S}, J \subset \mathcal{J}, L \subset \mathcal{L}$

Where $T(\cdot)$ is the transition function defined as:

$$T((a,z,s,j,l), A \times Z \times S \times J \times L) = I_{a'(a,z,s,j,l) \in A} \sum_{z' \in Z} \pi_z(z,z') I_{s \in S} I_{j \in J} \sum_{l' \in L} \mu^{l,l',o}(a,z,s,j,l)$$

Note that s and j are fixed given the ex-ante heterogenity along these dimensions.

D Computational Appendix

In the main paper we discussed how porting the the solution of the model from MATLAB to lower-level programming languages and to CUDA/C++ in particular could lead to speed-ups in the range of 60 to 20,000, depending on the degree of parallelizability of the algorithm. In this section we first seek to explore deeper what is meant by parallelizability along the computation and memory dimensions and how the economist can identify what in the main paper we referred to "memory bottlenecks". This is crucial as the benefits of this methodology rely on a clear understanding of several basic principles that we will discuss in the following section. Second, we will decompose the speed-up gains by employed intrinsic of the programming language: the Cooperative Groups API for within-kernel thread synchronization [65], a specific construction of the grid to facilitate coalesced memory accesses, [66] FMA (fused-multiply-add) instructions for improved accuracy and reduced instruction overhead, [67] extensive use of fp32 data types to exploit the optimized architecture of the RTX family around this format, [68] the use of the programmable L1 shared memory for reduced memory latency wherever sensible, [69] and lastly the capacity utilization. This is of importance as the obtained differentials underline the complexity of the programming model and help in rationalizing previous conflicting findings in the literature regarding the attainable speed-ups. Third, with this intuition in mind, we will show how applying these optimization techniques to the original Fernández-Villaverde and Valencia (2018) code can increase the attained speed-ups from x2-3 to around x1,600 (with respect to single-threaded C++).

D.1 A Tale of Two Algorithms

D.1.1 Case 1: Value Function Iteration

As argued above, the VFI presents itself as an example of an algorithm that is both computation and memory parallelizable.

[65] Check `https://developer.nvidia.com/blog/cooperative-groups/`.

[66] More info at: `https://developer.nvidia.com/blog/how-access-global-memory-efficiently-cuda-c-kernels/`

[67] Section 2.3 of the CUDA documentation `https://docs.nvidia.com/cuda/floating-point/index.html`

[68] A crucial specification of the GPUs are their core count per data type. Different categories of GPUs, mainly the gaming vs workstation branches, have a different configuration on their fp64 to fp32 ratio, which is typically 1:64 in gaming GPUs and 1:2 in workstation GPUs. For further information, check section 16.7. of the CUDA documentation: `https://docs.nvidia.com/cuda/cuda-c-programming-guide/index.html#cuda-enabled-gpus`.

[69] An explanation is provided in `https://developer.nvidia.com/blog/using-shared-memory-cuda-cc/`

For the sake of our discussion, suppose we have constructed the grid for the state space as a Cartesian product on $A \times Z \times S \times J \times L$. That is, if we were to linearize the obtained tensor, we would obtain N_a elements for given values of z, s, j, l, $N_a \times N_z$ elements for given s, j, l and so on. That is, the typical element (a, z, s, j, l) would have index $I_a + N_a \times I_z + N_a \times N_z \times I_s + N_a \times N_z \times N_s \times I_j + N_a \times N_z \times N_s \times N_j \times I_l$, where I_v corresponds to the index of the variable on its univariate grid and N_v is the number of elements per dimension. Clearly, the total number of elements is $N_a \times N_z \times N_s \times N_j \times N_l$, and the scaling of the grid is non-linear (N_d^d if we had equally sized dimensions).

Why is discussing the construction of the grid of importance? Recall how a typical memory access request works: the LD/ST unit of the SM (Streaming Multiprocessor) requests certain data. If it is found in the L1 cache, then we have a L1 cache hit and the data is collected. If not, the L2 is explored. If the data is located on the on-chip L2, we have a L2 hit. If not, we have a L2 cache miss and data is requested to the global memory (VRAM). This is the furthest a memory request can go and latency is added. Once the request is served, not only the desired memory address is delivered but a 32 byte (256 bits) chunk of data, which working with float32 bit types translates into the desired memory address plus the seven next closest elements in memory. Therefore, a sequential and aligned memory access pattern becomes pivotal in order to maximize memory bandwidth (how much memory can be serviced by second). When such is the case, it is said that we have a "coalesced" access pattern.

Why would this matter to the macroeconomist? In macroeconomic terms, let's simplify the problem and suppose that we are in the Aiyagari (1994) economy. Is there any reason to prefer $A \times Z$ over $Z \times A$? There is. It all boils down to the effect on the policy function of a marginal change from one point on the state space i to the next, $i + 1$. In the $A \times Z$ world, the change for index 0 to index 1 is that assets will increase to the next level, while keeping productivity fixed. In the $Z \times A$ world, this change is a change from the lowest productivity to the second lowest productivity. If we were maximizing for the policy function, which would be closer to a coalesced memory access pattern?

To illustrate this point, consider these two optimization problems:

$$V(a,z) = \max_{a'} u(\cdot) + \beta \mathbb{E}_{z'} V(a', z') \qquad V(z,a) = \max_{a'} u(\cdot) + \beta \mathbb{E}_{z'} V(z', a')$$

They are mathematically equivalent, yet computationally there is a major difference. Suppose we were solving for the optimal policy $a'(a,z)$ by golden section search. For a wide enough range of assets, there is a high probability that under both specifications the initial evaluation points would be very similar. However, as we approach the actual solution, the difference in memory access patterns between index i and $i+1$ on the state space will vary dramatically: in the (a,z) world, going from assets $a[i]$ to $a[i+1]$ for given $z[i]$ will likely have a minor effect on $a'(a[i], z[i]) - a'(a[i+1], z[i+1])$ for the typical element where $z[i+1] = z[i]$. However, in the (z,a) world, a jump from $z[i]$ to $z[i+1]$ could be substantial: think of a model where there are unskilled and skilled workers and there is a substantial wage differential. Therefore, the difference in $a'(z[i+1], a[i+1]) - a'(z[i], a[i])$ could be substantial, so that different points of the state space on a' are explored by two contiguous points on the state space, lowering the cache hit ratio and increasing the cycles of the warps being stalled waiting for memory.

To quantify this point, let's profile, using Nvidia's tool Nsight Compute, [70]. how many stall cycles there are precisely when performing the golden section optimization under each grid configuration:

Table D.4 clearly shows the trade-offs between the two choices for representing the state space. When performing the optimization process, in line with the intuition we developed above, in the case where the state space is $A \times Z$ there are 58,148 cycles where the instruction could not be executed due to threads waiting for memory. In turn, in the $Z \times A$ this number skyrockets to 1,083,295, a 1762.99% increase. Moreover, 84.94% of the performed memory requests are overhead due to strided or non-aligned memory access patterns or cache misses.

However, there is a trade-off to the $A \times Z$ choice. While the expectation operation is relatively coalesced in this case (index 0 needs to access N_a and index 1 needs to access $N_a + 1$ to compute the expectation w.r.t z_1), there is greater data locality when taking the expectation over z' in the case where grid set up is $Z \times A$ (in this case, both index 0 and 1 would need to access index 1,

[70] For an example on using the software, check `https://www.nvidia.com/en-us/on-demand/session/other2024-compnsight/`

Table D.4: Profiling Results of the VFI Algorithm in `https://github.com/markoirisarri/AiyagariModelCUDA` using Nvidia's Nsight Computing

	Optimization: Golden Section Search		Expectation $\mathbb{E}'_z V(\cdot)$		Execution Time
State-space	Memory Stall Cycles	% Excessive Requests	Memory Stall Cycles	% Excessive Requests	
$A \times Z$	58,148	22.8%	699,583	20%	172.63ms
$Z \times A$	1,083,295	84.94%	188,161	Below reporting threshold	189.15ms

Notes: Profiling results for the VFI algorithm in the Aiyagari (1994) model. Memory Stall Cycles represents how many cycles are spent by the processors waiting for the memory request to be serviced by the global memory (VRAM). % Excessive Requests indicates the proportion of memory requests that exceed those predicted by the profiler. This occurs when memory accesses are non-coalesced (i.e., memory accesses are not aligned, leading to inefficient use of the memory bus) or when there are cache misses due to the large problem size exceeding the L2 cache capacity. Overall timing of the VFI refers to the overall time it takes the kernel (function run on the GPU) to execute as reported by Nsight Compute. The dimensionality is $N_a = 100,000$ and $N_z = 10$. We intentionally keep the dimensionality high so that the L2 memory can not service all requests. Note that the algorithm employs Howard Improvement and the stall samples for optimization are collected at the optimization rounds while the ones for the expectation are recorded at every iteration.

which avoids a stride in global memory).

Ultimately, the difference in execution time between these two representations is about %10. Despite the seemingly low figure, it is important to note that the specific model at hand might guide the researcher in one direction or another. For example, in a model of occupational choice where K occupations perform their own optimizations the speed-ups of employing $A \times Z$ could be substantial. Back to our original $A \times Z \times S \times J \times L$, the macroeconomist needs to weigh the pros and cons of each potential choice for the representation of the state space. In our case, the choice for this specific order was motivated precisely by this discussion. The choice for A as the first dimension is motivated by the occupational types requiring their own optimization. The choice for Z as the second dimension is to retain as much data locality as possible when performing the expectation. Lastly, the choice for L as the last dimension allows to keep the location dimension constant per every $N_a \times N_z \times N_s \times N_j$ elements.

To sum up, we began this section arguing that the VFI is a "computation parallel, memory quasi-parallel" algorithm. The computationally parallel part was already clear: each point of the state space performs its own independent computations [71] . Regarding memory, we discussed how setting the grid up in a certain way can alleviate the memory read interdependencies that

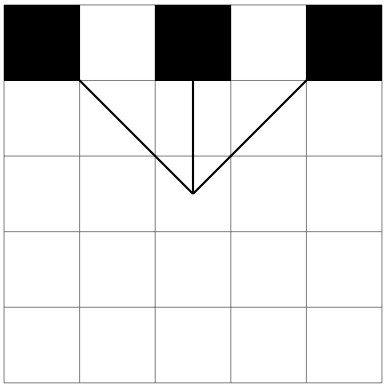
[71] An example of a non-computation parallel algorithm would be a panel simulation with multiple periods and a single agent. One would ideally like to parallelize along the time dimension, but the inter-temporal dependence does not allow this.

exist among threads so that about 20% of the accesses are excessive.

D.1.2 Non-Stochastic Simulation (Young (2010)'s Method)

The previous section discussed a common problem faced when solving models with GPUs: the fundamental optimization technique to have a coalesced memory access pattern (to the extent that it is possible). The aim of this section is to further discuss the negative effects on the attainable speed-ups that stem from memory bottlenecks, this time by looking at multiple threads (indexes) that want to read and write from and to the same location in memory. This is of relevance since it is precisely what is going on under the hood when solving macroeconomic models via Young (2010)'s method. To illustrate this point, figure D.8 provides an example of such a case:

Figure D.8: Example of a Multiple Memory Read and Write Access Pattern to the Same Address by Multiple Threads



Notes: The figure mimicks the behaviour of memory access patterns in solving for the invariant distribution using non-stochastic simulation. In this example, indexes (0,0), (0,2) and (0,4) are requesting to write their values to index (2,2).

In this example, there are multiple indexes ((0,0), (0,2), (0,4)) that aim to write their values to the same index ((2,2)). This requires first reading the memory address data and then writing the value of the requesting index on that value. This is the perfect example of a horse-race condition: without any kind of control, an asynchronous execution model (GPUs) would result in the fastest index reading and writing to the address, without any guarantees that the rest of the indexes have the updated value when they request it or that they will wait for the fastest index to be done writing the result to write their own. Not accounting for this necessity to synchronize the

memory read and writes can lead to catastrophic consequences D.5 [72]. Note that a sequential single-threaded implementation would suffer from no such problem since the memory access pattern is deterministic and synchronous, index $i + 1$ does not execute until i is done reading and writing the data.

Fortunately, CUDA counts with a set of instructions that impose this memory synchronization: the atomic operations. This set of instructions ensure that concurrent accesses to memory are safely coordinated, thus avoiding conflicts. For example, atomicAdd((2,2), (0,0)) ensures that no one other thread can access (2,2) until (0,0) is done adding its data to it. The notable downside to this is that it leads to a bottleneck in the memory pipeline, to the point where even the scheduler submitting the memory instruction is overloaded with such requests. If we run the Nsight Compute profiler again, 99.1% of the time instruction are waiting to be executed is due to unserviced global memory requests.

What would be the consequences of employing a non-safe "cumulative sum" $+=$ against the recommended atomicAdd()? Table D.5 provides the results under both scenarios:

Table D.5: Numerical Implications of non-safe Memory Operations in CUDA/C++

Safe Memory	Mass of Agents	K	Share Entrepreneurs	SWF	Share People in Madrid
No: $+=$	$1.07e-4$	$4.6e-3$	$3.54e-6$	$4.6e-4$	$1.46e-5$
Yes: atomicAdd()	1	48.9	3.45%	-3.52	33.58%

Notes: Computed aggregates of the economy under memory safe operations (atomicAdd()) and without ($+=$ operator). Safe memory operations means that no thread can access one memory address until threads that are currently writing to that address have finalized writing. Mass of agents denotes the normalized mass of agents in the economy. K is aggregate capital. Share of Entrepreneurs is the total overall share of entrepreneurs in the economy. SWF is hte utilitarian Social Welfare Function. Share of People in Madrid is the percentage of people in the model living in Madrid.

Clearly, ensuring safe memory operations is a key aspect that the macroeconomist thinking on porting the code to native CUDA should be aware of.

To close this brief section on Young (2010)'s method, we discussed how a seemingly parallel algorithm might not enjoy the same degree of speed-ups as the VFI did. In this case, we discussed how the effective serialization of memory access patterns does not allow the researcher to tap into the computational capabilities of GPUs.

[72] Note that this is not a problem specific to GPUs. For example, a multi-core implementation on a CPU would also suffer from the same horse-race conditions.

D.2 Decomposing the Speed-up Gains

As mentioned in the main body of paper, special care was placed on following the guidelines and best practices suggested by Nvidia engineers in Guide (2020) and Guide (2020). In particular, we argued how proper use of memory coalescing, choice of certain data types (in particular fp32 against fp64), employing FMA instructions and the use of the cooperative groups API was pivotal in achieving substantial speed-ups. The aim of this section is to decompose the speed-up gains by each of the employed language intrinsics. This is of importance as it can rationalize previous differing figures in the literature on the attainable speed-ups.

To fix ideas, a recap of the role played by each of these intrinsics is in place. We already discussed memory coalescing in D.1.1, the notion of keeping memory access requests sequential and aligned. Second, another crucial consideration is the choice between fp32 and fp64 data types. In a typical CPU implementation (say, a vectorized implementation in MATLAB), the choice of data types is not a critical performance consideration since all the cores on the CPU are able to process float and double data types. On GPUs, however, the data type choice is pivotal. Contrary to CPUs, GPUs have specific amounts of execution pipelines per data type, and these are specific to each generation and class of family. Critically, the main grouping is that of gaming cards and workstation cards. The former are designed to maximize the pixel fill rate, while the latter are designed around ensuring safe memory (ECC, Error-Correcting Code) and offering a higher precision. To service their priorities, the pipelines allocated within each SM (Streaming Multiprocessor) to each data type varies. While the gaming cards have a ratio of 1:64 fp64 to fp32 compute pipelines, the workstation ones have a 1:2 ratio. This has profound implications for the economist aiming to solve a model using GPUs, as one could expect nearly a two orders of magnitude performance differential based on data types alone.

Second is the use of the fast math compiler option, which has two major effects. First, it prioritizes FMA (fused multiply and add, $ab + c$) operations whenever possible. This reduces the instruction count by combining multiplication and addition into a single operation. It also improves precision by avoiding an additional rounding step: while the standard implementation does $rn(rn(ab) + c))$, FMA rounds once $rn(ab + c)$. Second, it replaces expensive math functions such as $pow(), log(), exp()$ by faster, hardware optimized counterparts. While this comes at

a precision cost, the speed-up gains can be substantial. As was the case with the choice of the data types, the macroeconomist must thoroughly test that the obtained solutions with this optimizations are within the tolerable precision margins (for example, by testing the obtained results against a double data type implementation without further optimizations).

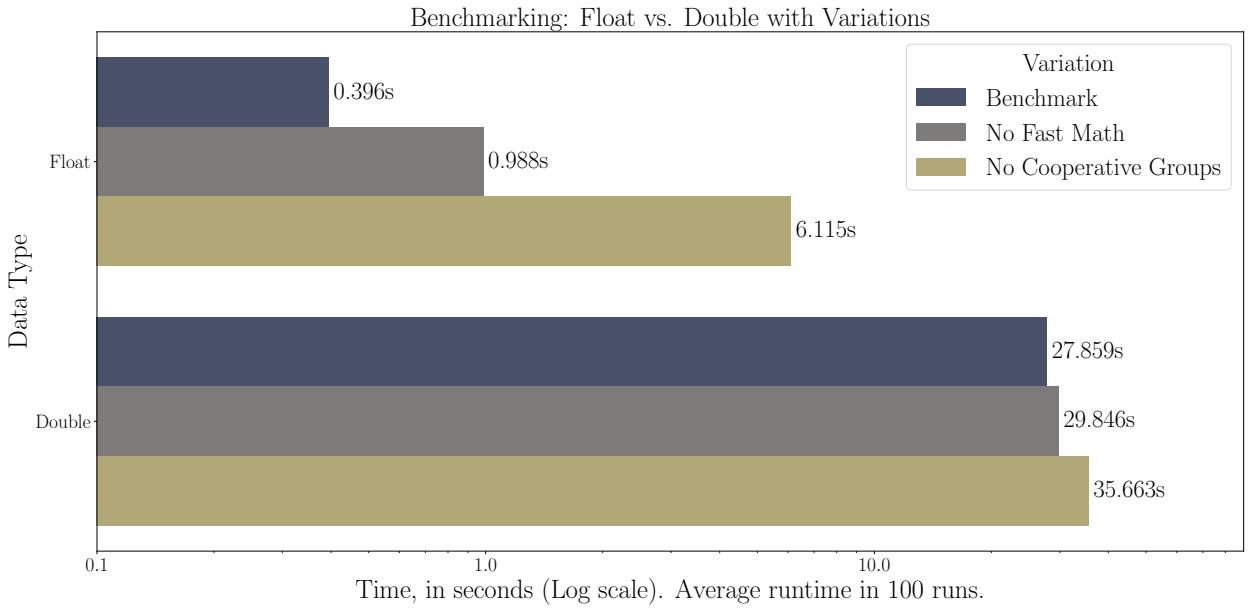
Lastly, the use of the Cooperative Groups API is a key innovation introduced in CUDA 9.0 (2018), particularly crucial for iterative algorithms like those used to solve DSGE models (e.g., VFI, Young 2010's method, PFI, EGM). The primary advantage of this API is the ability to synchronize threads within a kernel. Recall, as we discussed in D.1.2, that GPU execution is inherently asynchronous, meaning that threads run independently without waiting for others to complete their tasks. This leads to race conditions. For example, in VFI, thread i might advance to iteration $k+1$ while thread $i-1$ is still writing its result for iteration k to global memory. As a result, thread i might read an outdated value from thread $i-1$ (when computing the policy function for example), leading to errors. Cooperative Groups solves this issue by allowing threads to synchronize within the kernel by imposing synchronization barriers. In the VFI example, this ensures that no thread starts computing the expectation over entrepreneurial productivity shocks until all threads have finished computing their value and policy functions for the current iteration. Beyond addressing race conditions, this API also simplifies code structure. Before Cooperative Groups, the only way to ensure synchronization across iterations was to split the VFI into multiple blocks, each with their own function, thus forcing the CPU to manage synchronization. This resulted in significant overhead from frequent memory transfers between the GPU and CPU over the PCI-E bus, slowing down performance. With Cooperative Groups, the entire VFI algorithm can run within a single kernel using synchronization barriers, eliminating unnecessary host calls and reducing memory transfer costs. [73]

With the intuition behind each intrinsic in mind, we can now discuss their implications for performance. Figure D.9 provides data on the execution times of the VFI algorithm for this spatial model under different configurations over these intrinsics:

The first striking result is the variability in execution times within GPU computation based on

[73] In essence, this means that instead of splitting the VFI into two separate functions (one for the policy function computation and another for the expectation computation) and relying on the CPU (host) to manage synchronization through function calls, the entire process, including all K iterations, can now be handled entirely within the GPU kernel. This eliminates the need for costly back and forth communication between the CPU and GPU, significantly improving performance.

Figure D.9: Decomposition of the Speed-ups by Language Intrinsic, VFI Algorithm.



Notes: Decomposition of the speed-ups in CUDA/C++ by use of language intrinsics. The y-axis shows the employed data type, either fp32 (float, top) or fp64 (doubles, bottom). The x-axis denotes the time (in seconds, log scale) required by each variant. The reported time is the average execution time over 100 runs. Within each data type, the variants considered are: benchmark (fast math + cooperative groups), no fast math and no cooperative groups. The algorithm under consideration is the VFI. The employed dimensionality is $N_a \times N_z \times N_s \times N_j \times N_l = 100 \times 21 \times 2 \times 3 \times 20$ so that the double data type variation can fit in memory (8 GB VRAM). The employed GPU is an RTX 3070 (5888 fp32 pipelines).

which intrinsics are chosen. The difference in execution time between the fastest (first row, float + fast math + cooperative groups) and slowest (bottom row, double + no cooperative groups) is a whopping $\times 90.05$ *slowdown* in performance. And, importantly, this exercise is only considering a set of three intrinsics for a given implementation. Depending on the specific algorithm, as well as the proper use of other intrinsics such as shared memory, memory coalescing or occupancy could amplify this differential even further. This exemplifies the complexity of the programming model and rationalizes why there have been conflicting results on the attainable speed-ups in the literature (see the speed-ups results in Aldrich (2014) against those in Fernández-Villaverde and Valencia (2018)).

Shifting the focus to the specifics, we can see that relative to the benchmark configuration, which requires 0.396 seconds on average to perform 1500 VFI iterations, forgoing the fast math compiler option increases the execution time to 0.988, for a $\times 2.49$ slowdown. This is mainly due to the cost of the additional instructions required to solve the more complex math functions that

the fast math compiler option was replacing.

Second, not employing the within kernel thread synchronization enabled by cooperative groups results in a slowdown of $\times 15.44$, for a total execution time of 6.115 seconds and already a greater than one order of magnitude slowdown. This is mainly due to the overhead created by constant CPU-GPU memory transfers over the PCI-E lane.

Third, and most importantly, jumping now down to the second set of bars, we can see how the choice of data types matters dramatically. Among the tested variations, the data type choice is the single most important speed-up factor at a slowdown cost of $\times 70$, roughly equal to the total number of fp64 relative to fp32 pipelines (recall the 1:64 ratio). [74] Once conditioned on double types, we can see that the variability along the other employed intrinsics is relatively low. Interestingly, the overhead in CPU-GPU data transfers that the use of cooperative groups prevents presents itself as close to a fixed cost (5.71 vs 7.80 additional seconds in fp32 vs fp64 computations).

To sum up, the CUDA/C++ programming model, while powerful, has many layers of complexity and properly choosing the intrinsics that optimize performance for a given desired precision is a crucial task to be faced by any macroeconomist aiming to exploit the speed-up gains.

D.3 Applying the optimizations to Fernández-Villaverde and Valencia (2018)

In the introduction to this computational appendix, we claimed that applying certain optimization techniques could drastically improve the performance of the code. In this section, we will benchmark what speed-up gains are attainable through this set of optimization in the original code by Fernández-Villaverde and Valencia (2018). It is paramount to underline however that the extent to which these optimization techniques can improve performance depend critically on how the original code is written and follows the best practices guidelines in Guide (2020). To illustrate this point of how paramount it is that the original code abides by the best practices, we will also test the implications of breaking the coalesced memory patterns in Fernández-Villaverde and Valencia (2018)'s code (as we did in D.1.1) on the attainable speed-ups.

The spirit of this benchmarking exercise is as follows. First, we clone the github reposi-

[74] An amplifying factor over this pipeline ratio is the additional optimization rounds allowed for by a double data type. For strict convergence criterias, say $1e-6$, the fp32 precision counterpart may stop at $1e-4$ due to its inability to discern numbers any further, while the double counterpart will continue until reaching the desired precision. This underlines the importance of thoroughly testing every single function implemented on GPUs in float precision.

tory available at `https://github.com/davidzarruk/Parallel_Computing` (accessed 22/10/2024), containing the original code by the authors. Second, we time the execution time of the kernel performing the life-cycle VFI under several specifications. First, we keep the dimensionality as in the original code, with the dimensionality of assets at 1500, that of the shocks at 15 and 10 time periods. This is our benchmark specification for the benchmarking. Then, we consider the following scenarios:

- Benchmark + Utilization: The original code was executed on a GTX 860M. This implies that the maximum block size is 512. Here, since we have an RTX 3070 with different compute capability, rather than imposing the x-dimension of the grid at 30, we let cudaOccupancyMaxPotentialBlockSize() compute the optimal occupancy. This ensures that each SM has as many resident warps/threads as possible.

- Float: recall from our previous discussion that the GTX and RTX families of graphics cards are optimized around fp32 computation, as their primary use case is gaming. This implies that the available number of execution pipelines is far greater in fp32 precision than fp64. In particular, the GTX 860M has a 1:32 fp64 to fp32 ratio while the RTX 3070 has a 1:64 ratio. Utilizing floats rather than doubles therefore massively increases the utilization of both GPUs.

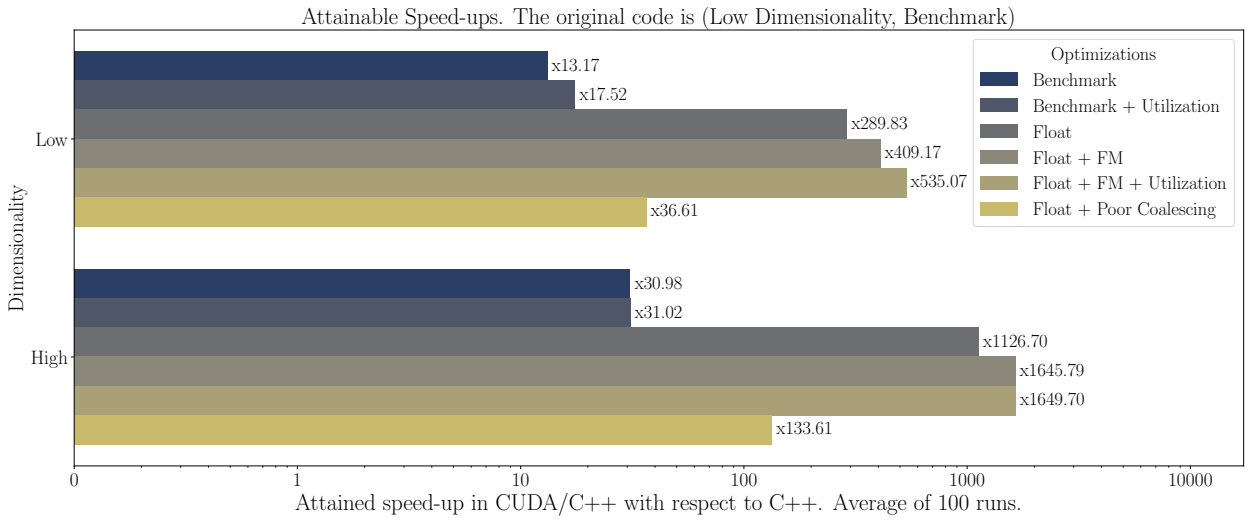
- Float + FM: this configuration employs float data types and the -use_fasth_math NVCC compiler option. We discussed the implications and precautions to be had when enabling this option in the previous section.

- Float + FM + Utilization: in this case we employ fp32 data types, fasth math instructions and optimize utilization by the same method described above.

- Float + Poor Coalescing: in this case we employ fp32 data types and we tamper with the code in order to have the grid set up as $X \times E$ rather than $E \times X$. We discussed the importance of memory coalescing in D.1.1.

Having described each of the scenarios under consideration, figure D.10 presents the obtained results:

Figure D.10: Attainable speed-ups in the original Fernández-Villaverde and Valencia (2018)'s code, by employed optimizations and dimensionality (CUDA/C++ vs single-threaded C++).



Notes: Attainble speed-ups in the original code by Fernández-Villaverde and Valencia (2018), by employed optimizations and dimensionality of the problem. The y-axis shows the dimensionality under consideration. Low refers to the original dimensions of 1500 for assets, 15 for the shock and 10 for age. In High we increase these to 15000, 16 and we keep the age at 10. The x-axis denotes the attained speed-up in CUDA/C++ relative to the single threaded C++ implementation as the average over 100 runs, by each considered variation. Within each dimension, the variants considered are: benchmark (original code), benchmark + utilization (original code with optimized execution configuration), Float (benchmark code but employing fp32 data types), Float + FM (benchmark code with fp32 data types and fast math), Float + FM + Utilization (previous one + optimized execution configuration) and Float + Poor Coalescing (original code in fp32 data types and tampered coalescing). The algorithm under consideration is a life-cycle VFI. The employed GPU is an RTX 3070 (5888 fp32 pipelines) and the CPU is a 10600K @4.5ghz.

Several comments are in place. First, compared to the original paper, we find a speed-up of x13.17 rather than x2-3 when employing CUDA as opposed to single-threaded C++ in the original code. Note however that this is setup specific as neither our GPU, CPU or Memory are the same as the ones employed by the authors. Therefore, this is the reference point we will employ when discussing our results.

Second, we can see that employing optimal utilization increases the speed-up to x17.52. Since the number of shocks is set to 15 and the dimension of the block in the assets dimension is set to 30, this results in blocks of size 450. Even if we were using fp32 rather than fp64, since the SMs of the RTX 3070 have a maximum number of resident threads of 1536, a maximum of 3 blocks would be allocated to each SM, which would leave them underutilized. Additionally, note that 450 is not divisible by the number of threads per warp, 32, which would further decrease performance as there

would be pipelines that will remain idle. Here, we let cudaOccupancyMaxPotentialBlockSize() decide the optimal block sizes in order to achieve the maximum occupancy.

Third, we can see that the attainable speed-up when employing fp32 data types rather than fp64 jumps quite notably to x289.83. This is due to the fact that employing fp32 data types massively increases the utilization of the GPU as now the 5888 fp32 pipelines operate in contrast to the 92 available for fp64.

Fourth, if we further combine the float data types with fast math instruction and optimal utilization, a speed-up of x535.07 is attained. This builds on the intuition from the previous paragraphs and previous section.

Fifth, and importantly, tampering with the proper memory coalesced access structure in the original code and changing the structure of the grid to $X \times E$ rather than $E \times X$ massively decreases the obtained speed-up from x535.07 to x36.61. This highlights the complexity of the CUDA programming language in and how the software to hardware mapping must always be had in consideration when writing the code.

So far, we achieved a maximum speed-up of x535.07 with respect to C++ when employing CUDA/C++, a substantially higher figure than our initial x13.17 and the x2-3 attained by the original authors. Is this as far as we can go? No. Note that at this low dimensionality, a total of 1500×10 threads are executed in parallel (the age dimension is not parallelizable). In the case of our RTX 3070, which has 5888 fp32 pipelines, this results in only 22500 resident threads out of the maximum 70656. Underutilizing this occupancy decreases performance as while a warp is waiting for global memory (or other stall motives) the scheduler of the SM could launch another resident warp in order to hide latency.

When jumping to the second set of bars under the "High" dimensionality configuration, we can observe that now the maximum speed-up jumps to x1649.70 (3.08 times more than at low dimensionality), which is a comparable ratio to the 70656 / 22500 (x3.14). This stresses the importance in GPU computing of achieving the maximum utilization. Importantly, it is not just about filling the entire 5888 fp32 lanes, but rather maximizing the number of resident warps. This allows for the "latency hiding" mechanism to be exploited. That is, while one warp is stalled waiting for global memory, for instance, the scheduler of the SM can launch another warp at virtually no cost. Note that for any dynamic spatial model this is going to be the relevant case as the dimensionality of

the problem is guaranteed to fully utilize the GPU.

To close this section, we have documented how similar speed-ups $(x1649.70)$ to ours $(> x2000)$ can be achieved when applying certain optimization techniques. However, and more importantly, we would like to emphasize how these sections have served to discuss the complex programming model that CUDA/C++ is, and how the attainable speed-ups of employing it ultimately depend on the choices made by the programmer and the understanding of the software to hardware mapping. For example, we saw how a misunderstanding of a single high-priority optimization, memory coalescing, brings down this speed-up from $x1649.70$ to $x133.61$. Also, note that here we have discussed only a subset of optimizations applied to a particular code for a particular algorithm, and that each implementation requires a careful consideration of the entire set of rich features offered by CUDA, of which we have only had time to discuss a few. Lastly, our results should not be taken as an evaluation of Fernández-Villaverde and Valencia (2018)'s implementation. Given their choices on data types, compiler options and available hardware, the implementation by the original authors may well have been the optimal (or close to the optimal) one.

E Calibration Details

This appendix further discusses the details regarding the calibration of the parameters of interest. Every calibrated parameter and the rationale behind the chosen procedure will be discussed.

Table E.6: Calibrated Parameters in the Model

Object	Num Elements	Internally/Externally	Definition	Target
			State Space	
L	1	Externally	Number of Locations	Top 20 UA by Population
J	1	Externally	Number of Sectors	3 Skill-based Sectors
S	1	Externally	Skill Levels	Low Skilled and High Skilled
D	1	Internally	Dimensionality, $N_a \times N_z \times N_s \times N_j \times N_l$	VRAM Capacity, 8GB ($150 \times 21 \times 2 \times 3 \times 20$)
			Skill-Sector Distribution	
$\Lambda_{s,j}$	$S \times J$	Externally	Exogenous Skill-Sector Distribution	MCVL Dataset
			Households	
$a_{s,l}$	$S \times L$	Internally	Amenities	Population Dist. by Skill
β	1	Internally	Discount Factor	Capital to Output Ratio 3
σ	1	Externally	Intertemporal Elasticity	Fixed at 1.5
α_c	1	Externally	Weight Consumption CB Utility	Fixed at 0.7
			Migration	
$\tau_{l,k,w}$	1	Internally	Utility Moving Costs	Share of movers 0.43% (Workers) (MCVL 2013-2018)
$\tau_{l,k,e}$	1	Internally	Utility Moving Costs	Share of movers 0.32% (Entrepreneurs) (MCVL 2013-2018)
v	1	Externally	Scale Parameter	Follow Giannone et al. (2023)
			Housing Sector	
$\bar{H}_{-0}$	$L-1$	Internally	Exogenous Housing Supply	Housing Cost w.r.t Madrid
$\bar{H}_0$	1	Internally	Exogenous Housing Supply	Cost in Madrid 30% of Avg. Low Skilled Gross Salary
η_l	L	Externally	Housing Supply Elasticity	Follow Saiz (2010)
			Production	
F_j	J	Internally	Fixed Cost Production	Target Stock of Entrepreneurs by Sector
λ_j	J	Internally	Leverage Ratio	Target Avg. Revenue-Weighted Debt-to-Assets Ratio
$\Phi_{s,j}$	J	Internally	Multiplier High Skilled	Match share High Skilled Entre. by Sector
ϕ_j	J	Internally	Productivity Scaler	Match profit rate by sector, SABI data
$A_{j,l}$	$J \times L$	Internally	Productivity UA-Sector	Target production distribution across UA-Sectors
$\alpha_{j,i}$	$3 \times J$	Externally	Input Shares CES	Estimate from FOC in the Data
$\Omega_{\text{SKILL},j}$&$\Omega_{\text{SIZE},J}$	$2 \times J$	Externally	Agglomeration Forces	Employ FOC w.r.t. H labour
γ_j	J	Internally	Weight of Sector in GDP	Official Statistics
ϵ_y	1	Externally	Elasticity of Subs. Production	1.5
ϵ_Y	1	Externally	Elasticity of Subs. Intermediate Goods	1.5
μ_j	J	Internally	DRS by Sector	1 - profit share SABI data
δ	1	Externally	Depreciation Rate	Follow Standard 8%
			Entrepreneurial Process	
ρ_z	1	Internally	Autocorrelation	Generate 6.6% Entry Rate
σ_z	1	Internally	Volatility	Avg. K Weighted MPK Premium 1.5%
			Government Block	
$\tau_{l,L}$	L	Externally	Labour Tax	30% Marginal Rate at Avg. Salary
$\tau_{l,K}$	L	Externally	Profit Tax	15% Effective Corporate Tax Rate
G	1	Internally	Share of Government Expenditure	SS value coming from Labour and Profit taxes

Notes: The table displays the calibrated parameters in the model. The first column denotes the calibrated set of parameters. The second column indicates the cardinality thereof. The third column clarifies whether they have been calibrated externally (either estimated using data or fixed to some common value in the literature) or internally (calibrated within the model to match the targeted moments via SMM). The fourth column provides a definition of the set of parameters. Lastly, the fifth column states the target moment in the data.

E.1 State-space

E.1.1 Set of Locations L

The chosen set of locations are the biggest 20 Urban Areas of the Spanish State by population. These UAs concentrate 49% of the Spanish population and about 74% of the economic activity. Below this threshold, the share of observations we have on the smaller UAs falls below 0.5% of the sample. At the same time, a choice of 20 guarantees that the dimensionality of the problem fits the memory constraints of our GPU (8GB of VRAM) so that we can still provide an accurate solution.

E.1.2 Number of Sectors j

Due to the heterogeneous sectoral spatial distribution and their sensitivity to the agglomeration forces, multiple sectors are present in the model. A choice of three different sectors is followed, as a compromise between tractability and the ability to capture this underlying heterogeneity.

The sectors are constructed based on the skill distribution of both workers and incorporated entrepreneurs in the 1st NACE sectors present after the data cleaning process. Accordingly, we classify these as "low skilled sectors", "medium skilled sectors" and "high skilled sectors".

Table E.7: Re-classification of the 1st Digit CNAE 2009 Sectors in the Model, based on Skill Categories

1st Digit CNAE 2009	Model
Construction	Low
Hospitality	Low
Transportation	Low
Manufacturing	Medium
Wholesale & Car Repairs	Medium
Arts & Enterteinment	Medium
Telecommunications	High
Professional Services	High

It is worthwhile noting that during the period under investigation (2013-2018), 86.96% of individuals did not switch the sector of their main activity. This is what motivates our choice in the model to have fixed sectoral types. Figure $(A.2)$ on the main text provides the details on the skill

distribution of incorporated entrepreneurs and workers across sectors.

E.1.3 Number of Skills s

The rationale behind the chosen two levels of skill is the same as with the sectors: a compromise needs to be found between tractability and sectoral and occupational composition.

Given the limited data quality of the education variable in the MCVL (as argued by the documentation of the dataset itself), Roca and Puga (2017) is followed and the skill category is constructed from the occupational categories of the individuals. The obtained estimates are presented in the following table:

Table E.8: Model Classification of the Skill Categories "Low" and "High" based on the Occupational Categories in the MCVL Dataset during the period 2013-2018.

Original MCVL Occupation Classification	Model Skill Level	Unconditional Overall Share
"Engineers, college graduates and senior manager"	"High Skilled"	12.53%
"Technical engineers and graduate assistants"	"High Skilled"	5.15%
"Administrative and technical managers"	"High Skilled"	7.38%
"Non-graduate assistants"	"Low Skilled"	8.96%
"Administrative officers"	"Low Skilled"	19.83%
"Subordinates"	"Low Skilled"	6.73%
"Administrative assistants"	"Low Skilled"	12.63%
"First and second class officers"	"Low Skilled"	17.41%
"Third class officers and technicians"	"Low Skilled"	6.22%
"Labourers"	"Low-Skilled"	3.15%

Overall, the share of high-skilled in the Spanish economy is around 25%, corresponding to the three uppermost categories. The remaining 75% are classified as low skilled following Roca and Puga (2017).

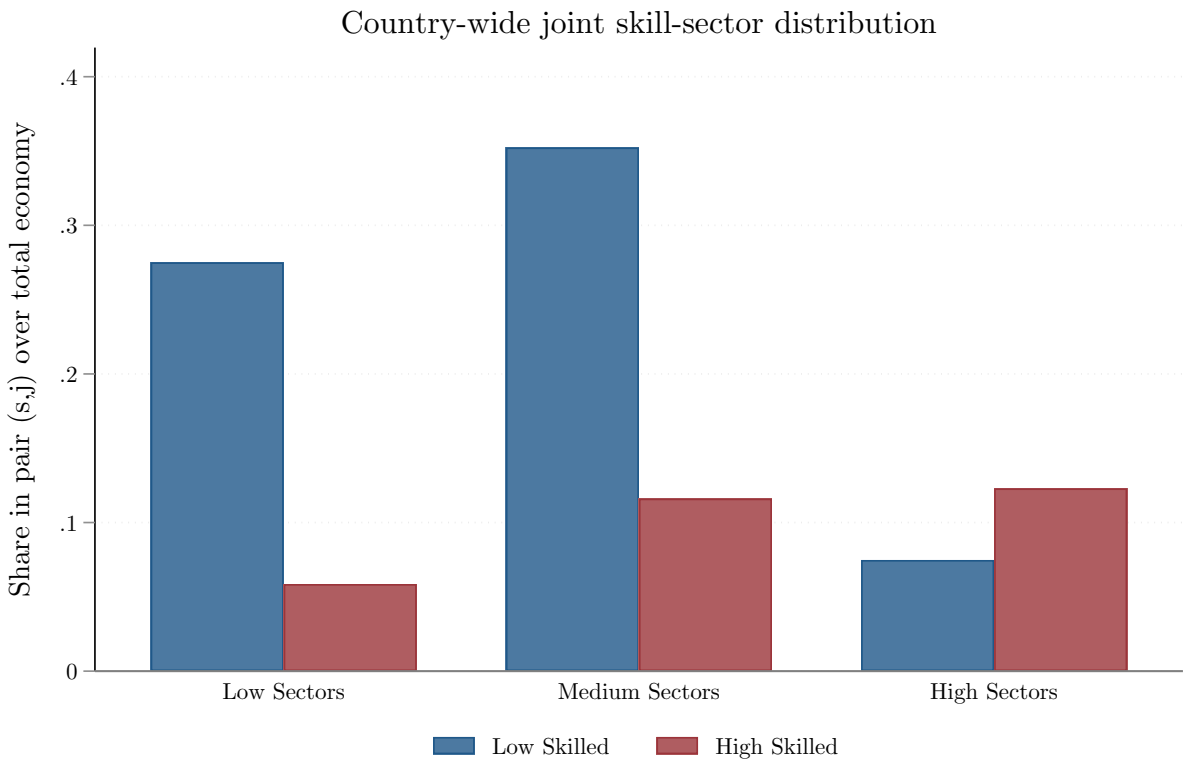
E.2 Households

E.2.1 Exogenous Distribution over Skills and Sectors $\Lambda_{s,j}$

One of the key assumptions regarding the household side was that the sectoral j and skill s types were fixed upon birth. This in turn requires data on the joint skill and sector distribution in the data. Figure $(E.11)$ shows the joint distribution of skill and sector in the economy.

As can be observed, the relative supply of low skilled to skilled varies substantially by sector. While the low sectors have about 5 unskilled workers per every skilled worker, in the high sectors

Figure E.11: Joint Skill-sector Distribution



Notes: Joint skill-sector distribution in the Spanish State's economy. Each bar represents a (s,j) combination. Blue bars denote the share of low skilled in a given sector, while the red bars do likewise for the skilled.

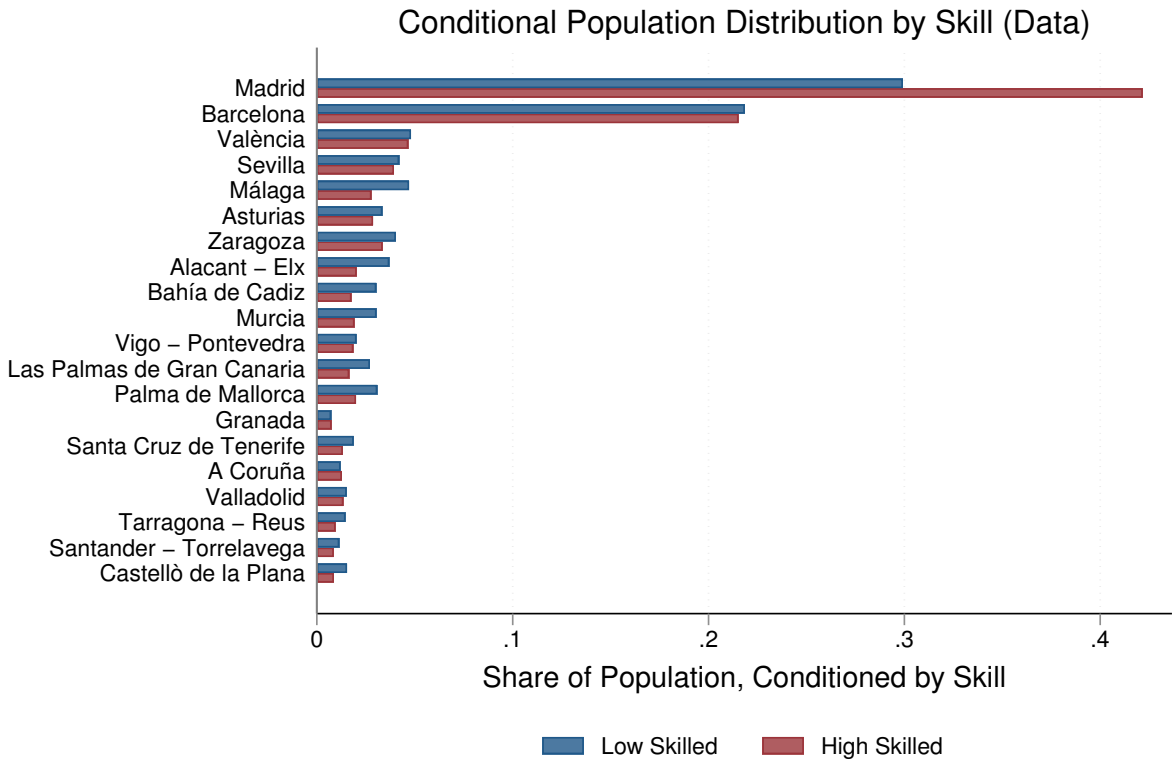
the ratio flips and there is more than one skilled worker per unskilled. This heterogenity in relative supplies has implications for the endogenous wages by sector in the model economy.

E.2.2 Amenities $a_{s,l}$

Amenities are an important parameter that needs to be calibrated internally in order to match the distribution of population by skill types. Matching this distribution is important since it directly regulates both the size and skill agglomeration forces. The necessity to match the distribution of unskilled and skilled stems from the fact that given the concavity of the value function not doing so would induce the skilled entrepreneurs who are at a higher point of the value function to move disproportionaly more than their less skilled counterparts. Figure $(E.12)$ shows the conditional population distributions being targeted by the amenities in the model.

One may notice that the high skilled are substantially more spatially concentrated than their

Figure E.12: Conditional Population across UAs Distribution by Skill

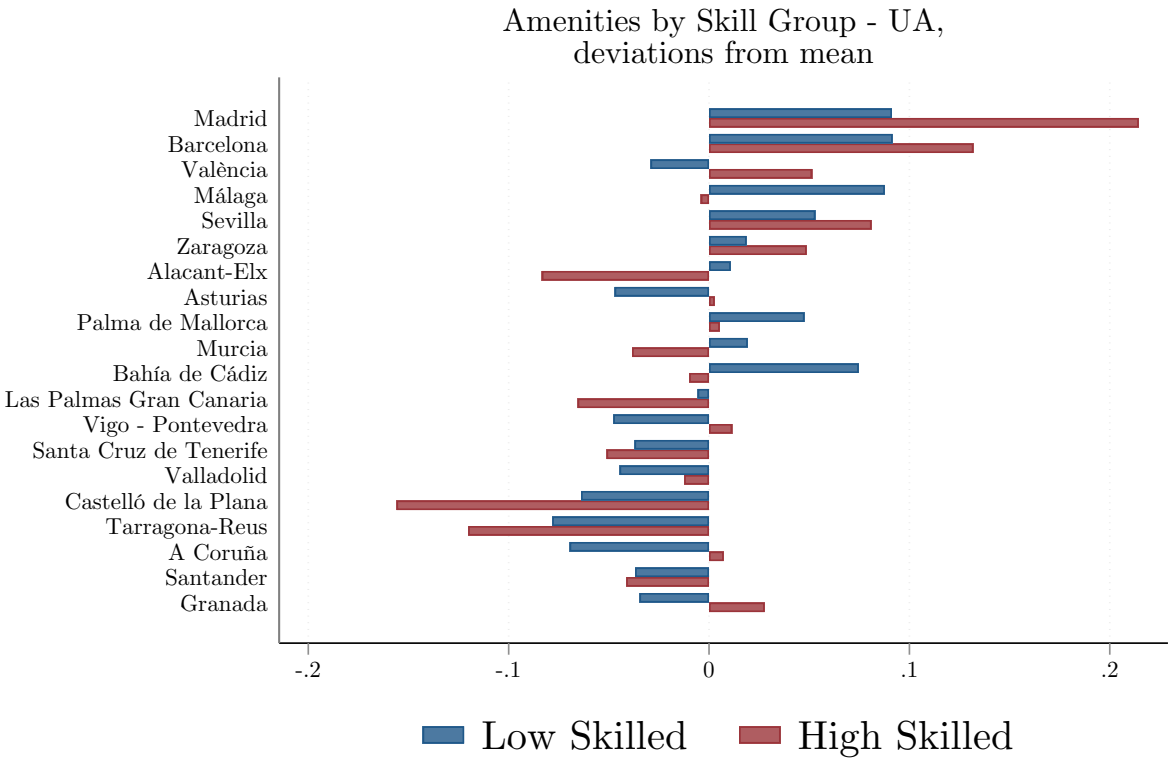


Notes: Conditional distribution of population by occupation in the MCVL data. The bars sum to 100% by occupation. The x-axis denotes the share of population of each skill type by UA. The y-axis indicates the UAs of the Spanish State.

less skilled counterparts. While Madrid and Barcelona (the largest 2 UAs) concentrate around 50% of the unskilled, the figure for the skilled goes well above to around 65%. The required amenities (in terms of deviations from the average amenities by skill) required to match that skill-wise population distribution is the following:

Intuitively, amenities should be higher for locations where given their economic offerings (wages, house prices, ...) there exists a necessity to attract even more population in the calibration. For example, València requiring less amenities than the mean for the low skilled could be interpreted as this UA already offering a good enough offering to the low skilled without the necessity for further amenities compared to the rest of the UAs.

Figure E.13: Amenities by Skill type in the Model (Deviations from Group Mean)



Notes: Internally calibrated amenities $a_{s,l}$ by skill type in the model. The blue bars represent the amenities for the low skilled, while the red ones do so for the high skilled. Amenities are reported in differences with respect to the mean of each group for better readability. The x-axis denotes absolute values. The y-axis lists the UAs of the Spanish State.

E.2.3 Migration Costs $\tau_{l,l',w}$ and $\tau_{l,l',e}$

Given the spatial nature of the model and the ability of individuals to move across space, it is crucial to match the share of movers across locations so that the model resembles the data. Recall that these costs are paid in terms of a fixed amount of utils. Since entrepeneurs enjoy higher incomes on average, a occupation-specific cost structure is necessary in order to match the mobility rate of these two groups in the data. For this exercise, the out-migration probabilities are targeted, which are reported in the following table:

In the case of the Spanish Economy, workers seem to be slightly more mobile across locations than their entrepreneurial counterparts. This could in part be due to the process involved in shutting down a firm (selling the assets, paperwork, firing employees...). The calibrated model is

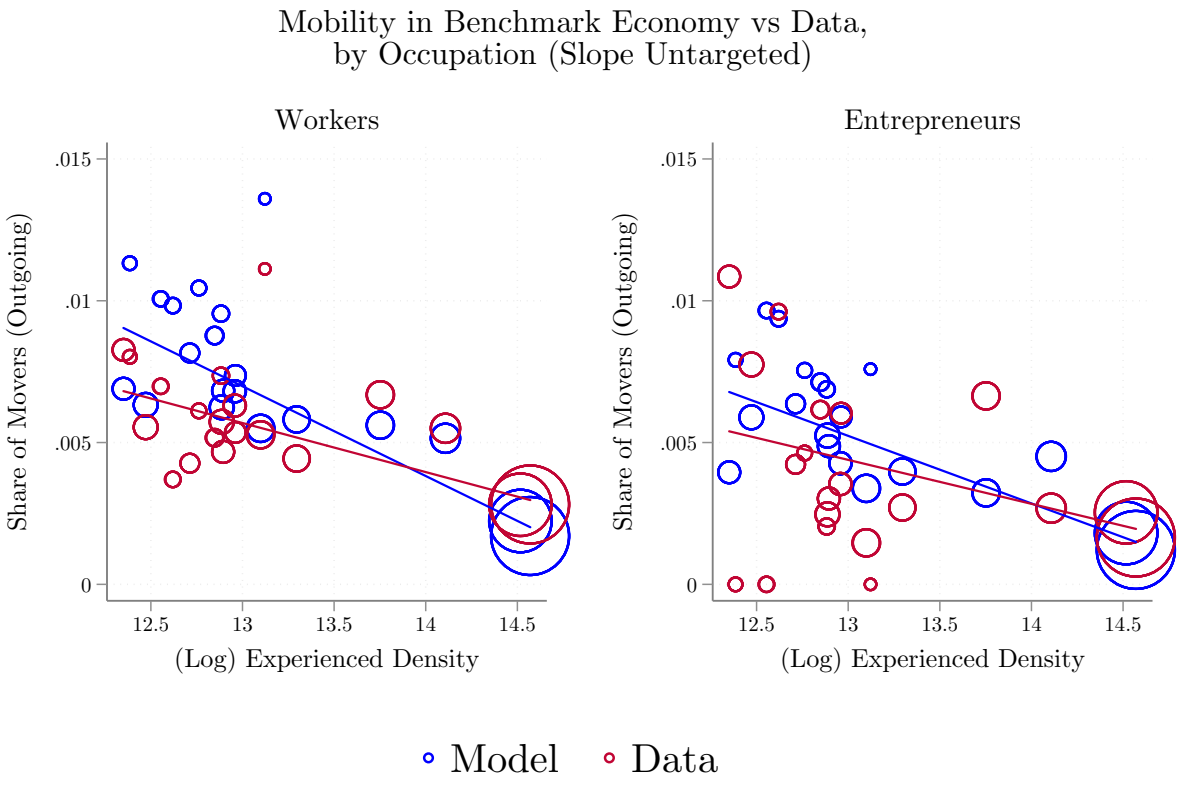
Table E.9: Mobility by Occupation in the Data

Group	Out-migration probability
Workers	0.45%
Entrepreneurs	0.29%

Notes: Out-migration probabilities by occupational group in the data (MCVL). The out-migration is defined as the fraction of people in a given UA in period t that move to another UA in period $t+1$.

able to properly capture the main relationship observed in the data, ie, that agents (both workers and entrepreneurs) move relatively more from smaller, less productive UAs to more productive, larger UAs. Figure $(E.14)$ shows precisely this.

Figure E.14: Mobility Patterns by Occupational Group in the Model and Data



Notes: Out-migration probabilities in the model (blue) and data(red). The out-migration is defined as the fraction of people in a given UA in period t that move to another UA in period $t+1$. The x-axis indicates the (Log) Experienced Density of the UA. The y-axis denotes the fraction. The left panel shows the relationship for workers while the right panel shows the one of entrepreneurs. Dots are scaled by the population of the UA.

Intuitively, larger, more productive UAs provide a larger value compared to their less produc-

tive counterparts. Therefore, in case of moving, the policy function $\mu_{l,l',o}$ puts more weight on the former set of UAs.

E.3 Housing Market

E.3.1 Housing Supply Elasticities η_l

Properly calibrating the housing supply elasticities η_l is paramount since it regulates the congestion forces in the model. In order to do so, we follow the procedure in Saiz (2010) which exploits geographical variation in the land availability near a UA in order to obtain heterogeneous estimates by UA. Intuitively, UAs that are sorrounded by either the ocean, steep terrain or rivers (Barcelona, San Francisco) are more geographically constained and therefore higher construction costs could be expected as opposed to geographically less constrained counterparts (Wyoming, Zaragoza).

This procedure requires detailed information on the location of the oceans, of inland bodies of water (weltands, rivers, springs...) and steep terrains (mountains, hills, ...) and housing stock and house prices. The employed datasets are the following:

- GEBCO's `https://www.gebco.net/` bathymetric data on the location of the oceans and elevation. Data for the year 2023 is taken. It provides global coverage of elevation at a 15 arc-second resolution.
- CORINE land coverage data `https://land.copernicus.eu/en/products/corine-land-cover`. Data for the year 2018 is taken.
- UA Atlas' data on house price and housing stock by UA `https://atlasau.mitma.gob.es/#bbox=-368050,5329651,135182,80768&c=indicator&i=siuare.sare001&selcodgeo=25&view=map5` Data for the years 2013-2018 is employed.

Step 1: Obtaining the Location of the Oceans and Seas and Steep Terrain

The first step is to load the raster image "gebco 2023 n44.9011 s26.488 w-19.0723 e5.3613.tif". This is a cut-down version of the global coverage that focuses on the area encompassed by the Spanish State (Canary Islands included).

From this, one can obtain the ocean and water bodies by selecting those coordinates with above 0 meters of elevation (note that the Spanish State has almost no points at below sea level). In order to compute the slopes/steepness of the terrain, the "gemgis.raster.calculate_slope()" function is employed on the raster image. This method employs a "Planar Method", which consists on computing the average rate of change of the surface in the x and y directions. To this end, for each point being evaluated, a 3x3 grid cell is constructed and the rates of change within this grid are computed in degrees. Following Saiz (2010), those points with a stepness above 15% are selected. The end result is provided in $(E.15)$, which shows the inferred ocean and sea location and steep terrain.

Overall, it has to be noted that most of the urban cores of the state are located next to the coast, which by just looking at the geographical map may not be obvious. In terms of steep terrain, there are five major mountain ranges in the country: the Pyrenees (limiting with France on the north), the Cantabrian mountains in the center-north, the Iberian System in the north-east, the Central System around Madrid and lastly the Baetic Mountain Range in the center-south. The Canary Islands, to the extent that they are volcanic islands, also feature a substantial proportion of steep terrain.

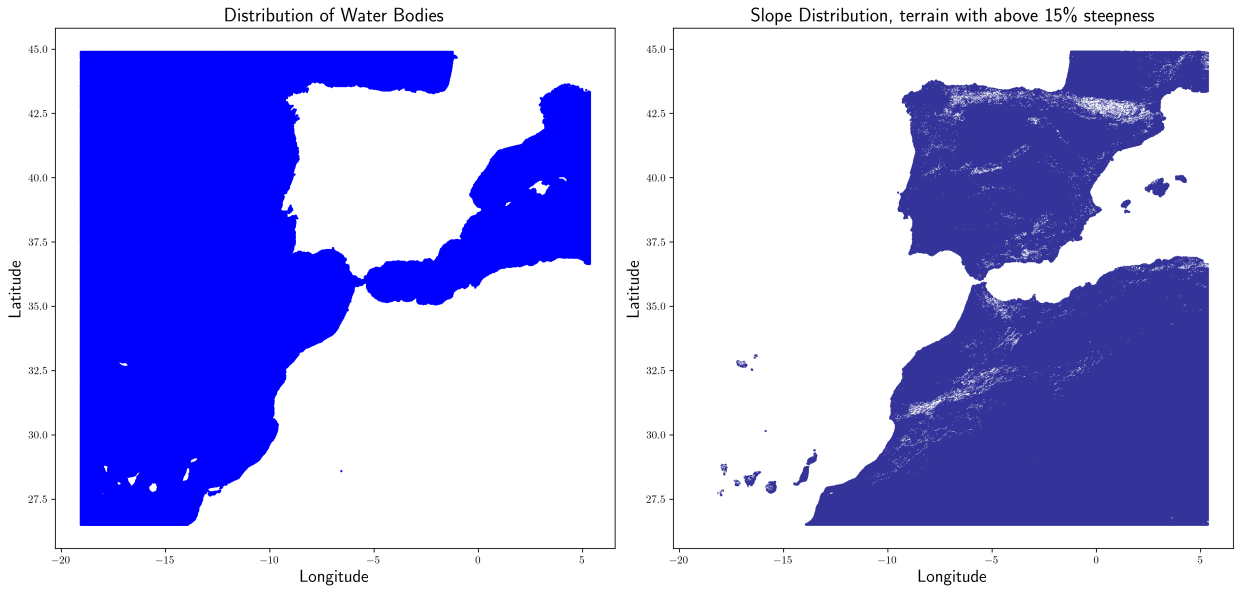
Step 2: Identifying the Inner Water Bodies

Since GEBCO's data provides no information on the inland water bodies, a separate satellite imagery dataset is required. The CORINE land cover dataset satisfies this condition as it provides a detailed classification of land coverage by categories.

The provided data on inland water bodies in CORINE includes: inland marshes, peat bogs, salt marshses, water courses, water bodies, coastal lagoons and estuaries. In this case, the provided ".shp" file is read directly by relying on the GeoPandas Python library. The data is filtered based on whether the code on what type of body covers teh water coincides with any of the above. Figure $E.16$ provides the spatial distribution of these inland water bodies:

The mainland of the state is home to six major rivers, which flow in a north-south order. The Ebro River originates in the Cantabrian Mountains and flows to the Mediterranean Sea, while the Miño River rises in the Galician Mountains and empties into the Atlantic Ocean. The Duero River begins in Soria and also flows toward the Atlantic. In addition, the Tajo River starts in Teruel and makes its way to Lisbon, and the Guadiana River flows from the Laguna de Ruideras to the

Figure E.15: Ocean and Sea Water bodies next to the Spanish State and its Steep Terrain



Notes: The left panel shows the ocean or sea coverage around the country. The right panel shows in white those areas of the country with above 15% degrees of steepness. According to the procedure in Saiz (2010), these inland areas are very costly to build upon.

Atlantic Ocean. Lastly, the Guadalquivir River originates in Quesada and reaches the Atlantic. Interestingly, the Spanish State counts with around 1,200 dams, mostly as an inheritance from the Francoist regime during the second half of the XX. century. [75]

Step 3: Putting everything together

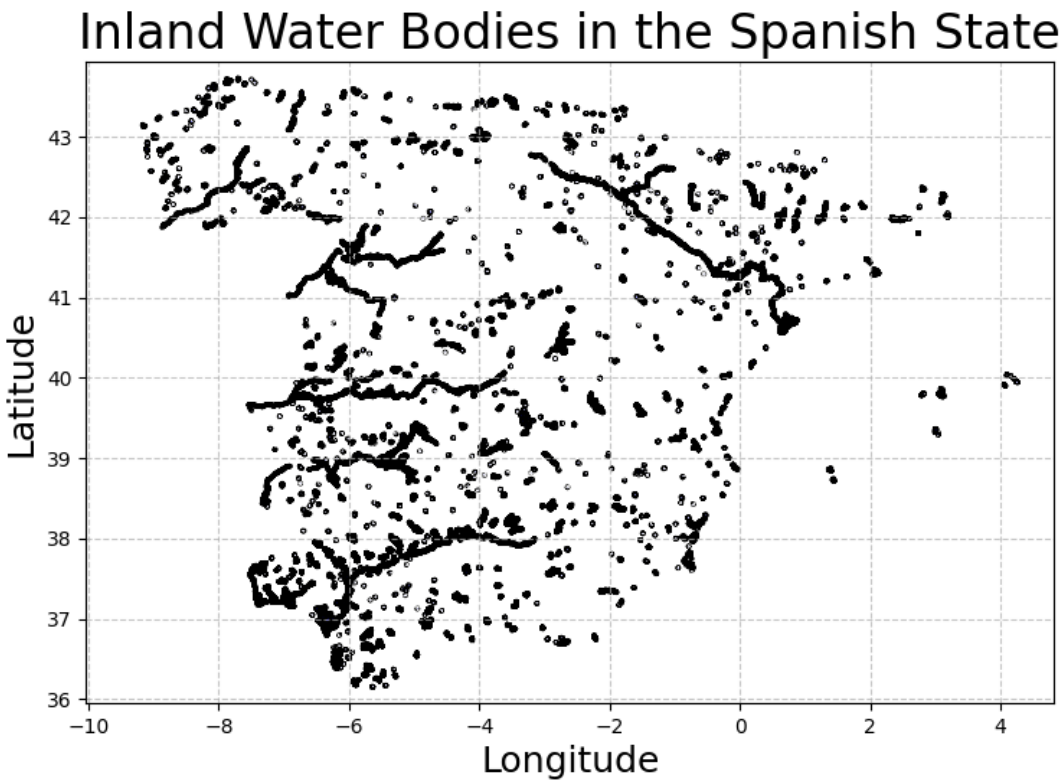
Once the coordinates of the water bodies and steep regions have been identified in the data, the next step is to compute what area around each UA is surrounded by such hard-to-build-on terrain.

In Saiz (2010) the buffer zone considered in order to compute the share of the land that is not easily buildable is 50km. This results problematic for a smaller country than the USA such as the Spanish State, as it leads to an overlap between buffer zones. Hence, the buffer zone is reduced to a radius of 20km, which is sufficient to cover all UAs of the Spanish State comfortably.

Once the 20km radius is constructed around each UA, the share of the circle made up of either

[75] The motivation for constructing these reservoirs stems from the warm Mediterranean climate, where authorities sought to store water during the months of highest rainfall to ensure reserves for the drier summer period. However, the technical standards of these projects, as well as their impact on local fauna, hydrology, and climate, have been increasingly called into question.

Figure E.16: Inland Water Bodies



Notes: Inland water bodies of the (mainland) Spanish State. These include the following categories of the CORINE Land Cover dataset: Inland Marshes, Peat Bogs, Salt Marshes, Salines, Interdital Flats, Water Courses, Water Bodies, Coastal Lagoons and Estuaries.

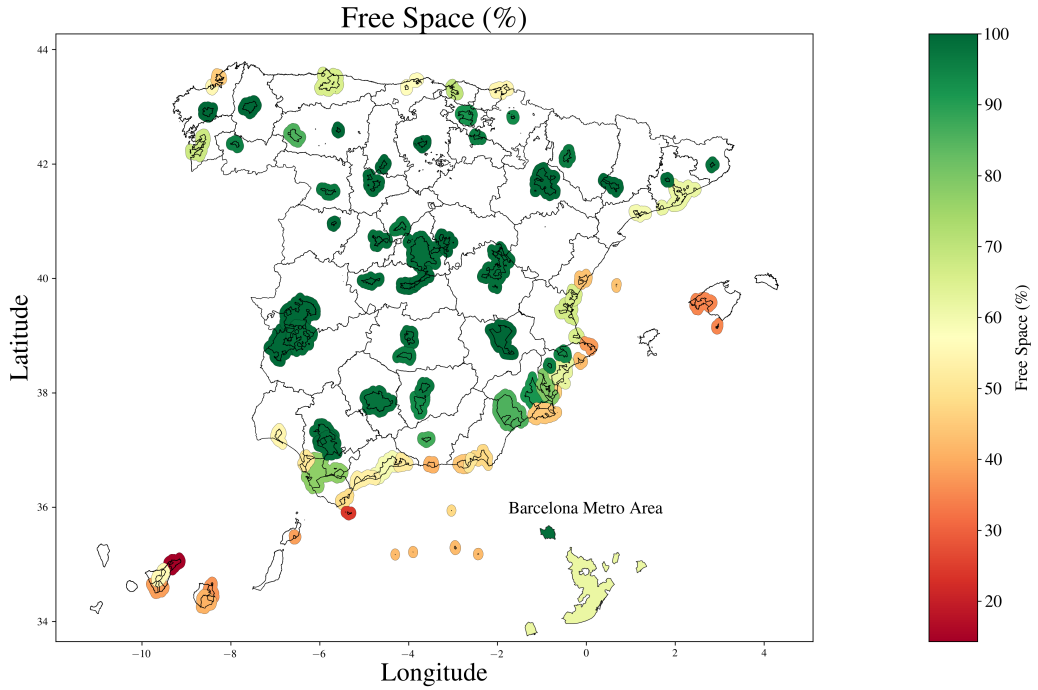
(i) ocean or sea water (ii) steep terrain or (iii) inland water bodies is computed. To the remaining area that is not taken up by any of these categories we denote "free space" and this is a proxy for the exogenously given geographical area where it is relatively unexpensive to construct.

As clearly ilustrated in figure $(E.17)$, the availability of terrain is primarily conditioned by the closeness to the sea. All interior UAs enjoy on average an availability of terrain over 70%, with most of them, including the largest (Madrid, Zaragoza, Sevilla...) reaching values above 90%. In contrast, most of the UAs located next to the coast hover around 30-40% of land availability in the 20km buffer zone.

Estimating the Housing Supply Elasticities η_l

The previous procedure allowed us to identify the area around each UA that is geographically constrained for construction. However, we are not yet done as we still need to estimate the hous-

Figure E.17: Area around each UA (20km Buffer Zone) not taken up by Water or Steep Terrain



Notes: Free space within a buffer zone of 20km in radius around each UA. Free Space is the total area of the circle minus the part taken by water (be it ocean/sea or inland water) or by steep terrain (hills, mountains...). The halos around each UAs are buffers in order to improve readibility.

ing supply elasticities, for which the obtained measure will be a primary component. In order to estimate the actual parameter of interest, η_l, we require data on how the stock of housing responds to prices.

To that end, the appraised home value of the total stock of housing from the UA Atlas will be employed as a proxy for the price of housing. As for the housing stock, we will take the value on the total amount of households by UA as reported in the Census of 2011 and 2021. Let us denote the former by $P_{\{2011,2021\}}$ and the latter $H_{\{2011,2021\}}$ for the years 2011 and 2021, respectively. Then, the standard housing supply elasticity estimation would proceed as follows, in its most basic form:

$$\Delta \log P_l = \beta_0 + \beta_1 \Delta \log H_l + \epsilon_l$$

Where H_l may be instrumented by a variable that is correlated with the size of the municipality but uncorrelated with the error term (such as the average temperatures in January, a bartik instrument of industry composition etc) given the endogeneity through the demand side (Saiz,

2010). Which would generate a location homogeneous elasticity of housing supply $\eta_l = \eta = \frac{1}{\beta_1}$. Fortunately, we can employ our estimated measure of free space in order to generate heterogeneous estimates of the housing supply elasticity. In particular, the specification to be run will be the following:

$$\Delta \log P_l = \beta_0 + \beta_1 \Delta \log H_l + \beta^{\text{Land}} \times (1 - \Lambda_l) \Delta \log H_l + \epsilon_l$$

Where $(1 - \Lambda_l)$ stands for the share of unavailable land for development. This specification now yields housing supply elasticities $\eta_l = \frac{1}{\beta_1 + \beta^{\text{Land}} \times (1 - \Lambda_l)}$, which allows for heterogeneity across locations by exploiting the differing degrees of land availability. Thus, those UAs with a higher $\uparrow (1 - \Lambda_l)$ will have a lower estimated housing supply elasticity η_l.

The final estimated elasticities by UA are shown in figure $(E.18)$. In line with the intuition obtained from the map above showing the percentage of free available space by UA, those UAs with higher housing supply elasticities are precisely those located in the interior of the country (Zaragoza, Valladolid, Sevilla, ...). In contrast, those UA located in isles (Palma de Mallorca, Santa Cruz de Tenerife, Las Palmas...) face the highest land availability constraints and have therefore the lowest housing supply elasticities.

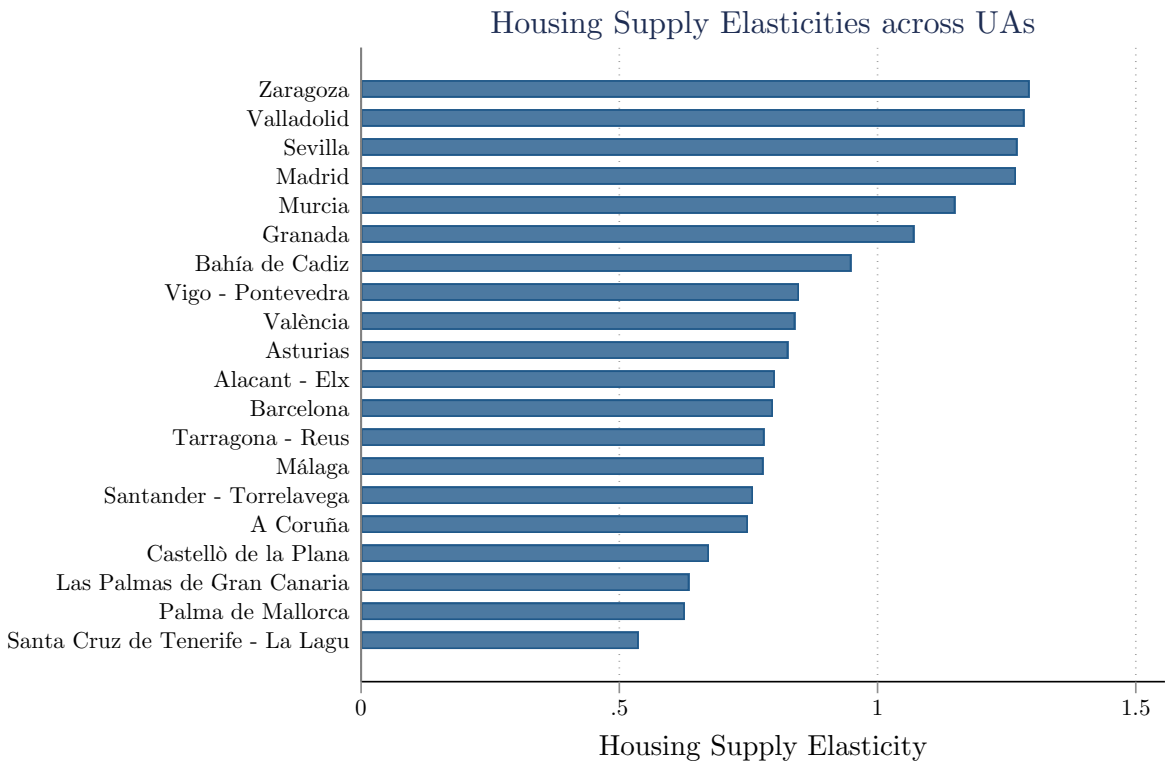
E.3.2 Relative Ratio of Exogenous Housing Supply $\{\bar{H}_l\}$

Along with the housing supply elasticities η_l, another pivotal set of parameters that govern the housing supply curve of the construction sector $H_l^s = p_{h,l}^{\eta_l} \bar{H}_l$ are the exogenous endowments of land per location $\bar{H}_l$.

In order to inform this set of parameters, the chosen approach is to first calibrate internally the endowment of Madrid so that a unit of housing $h = 1$ costs 30% of the average salary in Madrid (across sectors and skill levels). Then, we calibrate internally the remaining $L - 1$ exogenous endowments $\{H_l\}_{l=2}^{L}$ in order to match the observed price differentials per squared meter across UAs with respect to Madrid. Figure $(E.19)$ presents the data on the obtained price ratio of squared meters.

Overall, the largest UAs of the State, Madrid and Barcelona, are the most expensive, along with the capital of the Balearic Islands, Palma de Mallorca.

Figure E.18: Estimated Housing Supply Elasticities by UA



Notes: Estimated housing supply elasticities following the approach in Saiz (2010). The estimated specification is $\Delta \log P_l = \beta_0 + \beta_1 \Delta \log H_l + \beta^{\text{Land}} \times (1 - \Lambda_l) \Delta \log H_l + \epsilon_l$, where $(1 - \Lambda_l)$ stands for the share of unavailable land by UA in a 20km radius. $\Delta \log P_l$ is the log difference in prices between 2021 and 2011 (defined as the appraised value of the total stock of housing per squared meter by UA). $\Delta \log H_l$ is the log difference in the housing stock defined as the total stock of households from the Census in 2021 and 2011. The figure shows the estimates $\eta_l = \frac{1}{\beta_1 + \beta^{\text{Land}} \times (1 - \Lambda_l)}$ by UA.

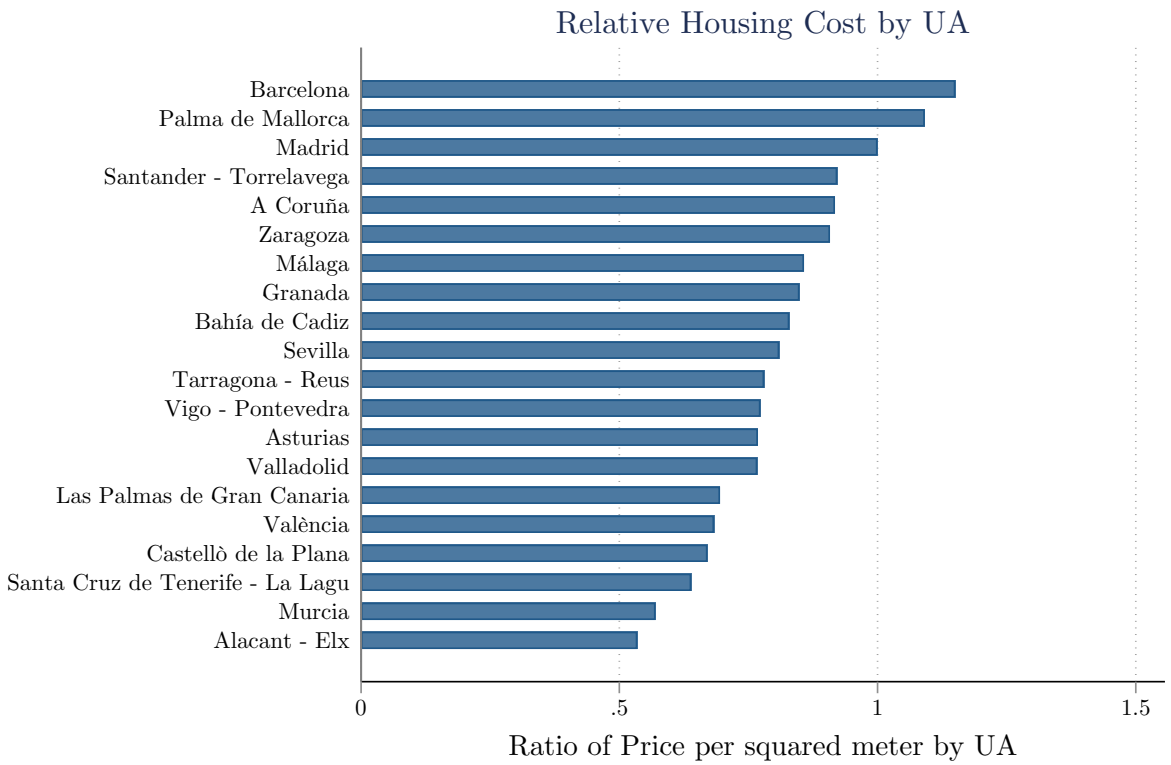
E.4 Production Parameters

E.4.1 Fixed Costs of Production

As argued in the model section of the main paper, the per-period fixed costs of production F_j play the crucial role of regulating how many entrepreneurs there are per sector in the economy. Intuitively, given the financial frictions, profits are determined (mainly) by the productivity of the entrepreneur z, their asset level a and the productivity of the UA. The fixed costs bring down the profit level across the board in order to guarantee that only the desired fraction of the population find it profitable to operate a business.

These are calibrated internally in order to match the following country-wide stock of en-

Figure E.19: Targeted Price ratios to inform Exogenous Housing Endownments



Notes: Ratios of the price of a squared meter of housing stock by UA with respect to Madrid. Data is taken from the appraised value of the total housing stock from the UA Atlas for the year 2011.

treprenerus per sector:

Table E.10: Targeted Moments for calibrating the Fixed Costs of Production

Sector	Share Entrepreneurs (over total Pop.)
Low Sectors	1.1%
Medium Sectors	1.4%
High Sectors	0.8%

Notes: Employed moments for targeting the share of entrepreneurs by sector at the country level. The data is the average stock of entrepreneurs by sector over total population obtained from the MCVL for the sample period 2013-2018.

Overall, the the share of incorporated is 3.3% in the economy. The Medium sector contains the highest proportion thereof, followed by the Low and High sectors. The endogenous spatial distribution of this entrepreneurship is not calibrated and left as an endogenous object for the model to capture.

E.4.2 Leverage Ratios λ_j

The crucial parameter governing the strength of financial frictions are the leverage ratio parameters λ_j by sector. Intuitively, as argued in the model section of the main paper, a higher value for this parameter implies that entrepreneurs are closer to being able to finance their unconstrained level of capital.

Since the ratio borrowed funds in the model over the assets of the firm are equal to $\frac{\{\max 0, k-a\}}{k}$, this creates a direct parallel to the debt-to-assets ratios in the data. In order to account for the size distribution of firms, the average sector-wise revenue-weighted debt-to-assets ratios are targeted. The following table presents the moments in the data employed to calibrate the λ_j internally:

Table E.11: Targeted Moments for calibrating the Leverage Ratios in the Model

Sector	Revenue-weighted debt-to-assets ratio
Low Sectors	56.7%
Medium Sectors	55.2%
High Sectors	51.2%

Notes: Employed moments for targeting the leverage ratios λ_j in the model. The data is the average revenue-weighted deb-to-assets ratio by sector obtained from SABI for the sample period 2013-2018.

The obtained λ_j parameters that correspond to targeting these moments internally are in turn: $\lambda_{\text{Low}} = 4.35$, $\lambda_{\text{Medium}} = 2.97$, $\lambda_{\text{High}} = 3.08$. Although these values might seem high compared to other papers in the literature, these results have to be understood in conjunction with the calibration for the volatility of the entrepreneurial process, which we shall discuss shortly.

E.4.3 Production Multipliers for the High Skilled $\Phi_{s,j}$

A key observation from figure $(A.2)$ was that both the workers and incorporated entrepreneurs are sorted differentially across sectors as a function of skill. The joint fixed distribution over skills s and sectors j ensures that the skill distribution of workers across sectors is consistent with the data. That notwithstanding, the selection into entrepreneurship is endogenous and therefore an additional parameter is required that governs the skill composition of entrepreneurs by sector in order to match the data.

That role is played by the sector-specific productivity boosts to high skilled entrepreneurs, $\Phi_{s,j}$. Intuitively, a higher productivity boost to the high skilled increases the returns of entrepreneur-

ship for this group. This set of parameters is calibrated internally in order to match the following moments in the data:

Table E.12: Targeted Moments for calibrating the Productivity Boost to High Skilled Entrepreneurs in the Model

Sector	Share of high skilled incorporated by sector
Low Sectors	28.7%
Medium Sectors	38.8%
High Sectors	68.8%

Notes: Employed moments for targeting the productivity boosts of the high skilled entrepreneurs $\Phi_{s,j}$ in the model. The data is the share of high skilled incorporated entrepreneurs (as defined in table $(E.16)$ by sector obtained from SABI for the sample period 2013-2018.

The required values to match the data are, respectively, $\Phi_{h,\text{Low}} = 1.28937$,$\Phi_{h,\text{Medium}} = 1.0652$,$\Phi_{h,\text{High}} = 1.0976$.

E.4.4 Productivity Scalers ϕ_j

In the presence of financial frictions the profit share out of revenues is no longer equal to $1 - \mu_j$ but rather a function of the average strength thereof across the entire distribution. For this reason, to guarantee that the share of entrepreneurs, a given average MPK Premium and the profit shares can be matched simultaneously, an additional parameter is required, the productivity scalers ϕ_j [76]. The intuition behind this is the following. The per-period fixed costs F_j in conjunction with the additive productivity scalers ϕ_j mainly determine the extensive margin and the profit rate. If the profit rate is too high for a given μ_j and σ_z, then there is too much entry and $\uparrow F_j$, which decreases the number of entrepreneurs and the profit rate. The additive ϕ_j essentially regulates how much the average entrepreneur by sector is receiving in profits out of revenues [77].

These parameters are calibrated internally in order to match the profit rate by sector as observed in the SABI dataset for the period 2013-2018. The employed definition of profits is such that it matches the model as closely as possible and is equal to: $\frac{\text{Profits}}{\text{Revenues}} = \frac{\text{(Revenues - Wagebill - (Ebitda - Ebit) - Materials)}}{\text{Revenues}}$

[76] The intuitive solution would be to use the DRS parameter μ_j in order to regulate the profit rate. However, in the presence of financial frictions it no longer uniquely determines the profit rate. Since altering the μ_j has large effects on the MPK Premium, extensive margin and profit rates, this alternative calibration strategy is pursued to guarantee that the model matches all three moments reliably. The μ_j parameters are set at their unconstrained level, which is equal to 1 minus the profit rate by sector as observed in the SABI dataset for the period 2013-2018.

[77] If the producitivity scaler entered additively, then the entire mass of productivities would move proportionally and we could not find the average entrepreneur for whom the profit rate is the desired one.

Said profit rates are provided in the following table:

Table E.13: Targeted Moments for calibrating the Productivity Scalers by Sector in the Model

Sector	Profit rate by sector in SABI
Low Sectors	24.1%
Medium Sectors	18.0%
High Sectors	29.7%

Notes: Employed moments for targeting the productivity scalers by sector ϕ_j in the model. The data is the average profit rate by sector in the SABI dataset over the period 2013-2018. The definition of profits employed in the data is that of the model and is equal to $\frac{\text{Profits}}{\text{Revenues}} = \frac{\text{(Revenues - Wagebill - (Ebitda - Ebit) - Materials)}}{\text{Revenues}}$.

The obtained productivity scalers are, respectively, $\phi_{\text{Low}} = -0.505$, $\phi_{\text{Low}} = 3.765$, $\phi_{\text{Low}} = -0.876$.

E.4.5 Exogenous UA-Sector specific Productivities $A_{j,l}$

In the model section of the main paper we emphasized how the productivity of the different UA-Sectors is endogenous through the agglomeration forces. However, these agglomeration forces build upon the exogenous productivity of each location $A_{j,l}$. Thus, for the same agglomeration forces, exogenously more productive locations will attain higher productivity levels.

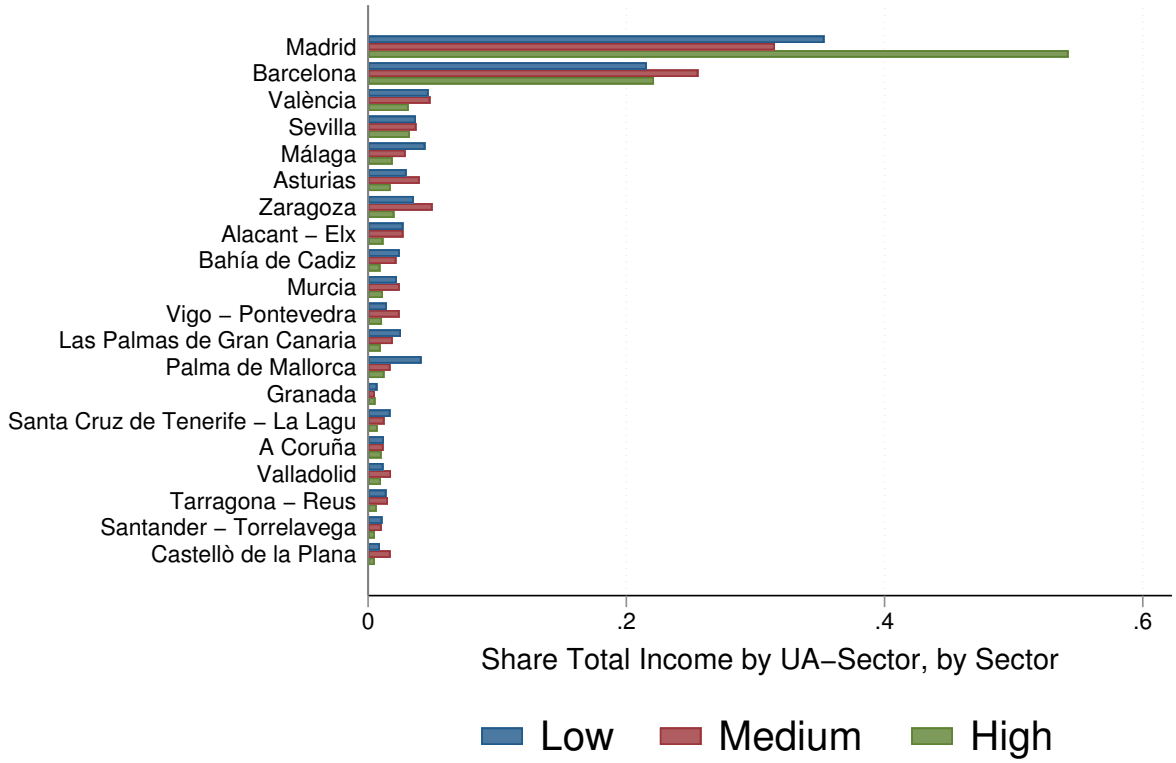
We address the challenge of estimating these productivities by calibrating them internally. The internal calibration requires disciplining the $J \times L$ exogenous location-sector productivities $A_{j,l}$ by employing some moment in the data. We employ either the share of total income by UA-Sector conditional on a given Sector from the MCVL dataset or the share of revenues of a given UA-Sector conditional on a given Sector from the SABI dataset. Both provide very similar estimates. This additional set of parameters is required as otherwise there is no guarantee that the estimated agglomeration forces are enough to explain the per capita differences in production across sectors and UAs.

We normalize Madrid to have a productivity of one and the rest of the UA-Sectors are calibrated internally so that they produce their given share. We employ the sector-conditioned distribution as this allows the normalization of the productivities in Madrid to function properly. [78] Figure E.20 presents the distribution of the targeted moment and figure E.21 displays

[78] Doing the overall UA-Sector production distribution over the entire production would not allow the normalization in Madrid to generate the same shares as in the data.

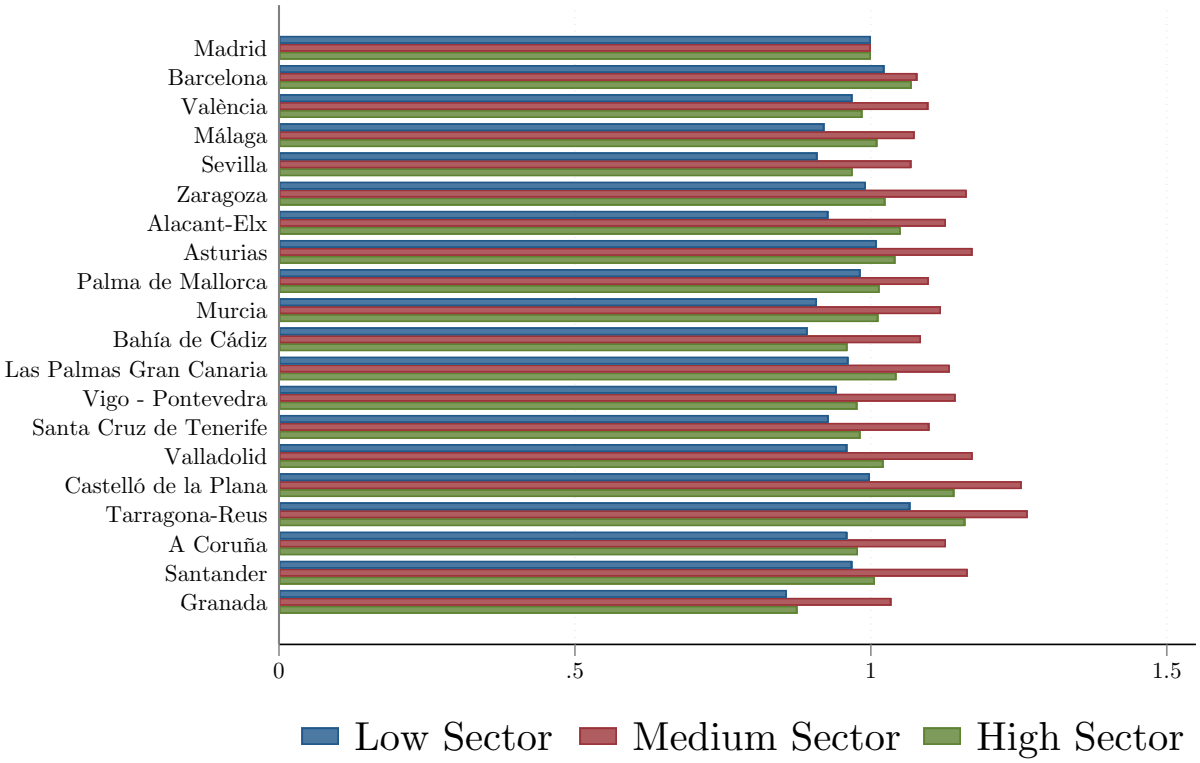
the obtained productivities. Figure E.22 reports the effective productivities by UA-Sector $A_{j,l} \times \left(1 + \left(\frac{L_l}{L}\right)^{\Omega_{\text{SIZE}}} \left(\frac{H_l}{U_l}\right)^{\Omega_{\text{SKILL}}}\right)$.

Figure E.20: Targeted moments for $A_{j,l}$: the Spatial Distribution of Total Income by UA-Sectors conditional on a Sector



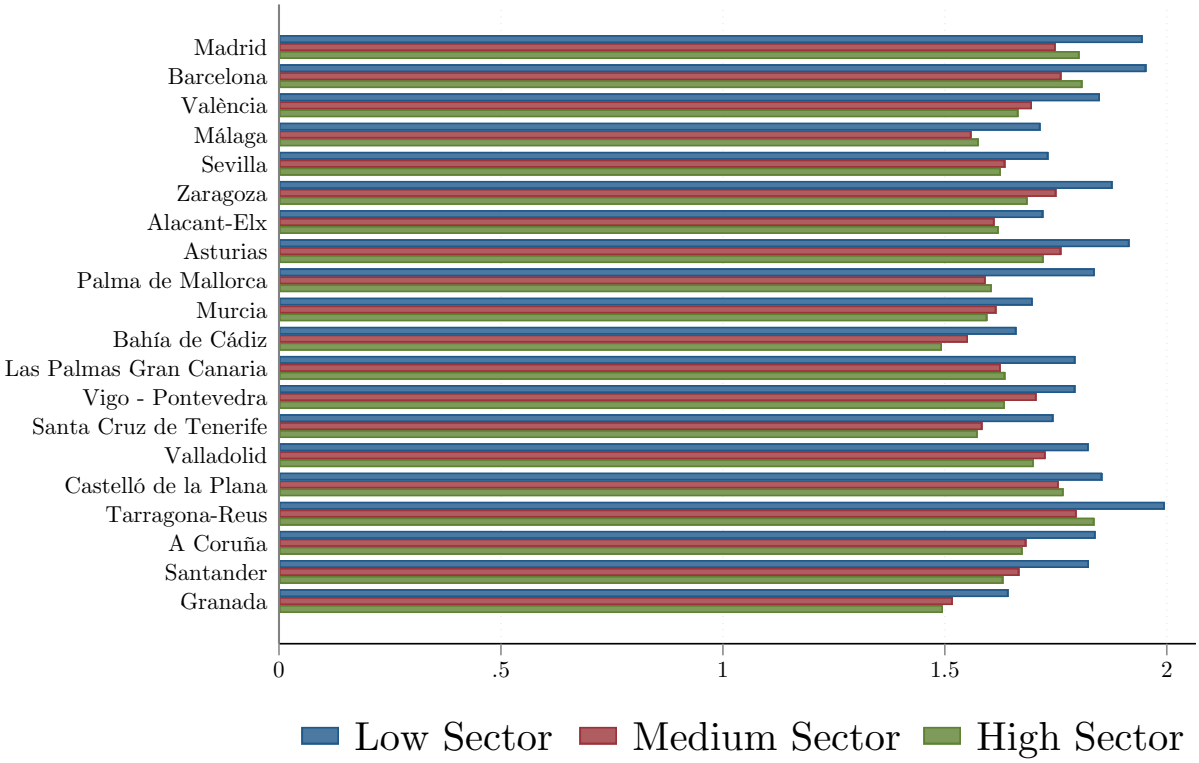
Notes: The figure displays the targeted moments to calibrate the $A_{j,l}$ exogenous productivities internally. The moment is the share of total income by UA-Sector over the total income of a given Sector. The blue bar represents conditional distribution for the Low Skilled Sectors. The red bar does likewise for the Medium Skilled sectors. Lastly, the green one denotes the high skilled sectors.

Figure E.21: Estimated Exogenous Productivities by UA-Sector $A_{j,l}$



Notes: Internally estimated exogenous UA-Sector productivities $A_{j,l}$. The y-label in the figure indicates the UA and the bars correspond to one of the skill-based sectors. All productivities are normalized with respect to Madrid (the capital) by Sector.

Figure E.22: Effective Productivities by UA-Sector $A_{j,l} \times \left(1 + \left(\frac{L_l}{L}\right)^{\Omega_{\text{SIZE}}} \left(\frac{H_l}{U_l}\right)^{\Omega_{\text{SKILL}}}\right)$



Notes: Effective productivities by UA-Sector $A_{j,l} \times \left(1 + \left(\frac{L_l}{L}\right)^{\Omega_{\text{SIZE}}} \left(\frac{H_l}{U_l}\right)^{\Omega_{\text{SKILL}}}\right)$. The y-label in the figure indicates the UA and the bars correspond to one of the skill-based sectors.

E.4.6 Input Shares in Production $\alpha_{i,j}$ by Input-Sector

As argued in the main paper, a major reason for including multiple sectors was to allow differential intensities in the use of inputs in the economy. In order to capture the intensity at which sectors operate the different inputs, their weights in production $\alpha_{i,j}$ need to be estimated for every input and every sector. The employed procedure makes use of the FOC in the unconstrained case.

Note that the assumed CES production function in the economy leads to the profit function:

$$\pi(\cdot) = p_j(z_t + \phi_j)\Phi_{s,j}A_{j,l}\left(1 + \left(\frac{L_l}{L}\right)^{\Omega_{\text{SIZE}}}\left(\frac{H_l}{U_l}\right)^{\Omega_{skill}}\right)\left(\alpha_{H,j}h_{f,j}^{\frac{(\xi-1)}{\xi}} + \alpha_{U,j}u_{f,j}^{\frac{(\xi-1)}{\xi}} + \alpha_{K,j}k_f^{\frac{(\xi-1)}{\xi}}\right)^{\mu_j\frac{\xi}{\xi-1}}$$
$$-W_{H,j,l}h_{f,j} - W_{U,j,l}u_{f,j} - (r_t + \delta)k_f - F_j$$

Taking the FOC:

$$\frac{\partial\pi}{\partial h_f} = p_j(z_t + \phi_j)\Phi_{s,j}A_{j,l}\left(1 + \left(\frac{L_l}{L}\right)^{\Omega_{\text{SIZE}}}\left(\frac{H_l}{U_l}\right)^{\Omega_{\text{SKILL}}}\right)\mu_j\alpha_{H,j}\left(X\right)^{\mu_j\frac{\xi}{\xi-1}-1}h_{f,j}^{\frac{-1}{\xi}} - W_{H,j,l}$$

$$\frac{\partial\pi}{\partial u_f} = p_j(z_t + \phi_j)\Phi_{s,j}A_{j,l}\left(1 + \left(\frac{L_l}{L}\right)^{\Omega_{\text{SIZE}}}\left(\frac{H_l}{U_l}\right)^{\Omega_{\text{SKILL}}}\right)\mu_j\alpha_{U,j}\left(X\right)^{\mu_j\frac{\xi}{\xi-1}-1}u_{f,j}^{\frac{-1}{\xi}} - W_{U,j,l}$$

$$\frac{\partial\pi}{\partial k_f} = p_j(z_t + \phi_j)\Phi_{s,j}A_{j,l}\left(1 + \left(\frac{L_l}{L}\right)^{\Omega_{\text{SIZE}}}\left(\frac{H_l}{U_l}\right)^{\Omega_{\text{SKILL}}}\right)\mu_j\alpha_{K,j}\left(X\right)^{\mu_j\frac{\xi}{\xi-1}-1}k_f^{\frac{-1}{\xi}} - (r_t + \delta)$$

Where the compound term X is:

$$X = \alpha_{H,j}h_{f,j}^{\frac{\xi-1}{\xi}} + \alpha_{U,j}u_{f,j}^{\frac{\xi-1}{\xi}} + \alpha_{K,j}k_f^{\frac{\xi-1}{\xi}}$$

By combinig the three FOCs together, one can obtain that:

$$u = \left(\frac{w_{u,j}}{w_{h,j}} \frac{\alpha_{h,j}}{\alpha_{u,j}} \right)^{-\epsilon} h \qquad\qquad u = \left(\frac{w_{u,j}}{R} \frac{\alpha_{k,j}}{\alpha_{u,j}} \right)^{-\epsilon} k$$

The final step is to employ the data counterparts in order to inform the parameters of the model. By relying on SABI data for the small firms over the period 2013-2018, one can obtain the revenue-weighted average wagebill, capital and capital costs by sector. Since the SABI provides no data on the distinction between skilled and unskilled workers, we complement the SABI data with the MCVL in order to construct estimates of the wagebill and workforce that correspond to each of these groups [79].

Following this procedure leads to the following estimates of the weights in production of the different inputs by sector $\alpha_{i,j}$:

Table E.14: Estimated Weights in Production by Input-Sector $\alpha_{i,j}$

Input Sector	Low	Medium	High
Low Labour	0.426	0.377	0.217
High Labour	0.25	0.311	0.464
Capital	0.323	0.311	0.318

Notes: Estimated input-sector production weights $\alpha_{i,j}$ by sector. The methodology relies on combining the FOCs the CES production function in the unconstrained case in order to isolate these weights in terms of measurable objects in the data. A combination of the SABI dataset and MCLV datasets over the sample period 2013-2018 is employed in order to construct the required wages by skill, low and high skilled stocks of labour, capital costs and capital stocks by sector.

Overall, the weight in production of low skilled labour is monotonically decreasing in the sectors' skill while that of the high skilled labour is monotonically increasing. Interestingly, the weight of capital remains stable across all three sectors.

[79] In particular, we assume that the wagebill and the workforce are distributed by sector in the same vein as in the MCVL. Thus, for example, if 60% of the workforce and 70% of the wagebill corresponds to the high skilled in the high sector, the average employment and average wagebill in the SABI dataset is distributed among the skill types in this manner.

E.4.7 Size and Skill Agglomeration Forces by Sector $\Omega_{\text{size, j}}$ and $\Omega_{\text{skill, j}}$

A central element of the productivity by UA-Sector in the model were the size and skill agglomeration forces. These are crucial in order to understand each sector's sensitivity to urbanization economics and knowledge spillover effects (just to name a two potential sources of agglomeration economies). In order to obtain estimates for these we proceed based on Giannone (2017).

The methodology starts with the observation that from the FOC of the profit function with respect to high labour:

$$\frac{\partial \pi}{\partial h_f} = p_j(z_t + \phi_j)\Phi_{s,j}A_{j,l}\left(1 + \left(\frac{L_l}{L}\right)^{\Omega_{\text{SIZE}}}\left(\frac{H_l}{U_l}\right)^{\Omega_{\text{SKILL}}}\right)\mu_j\alpha_{H,j}\left(X\right)^{\mu_j\frac{\xi}{\xi-1}-1}h_{f,j}^{\frac{-1}{\xi}} - W_{H,j,l}$$

Here we obtain the sector-location wages of the high skilled related to the (i) agglomeration forces in the UA-Sector (ii) the currently employed level of high skilled workers (iii) the overall level of production (iv) additional productivity terms. The main idea consists on re-writing this equation as: [80]

$$\underbrace{\log w_{h,j,l} + \left(\frac{1}{\epsilon}\right)\log h_{l,j} - \frac{(\mu\epsilon - \epsilon + 1)}{(\epsilon - 1)}\log\left(\sum_{i=0}^{i=N_{l,j}-1}\frac{y_{i,l,j}}{N_{l,j}}\right)}_{\text{DEPENDENT VAR}} = c + \underbrace{\Omega_{skill}\log\left(\frac{H_l}{U_l}\right)}_{\text{SKILL AGGLOMERATION}} + \underbrace{\Omega_{size}\log\left(\frac{L_l}{L}\right)}_{\text{SIZE AGGLOMERATION}} + \epsilon_{l,j}$$

By re-arranging the terms, we end up with an specification that relates the wages (controlling for employment and production levels) to the agglomeration forces. Given that we have data on all necessary components, we can run this specification and obtain estimates of the agglomeration forces $\hat{\Omega}_{\text{size, j}}$ and $\hat{\Omega}_{\text{skill, j}}$. The $w_{h,j,l}$ will be the average location-sector annual wage obtained from the MCVL. $h_{l,j}$ is the average-location sector stock of high skilled labour by firms in SABI weighted by the revenue of each firm and constructed by assuming that the workforce composition by skill is the same by sector-location as in the MCVL. $\left(\sum_{i=0}^{i=N_{l,j}-1}\frac{y_{i,l,j}}{N_{l,j}}\right)$ is the average revenue by sector-location from SABI. $\left(\frac{H_l}{U_l}\right)$ is the high skilled to low skilled ratio by location from the MCVL.

[80] Here we are effectively removing the 1+ that made agglomeration forces increase the productivity over the base level of the UA in order to have a linear specification on the agglomeration force parameters Ω_{skill} and Ω_{size}.

Lastly, $\left(\frac{L_l}{L}\right)$ is the share of inhabitants at each location from taken from the MCVL.

To alleviate endogeneity concerns, $\left(\frac{H_l}{U_l}\right)$ is instrumented with a Bartik instrument on the stock of inmigrants and $\left(\frac{L_l}{L}\right)$ is in turn instrumented with the estimated housing supply elasticities η_l. The obtained estimates are reported in the following table:

Table E.15: Estimated Size and Skill Agglomeration Forces $\Omega_{size,j}$, $\Omega_{skill,j}$ by sector

	IV		OLS	
Skill Sectors (Low, Medium, High)	Ω_{size}	Ω_{skill}	Ω_{size}	Ω_{skill}
Low	-0.032*	.107	.007	.389***
Medium	.095**	.351***	.060***	.560***
High	-.019*	.417***	-.004	.479***

Notes: Estimated size and skill agglomeration forces by sector. The methodology is based on Giannone (2017).

A few comments are in place. First, regardless of whether the coefficients are estimated via IV or OLS the skill agglomeration forces appear to be stronger in absolute terms than their size counterparts. This is consistent with the results in Giannone (2017). Second, instrumenting the skill ratio and share of inhabitants qualititatively changes the ranking of the strength of the agglomeration forces across sectors. In the IV estimation, the size agglomeration forces are monotonically increasing in the skill level of the sector. Third, size agglomeration forces appear to only be relevant for the medium sector and negative for the low and high sectors. In the model, we truncate the agglomeration forces at zero thus not allowing for negative agglomeration forces.

E.4.8 Weights of the Intermediate Goods Sectors γ_j

Recall that from the assumption that the final good is a CES composite of the different intermediate goods in the economy, we can obtain the country-wide demand for each intermediate as:

$$Y_j = \left(\frac{\gamma_j P_j^{-\rho} Y}{P^{(1-\rho)}}\right) \forall j \in J$$

Since the overall price index P is normalized to unity in the economy, the demand for the amount of intermediate goods Y_j depends on the weight of the sector γ_j, the price of the intermediate good $P_j^{-\rho}$ and the overall demand level in the country Y. In particular, those intermediates

goods with a lower price $\downarrow P_j$ will enjoy greater demand $\uparrow Y_j$.

The weight parameters of the intermediate goods Y_j in the economy are thus calibrated internally in order to match the fraction of fiscal income by sector Spanish economy. Note that the object being targeted through the model's perspective is $\gamma_j = \frac{P_j Y_j}{PY}$ where recall that $P = 1$. The following table presents the results of the estimated parameters:

Table E.16: Internally calibrated weights of the Intermediate Goods γ_j and Targeted Shares in the Data

Sector	Obtained γ_j estimate	Weight of sector in the data
Low Sectors	0.2598	0.265252
Medium Sectors	0.334016	0.4588687
High Sectors	0.406085	0.275

Notes: Internally calibrated weights of the intermediate goods in the economy γ_j and the targeted shares of the sectors out of total fiscal income in the data. The data corresponds to the share of fiscal income of individuals by sector over total fiscal income as observed in the MCVL over the period 2013-2018.

E.5 Populaton and Skill Distribution across UAs

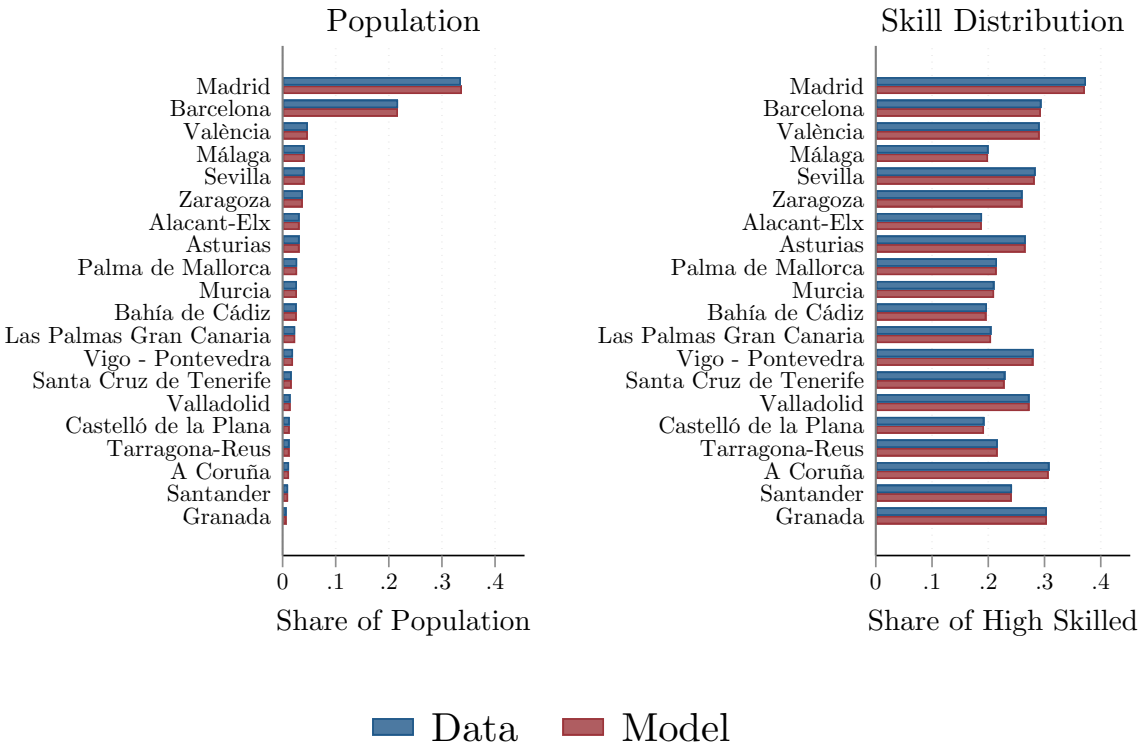
The obtained population and skill distribution across UAs is reported in figure E.23. This is an important targeted moment as it regulates the agglomeration forces in the model and hence productivity.

Interestingly, it appears to be the case that Zipf's law does not evidently apply to the size distribution of the Spanish state's UAs. While among the biggest UAs Barcelona is approximately half the size of Madrid, the remaining UAs appear to be more uniformly distributed than Zipf's law would suggest, pointing towards a more polycentric distribution on the lower end [81]. The distribution of skill across UAs is less skewed than that of population, as also shown in Rossi-Hansberg et al. (2019) or Glaeser and Resseger (2010). Interestingly, it is the case in the Spanish state's economy the correlation between city size and skill is entirely driven by the large UAs of Madrid and Barcelona [82].

[81] Note that the largest 20 UAs are selected for this exercise, and the overall fit of Zipf's law may depend on the chosen cut-off (Eeckhout, 2004).

[82] The overall correlation between city size and average skill is 0.6 in the data. However, this is entirely driven by the large UAs. Once Madrid and Barcelona are taken out, the correlation becomes negative, at -0.26. This stands in contrast with Glaeser and Resseger (2010), where a positive relationship is found.

Figure E.23: Population (left panel) and Skill (right panel) distributions in the Data (MCVL) and Model.



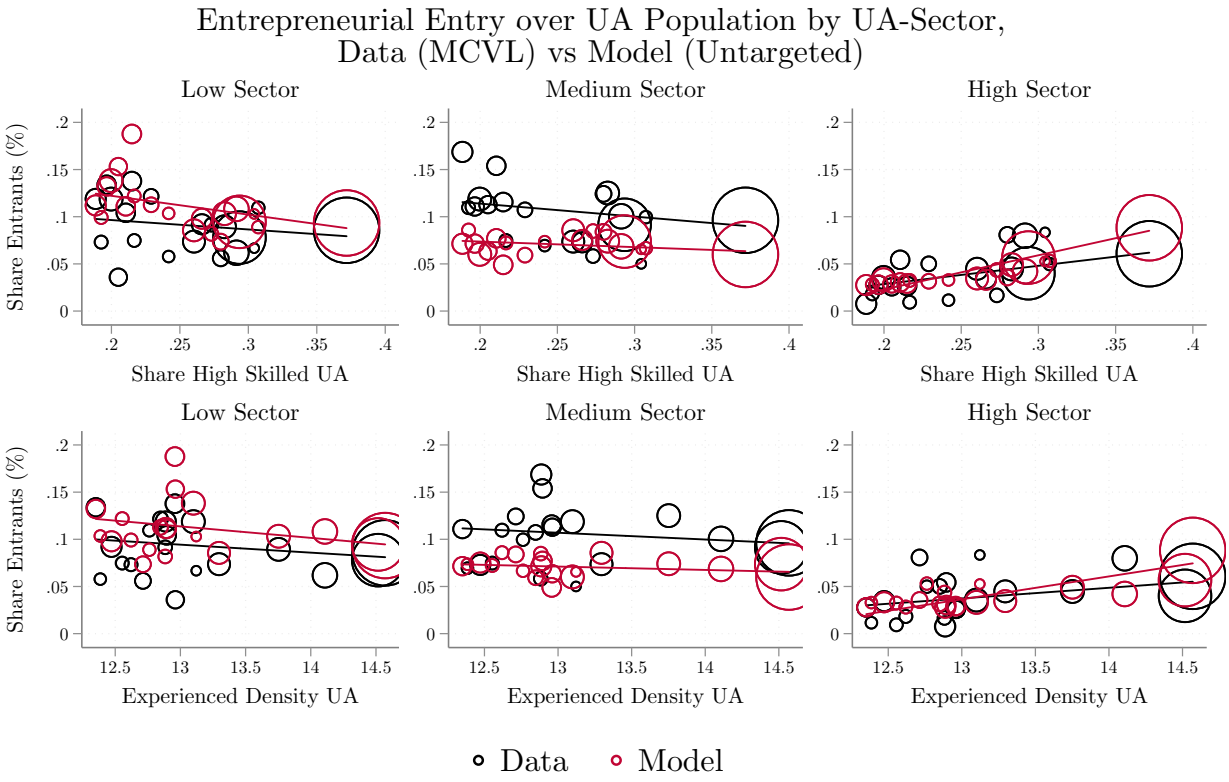
E.6 Entrepreneurial Entry

The Entrepreneurial Entry is defined as the fraction of new entrepreneurs (those transitioning from being workers to becoming entrepreneurs) over the population of the UA. Figure $(E.24)$ presents the results obtained in the model.

The interpretation of the panels is similar to figure (6), this time with the y-axis showing the fraction entrepreneurs by UA-Sector out of the population of each UA that are new entrants (those transitioning from workers to entrepreneurs). Thus, value of say 0.2% would indicate that out of the UA population in location l, 0.2% of the population transitioned from being a worker into entrepreneurship in a given sector.

Overall, the results are reminiscent of those in figure 6. There exists a negative relationship between the agglomeration proxies and entry into entrepreneurship both in the data and model for the low and medium sectors, while the relationship flips again in the case of the high sector.

Figure E.24: Entrepreneurial Entry: Data and Model



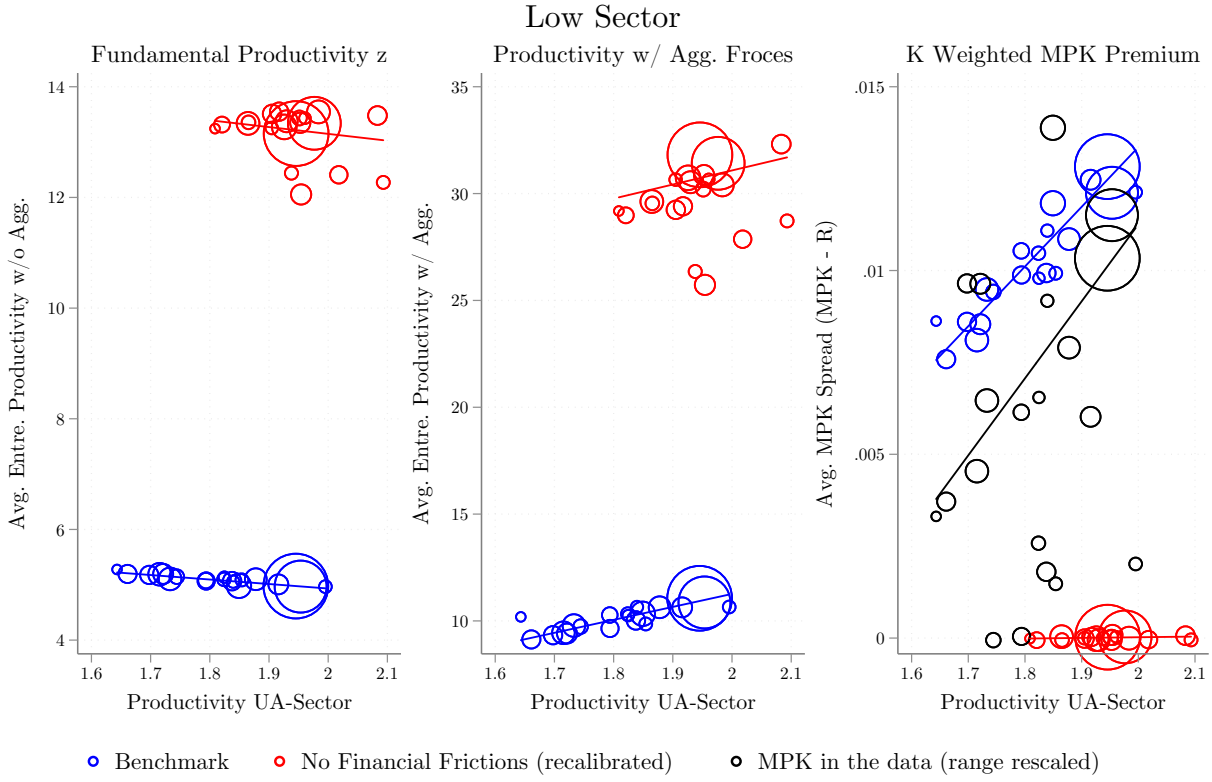
Notes: Entrepreneurial entry by UA-sector over the population of the UAs. Each column corresponds to one sector of the economy. The rows show the relationship between the entrepreneurial entry and a given variable of interest (x-axis). The y-axis is in percentage terms, and denotes the share of new entrepreneurs in a given UA-sector over the population of that UA-sector. The black lines and dots refer to the actual data and the red are the model counterparts. Fit lines are weighted by the population of each UA. The overall share of entrants in the economy is targeted, while slopes are untargeted. Dots are scaled by the population of the UA.

F Appendix to the Steady State Results

F.1 Heterogeneous Returns to Capital in the Low and Medium Sectors

This section corroborates the results in figure 8 for the Low and Medium sectors.

Figure F.25: Steady-State Results: Composition of Productivity by UAs and MPK Premiums, Low Sector.

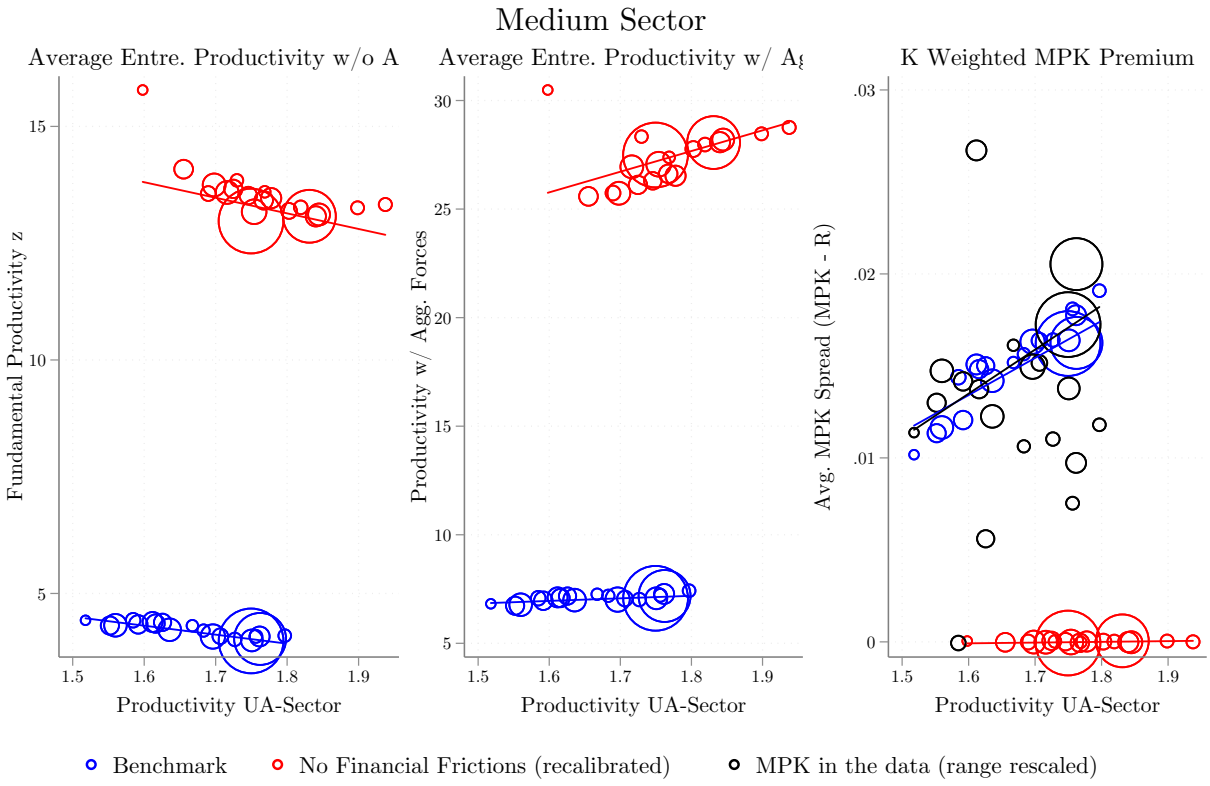


Notes: Steady-state results. Each panel displays one variable of interest and its relationship to UA-sector productivity. The size of the circles represents the size of the UAs. The first panel shows the average fundamental productivity (z) of entrepreneurs (unweighted) by UA-sector. The second shows the average productivity of entrepreneurs once agglomeration forces are included (unweighted). The last shows the MPK Premium by location. All results are for the low skilled sector. The results are displayed for the benchmark economy and the recalibrated no financial frictions counterfactual.

F.2 Profile of the Entrepreneurs in the Model across Space

In section 6, we discussed that the main force behind the obtained positive relationship between the productivity of the UAs and the average MPK Premium was the intensive margin, while there was no clear impact of the extensive margin. This section aims to provide further details on this

Figure F.26: Steady-State Results: Composition of Productivity by UAs and MPK Premiums, Medium Sector.

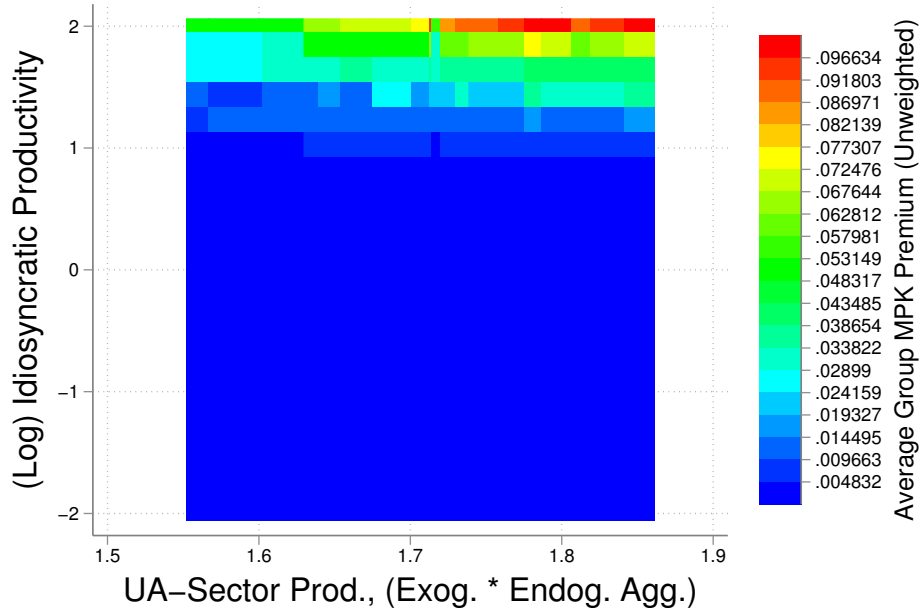


Notes: Steady-state results. Each panel displays one variable of interest and its relationship to UA-sector productivity. The size of the circles represents the size of the UAs. The first panel shows the average fundamental productivity (z) of entrepreneurs (unweighted) by UA-sector. The second shows the average productivity of entrepreneurs once agglomeration forces are included (unweighted). The last shows the MPK Premium by location. All results are for the medium skilled sector. The results are displayed for the benchmark economy and the recalibrated no financial frictions counterfactual.

finding. Figure F.27 displays the obtained MPK premiums by idiosyncratic productivity - UA combination in the model, while F.28 shows the mass at each idiosyncratic productivity - UA pair, conditioned by UA.

As we can observe, the average MPK of the marginal entrepreneurs does not vary substantially across UAs. At the same time, the relative mass of the marginal entrepreneurs shows no clear trend either. This suggests that the extensive margin plays no key role and therefore the observed heterogeneous returns to capital across locations are the product of the differences across the intensive margin. In fact, we can see clearly that as we move along the x-axis the average MPK premiums increase for every idiosyncratic productivity level (recall that the agglomeration

Figure F.27: MPK Premium (Unweighted) Distribution by Idiosyncratic Productivity - UA pair, Active Entrepreneurs Only.



Notes: Average unweighted MPK Premiums by idiosyncratic productivity - UA pair, for active entrepreneurs only. The x-axis denotes the average effective productivity (agglomeration forces included) of the UA across all sectors weighted by the revenue share of each. The y-axis denotes the log productivity of the entrepreneurs. The heatmap shows the unweighted average MPK Premiums for each idiosyncratic productivity - UA pair. Note that this is the average across all sectors and asset levels.

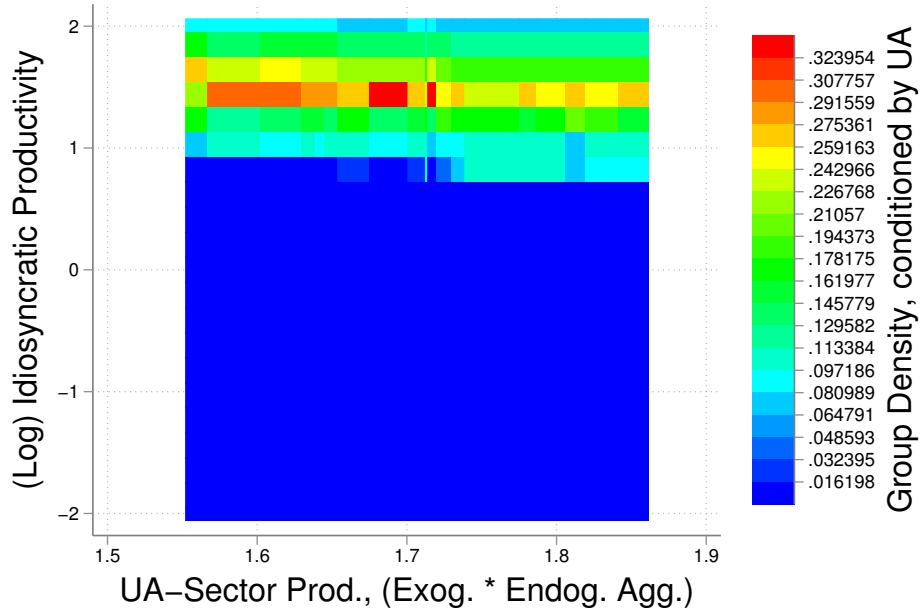
forces increase the effective productivity for a given idiosyncratic productivity). Since the relative masses of each group remain sufficiently similar, this intensive margin is the key driver.

F.3 Productivity distributon across UA-Sectors

In Combes et al. (2012), the authors, using rich French firm data, argue that selection forces can not explain the observed differences across locations, pointing to agglomeration forces as the main driver. Is that also the case in this model? Figure $(F.29)$ presents the idiosyncratic and effective (agglomeration forces inclusive) productivities in the model.

There are two main results worth discussing. First, one may see that by looking at the left figure the model features *negative* selection in terms of idiosyncratic productivity as the productivity of the UA increases. Second, once we focus on the effective productivity however (the one estimated in the data), the results are in line with those in Combes et al. (2012). Selection alone does not

Figure F.28: Mass of the Distribution by Idiosyncratic Productivity - UA pair, Active Entrepreneurs Only. Conditional distributions by UA.

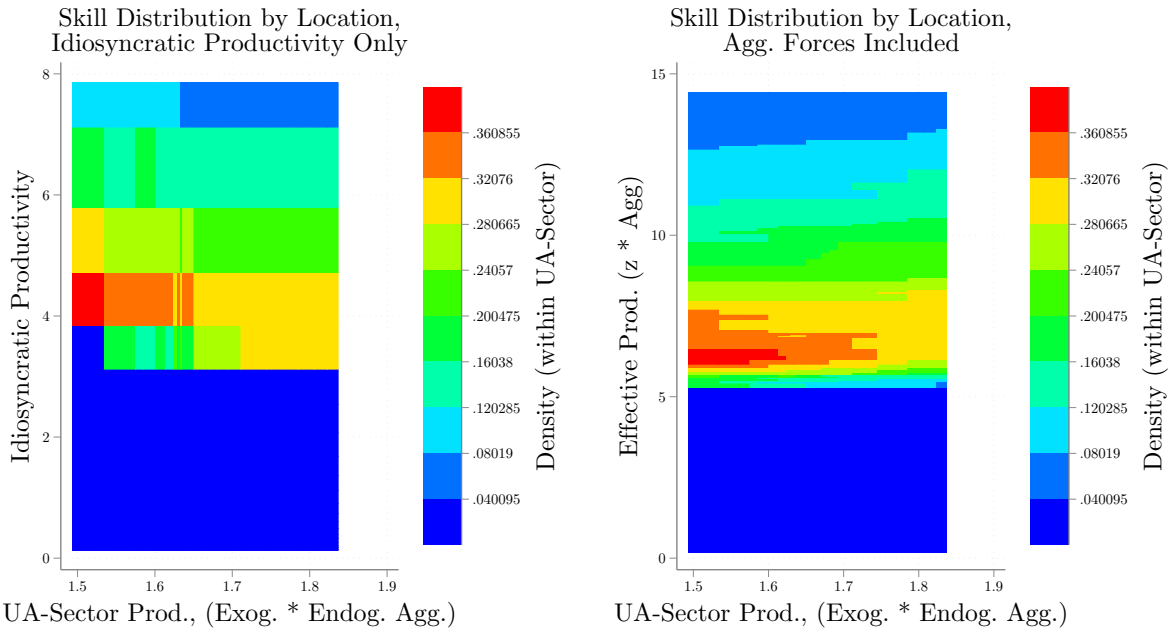


Notes: Mass of the distribution at each idiosyncratic productivity - UA pair, for active entrepreneurs only. The x-axis denotes the average effective productivity (agglomeration forces included) of the UA across all sectors weighted by the revenue share of each. The y-axis denotes the log productivity of the entrepreneurs. The heatmap shows the mass of the conditional by UA distribution for each idiosyncratic productivity - UA pair. Note that this is the average across all sectors and asset levels.

seem to explain the variation, as no subsntantial left-truncation is observed. Instead, as found by the authors, a right shift (entrepreneurs are on average more productive in more productive UAs) and dilation (the right tail of the distribution widens as we move toward more productive UAs) explains the differentials in productivities across cities.

To sum up, we discussed how the profit shares are increasing in UA-Sector productivities for a given point on the state space, how this leads to average heterogenous returns to capital across space and how the variation in effective productivites across locations is mainly explained by the agglomeration forces.

Figure F.29: Conditional distributions of Idiosyncratic and Effective Productivities by UA (High Sector)



Notes: Conditional spatial distribution of entrepreneurial idiosyncratic and effective (agglomeration forces inclusive) productivities by UA. The x-axis denotes the productivity of the UA-Sector. The y-axis denotes either the idiosyncratic (left panel) or effective (right panel) productivities. The colorbar shows the conditional mass per UA.

G Appendix Policy Exercises

G.1 Total Welfare Responses by Group

In the main body of the paper we presented the per capita welfare responses, defined as $\frac{\frac{\sum w^P V^P}{\sum w^P} - \frac{\sum w^{SS} V^{SS}}{\sum w^{SS}}}{\frac{\sum w^{SS} V^{SS}}{\sum w^{SS}}}$, where the w denote the weights in the invariant distribution over the state space of the relevant group (all population, workers or entrepreneurs) at either the steady-state (SS) or policy exercises (P) and the V stand for the value subject to the same observations. In this Appendix, we present the results for the alternative definition $\frac{\sum w^P V^P - \sum w^{SS} V^{SS}}{\sum w^{SS} V^{SS}}$.

The main difference between both is that while the first captures variation in per capita responses, the latter captures the aggregate response of the group. Thus, if the marginal entrepreneur that has entered in UA l is less productive, the former would decrease while the second will increase, as the aggregate stock of welfare held by entrepreneurs in l now increases due to the larger mass of the group (assuming the continuing entrepreneurs don't experiment a substantial welfare loss, which is unlikely given the entry of more inefficient types). Likewise, if after the policy counterfactual l has lost population and has retained the most skilled, the per capita welfare will increase while the aggregate stock of welfare held by the group will now decrease (assuming that the extensive margin loss dominates over the intensive margin gains of the ones remaining).

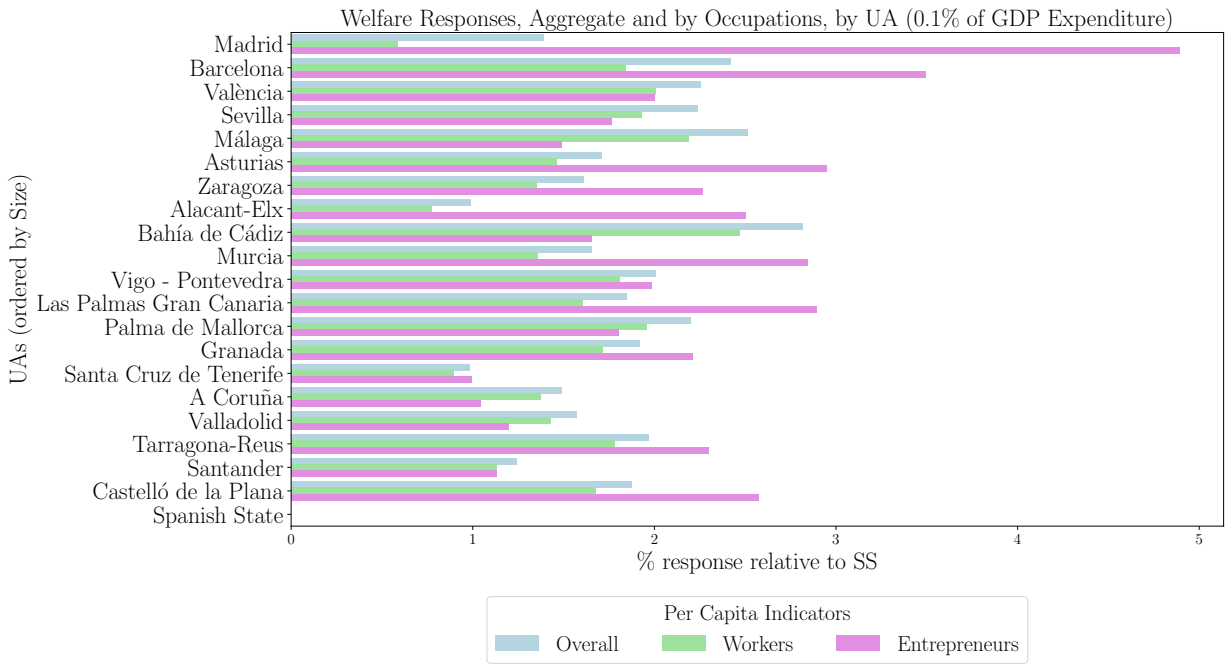
G.1.1 Total Welfare by UA in the Country-wide Policy

Figure G.30 shows the responses of welfare by location to the country-wide policy by UA according to this alternative definition. An obvious difference is the positive response of the welfare of the entrepreneurs. As discussed above, since this measure captures the overall stock of welfare held by each group, the entry of marginally less productive entrepreneurs still leads to an increase in this measure.

G.1.2 Total Welfare by UA in the Single-targeting Exercises

In the case of the single-targeting of UAs, we can now see that under this alternative definition of welfare response entrepreneurs benefit from the policy even when targeting UAs that lead to

Figure G.30: Response of Total Welfare, Aggregate and by Occupation, to an Untargeted Policy (0.1% of Country-wide GDP Expenditure)



Notes: Response of total welfare, aggregate and by occupations, to a Country-wide untargeted lump-sum transfer T such that total programme expenditure is 0.1% of GDP, by Urban Area. Urban Areas are ordered from largest in population size (Madrid) to smallest (Castelló de la Plana). The last row indicates the Country-wide response. Blue bars indicate the response of the total welfare by UA. Green bars do likewise for the total welfare response of workers. Lastly, the violet bars indicate the total welfare response of entrepreneurs. The welfare function under consideration is the utilitarian one, SWF = $\int V(\cdot)d\lambda(\cdot)$.

aggregate welfare losses.

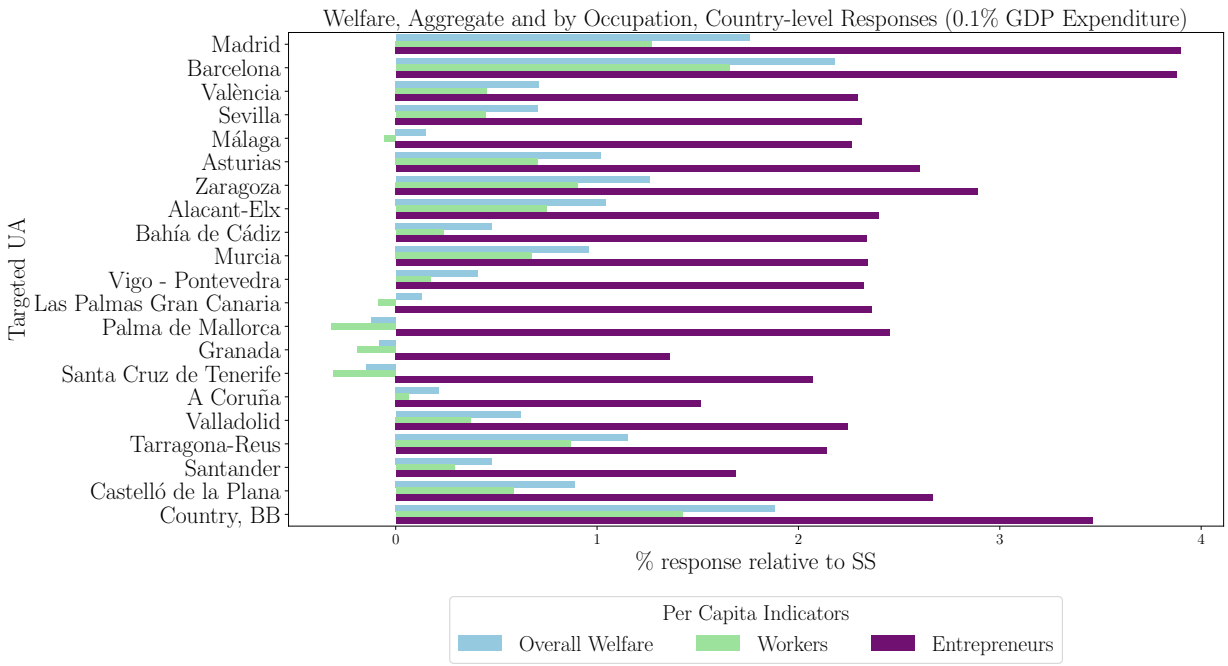
G.1.3 Total Welfare by UA in the Cumulative-targeting Exercises

G.2 Proportional Collateral Transfers

The policy in the previous section was inherently uniform across entrepreneurs, as everyone obtained the same amount T_l of funds. However, since the key friction in the economy are the borrowing constraints, $k < \lambda_j a$, a natural policy would be to consider a programme that alleviates this frictions by providing collateral.

In particular, the exercise we will be considering is a set of entrepreneur and location specific transfers $T_{e,l}\ \forall\ e \in L_p$ so that the constraint at the individual level $k < \lambda_j a$ becomes $k < (1+\phi)\lambda_j a$.

Figure G.31: Response of Country-wide Total Welfare, Aggregate and by occupation, to Place-based Entrepreneurial Transfers (0.1% of Country-wide GDP Expenditure)



Notes: Response of the country-wide total welfare, aggregate and by occupations, to place-based entrepreneurial transfers T_l such that total programme expenditure is 0.1% of GDP. The y-axis denotes which UA is being targeted by the policy. Blue bars indicate the response of the country-wide total welfare. Green bars show the response of the total welfare of workers. Violet bars display the total welfare response of entrepreneurs.

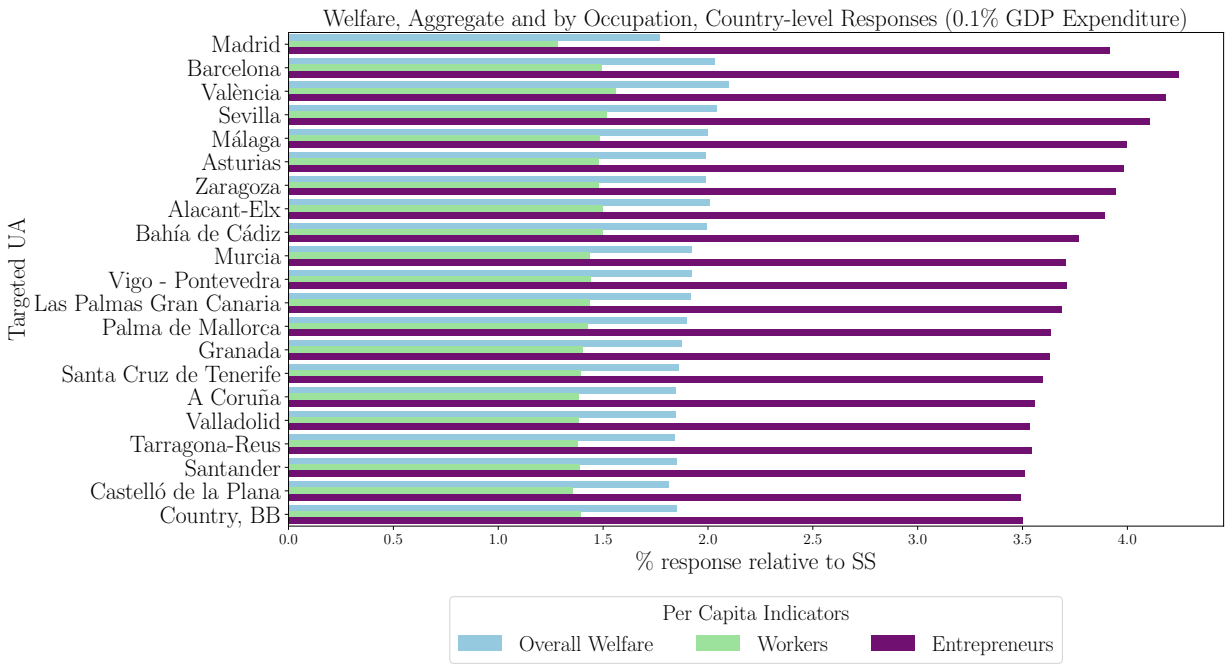
That is, entrepreneurs will receive at most ϕa assets from the government [83] in order to finance an additional $\phi\lambda_j a$ units of capital. The constraint of the government therefore reads:

$$\underbrace{\sum_{l=1}^{L}\sum_{j=1}^{J}\sum_{s=1}^{S}(\tau_L+\delta_P)w_{l,j,s}LS_{l,j,s}}_{\text{Revenues from Labour}}+\underbrace{\sum_{l=1}^{L}\sum_{j=1}^{J}\tau_K TBK_{l,j}}_{\text{Revenues from Profits}}=\underbrace{G^{SS}Y}_{\text{Exogenous Expenditure}}+\underbrace{\sum_{l,e\in L_P}T_{e,l}S_{e,l}}_{\text{Transfers Programme}}$$

As in the case with the lump-sum transfers, these proportional to collateral transfers are financed through an increase on the labour tax δ_p.

[83] In the case where the entrepreneur would be unconstrained with a partial amount of the resources only this partial amount is transferred.

Figure G.32: Response of Country-wide Total Welfare, Aggregate and by Occupation, to **Cumulative** Place-based Entrepreneurial Transfers (0.1% of Country-wide GDP Expenditure)



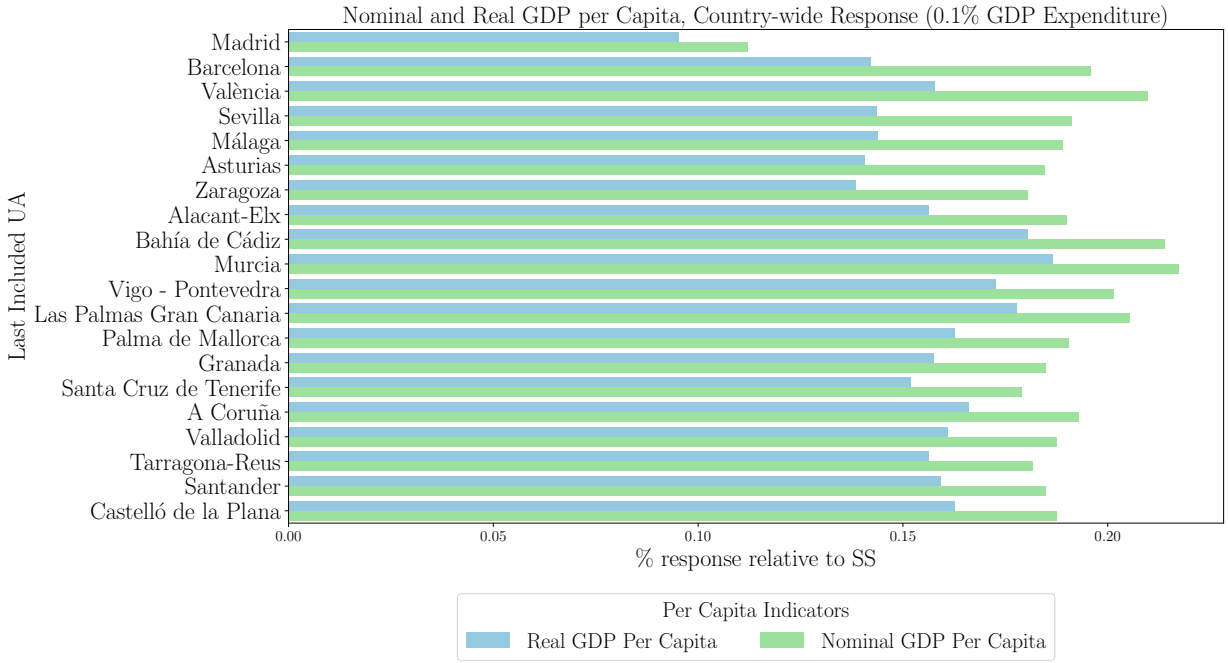
Notes: Response of the country-wide total welfare, aggregate and by occupations, to **cumulative** place-based entrepreneurial transfers T_l such that total programme expenditure is 0.1% of GDP. The y-axis denotes which UA is being included last in the set of UAs being targeted. Blue bars indicate the response of the country-wide total welfare. Green bars do likewise for the workers. Violet bars denote that of entrepreneurs. The last row is equivalent to targeting the entire country, since all UAs are included.

G.2.1 Cumulative Responses under the Proportional Transfers Programme

In this section we will study the long-run response of welfare and efficiency to these proportional transfers programme and compare the results with those of the lump-sum transfers programme. Figure G.33 presents the responses of efficiency under this programme:

A few comments are in place. First, one may observe that the overall magnitude of the responses is an order of magnitude smaller than that of the lump-sum transfers for the same expenditure level. While the untargeted lump-sum policy for instance led to a 1.2% increase in country-wide real GDP per capita, the proportional transfers lead to an increase of 0.16%. This is a direct result of the distribution of funds over entrepreneurs with this policy. Since, *ceteris paribus*, wealthier entrepreneurs are less constrained, what this policy effectively does is to concentrate the funds on the region of the state space where the returns to them are the lowest.

Figure G.33: Response of Country-wide Nominal and Real GDP per Capita to **Cumulative** Place-based Entrepreneurial **Proportional** Transfers (0.1% of Country-wide GDP Expenditure)



Notes: Response of the country-wide nominal and real (deflated by the local price index $p_{h,l}^{(1-\alpha_c)}$) GDP per capita to **cumulative** place-based entrepreneurial **proportional** $T_{e,l}$ such that total programme expenditure is 0.1% of GDP. The y-axis denotes which UA is being included last in the set of UAs being targeted. Blue bars indicate the response of the country-wide per capita real GDP while green bars do likewise for the nominal GDP per capita. The last row is equivalent to targeting the entire country, since all UAs are included.

Second, it is worthwhile to note that the "optimal" policy still targets a subset of UAs rather than the entire country. Since the marginal returns to these transfers now decrease rapidly, rather than spending all the resources on the wealthiest entrepreneurs in a single location the results suggest spreading this resources among wealthy entrepreneurs in several locations.

While the policy leads to an overall efficiency gain of 0.186% at the optimal set of UAs, regional disparities increase by 0.4% which is comfortably the worst relative ratio among all considered policies. Not only are the largest entrepreneurs receiving the bulk of the subsidies, it is also the largest entrepreneurs in the largest UAs that are concentrating most of the funds.

What would be the welfare implications of such policy?

The results point to a negative welfare response across the board. The intuition is simple: workers in the economy are paying an additional δ_p on the labour income tax to subsidize a policy

Figure G.34: Response of Country-wide per Capita Welfare, Aggregate and by Occupation, to **Cumulative** Place-based Entrepreneurial **Proportional** Transfers (0.1% of Country-wide GDP Expenditure)

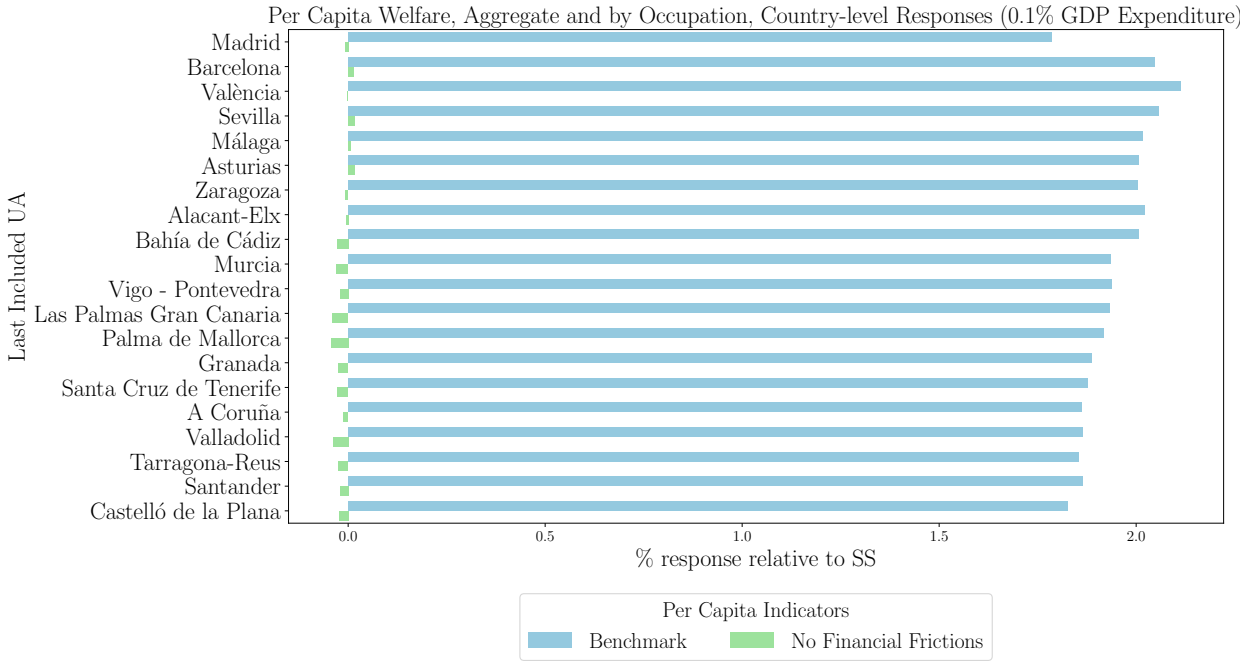
Notes: Response of the country-wide per capita welfare, aggregate and by occupations, to **cumulative** place-based entrepreneurial **proportional** transfers $T_{e,l}$ such that total programme expenditure is 0.1% of GDP. The y-axis denotes which UA is being included last in the set of UAs being targeted. Blue bars indicate the response of the country-wide per capita welfare. Green bars do likewise for the workers. Violet bars denote that of entrepreneurs. The last row is equivalent to targeting the entire country, since all UAs are included.

that has very uneven impact across the state space (large entrepreneurs concentrate most of the funds), very limited impact on efficiency (wage increases are very low) and almost no impact on the extensive margin entry decisions.

These results have important policy implications for investment subsidies, loans, or any other policy that is size dependent. If there are reasons to believe that larger firms have easier access to capital (Gilchrist et al. (2013), Shen (2023)) and that financial frictions are relevant ((Evans and Jovanovic, 1989), (Holtz-Eakin et al., 1994), (Banerjee and Duflo, 2014), (Schmalz et al., 2017)), then policies that target the region in the domain where entrepreneurs are more constrained are more likely to lead to both efficiency and welfare gains.

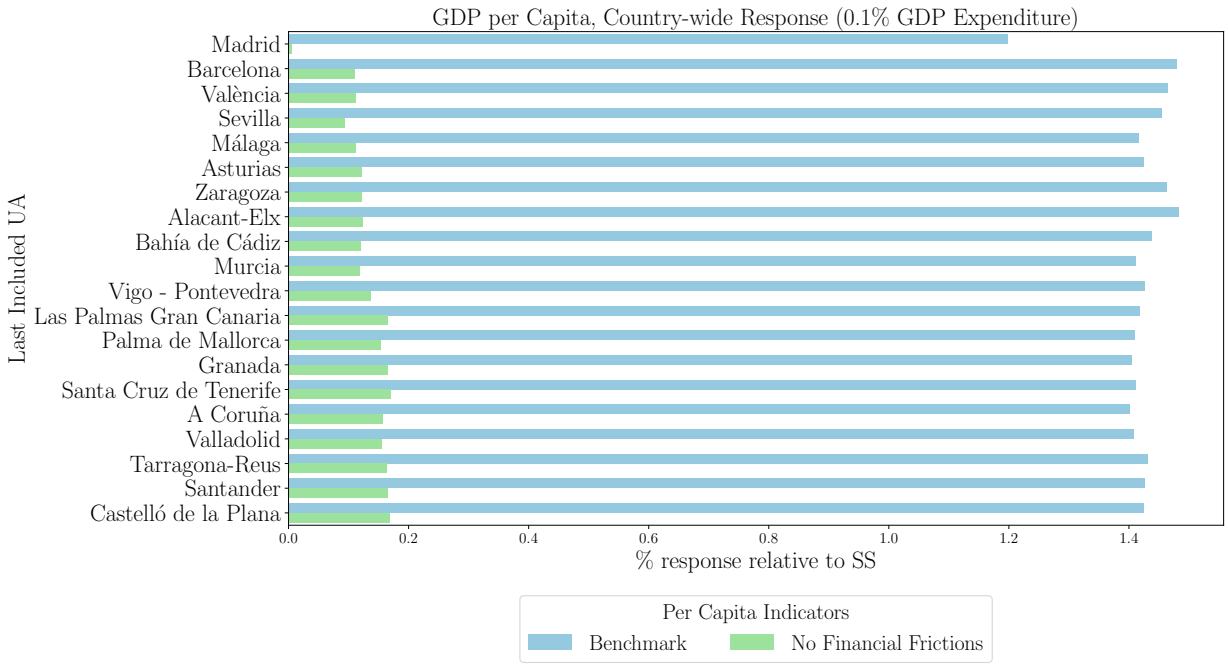
G.3 The Role of Financial Frictions

Figure G.35: Response of Country-wide per Capita Welfare, to **Cumulative** Place-based Entrepreneurial Transfers (0.1% of Country-wide GDP Expenditure), in the Benchmark Economy and the Counterfactual without Financial Frictions



Notes: Response of the country-wide per capita welfare to **cumulative** place-based entrepreneurial transfers $T_{e,l}$ such that total programme expenditure is 0.1% of GDP. The y-axis denotes which UA is being included last in the set of UAs being targeted. Blue bars indicate the response of the country-wide per capita welfare in the Benchmark Economy. Green bars do likewise for the No Financial Frictions counterfactual. The last row is equivalent to targeting the entire country, since all UAs are included.

Figure G.36: Response of Country-wide Nominal GDP per Capita to **Cumulative** Place-based Entrepreneurial Transfers (0.1% of Country-wide GDP Expenditure) under the Benchmark Economy and the Counterfactual without Financial Frictions



Notes: Response of the country-wide nominal GDP per capita to **cumulative** place-based entrepreneurial $T_{e,l}$ such that total programme expenditure is 0.1% of GDP. The y-axis denotes which UA is being included last in the set of UAs being targeted. Blue bars indicate the response of the country-wide per capita nominal GDP in the Benchmark economy. Green bars do likewise for the No Financial Frictions counterfactual.