Abstract

We consider an economy where the production technology has constant returns to scale but where in the decentralized equilibrium there are aggregate increasing returns to scale. The result follows from a positive contracting externality among firms. If a firm is surrounded by more firms, employees have more opportunities outside their own firm. This improves employees' incentives to invest in the presence of ex post renegotiation at the firm level, at no cost. Our leading result is that if a region is sparsely populated or if the degree of development in the region is low enough, there are multiple equilibria in the level of sectorial employment. From the theoretical model we derive a non-linear first-order censored difference equation for sectoral employment. Our results are strongly consistent with the multiple equilibria hypothesis and the existence of a sectoral critical scale (below which the sector follows a delocation process). The scale of the regions' population and the degree of development reduce the critical scale of the sector.

JEL Classification: D23, J41, O18, R12.

Contracting Externalities and Multiple Equilibria in Sectors: Theory and Evidence

Diego Rodríguez-Palenzuela[*][†]
Neus Bover-Fonts[‡]
Juan José de Lucio[§]

November 1998

[*]Universitat Pompeu Fabra. I gratefully acknowledge financial support from the Spanish Ministry of Science PB93-0398, and from Centre per la Recerca de l'Economia Internacional, CREI, at Universitat Pompeu Fabra.

[†]We thank Antonio Cabrales, Gary Charness, Antonio Ciccone, Patrik Rey and Konrad Stahl for helpful conversations and seminar participants at UPF and at the 1997 ASSET Meeting. The usual disclaimer applies.

[‡]Caminus Research

[§]Consejo Superior de Cámaras de Comercio.

1 Introduction

There has been in the last years increasing interest in models of the agglomeration of economic activity. The largest fraction of papers that have attempted to explain agglomeration have recurred to explanations based on fixed costs of starting economic activities, that is, based on aggregate or disaggregate technological increasing returns to scale. For an introductory survey to theories of localization of economic activities, see chapter 2 in Krugman (91).

A second branch of the literature has focused on externalities that do not affect production technologies but affect transaction technologies. In Acemoglu (96) search externalities arising from decentralized factor markets lead to benefits of agglomeration, in a context where agents make their factor investments before matching takes place. Acemoglu (97) generalizes the previous results and introduces an endogenous innovation decision. The combination of transacting externalities and the complementarity of investments at the firm level implies the multiplicity of Pareto ranked equilibria.. Importantly, multiplicity of equilibria does not arise in the benchmark frictionless economy.

Our paper belongs to this second branch of the literature, with transacting externalities. In the theoretical side we construct a source of contracting externalities that have not been considered before. In the empirical side we focus on the existence of multiple equilibria in sectors and regions, since multiplicity of equilibria is the leading implication of models with transacting externalities.

We consider an economy where the production technology has constant returns to scale but where in the decentralized equilibrium there are aggregate increasing returns to scale. The result follows from a positive *contracting* externality among firms. If a firm is surrounded by more firms, employees have more opportunities outside their own firm. Increasing outside options improves employees' incentives to invest in the presence of ex post renegotiation at the firm level, *at no cost*: outside opportunities are not used in equilibrium.

In our model there are no direct or technological externalities across firms.

Our theory starts from a simple observation. As the scale of an economy increases, the number of inefficient reallocations of factors increases exponentially. Inefficient reallocations serve as outside options to agents when bargaining for the terms of trade. Although these options do not directly contribute to social welfare, they change agents' incentives to engage in human capital investments. With transaction costs deriving from inefficient ex post bargaining, we find that in the decentralized equilibrium there are generally increasing returns to scale.

In summary, if and only if transaction costs (arising from agents' inability to commit on future wages) matter, outside opportunities have an option value for each of the players. Outside opportunities replicate naturally as the size of the economy increases, but players' competition to take them remains constant (and nil), since they are a worse alternative than staying in the original team. To repeat, under transaction costs, there is a benefit that freely accrues to each player as the economy grows: her options outside her team. This improves her incentives to invest and, if players' investments are complementary, the effect is mutually reinforced.

We lay out a model that shows this mechanism precisely. To keep the model tractable, we introduce specific assumptions on the production function. We distinguish between *team production*, and *ex post production*. Team production is the output of a team that "trained jointly" and whose team members sunk investments in human capital that are specific to physical assets. In particular, we introduce assumptions such that the optimal size at team production is equal to two members. This allows us to avoid complex multilateral bargaining solutions like the Shapley Value or the one in Stole and Zwiebel (96).

Ex post production means that firms can hire additional workers (beyond the optimal size of two) at the production stage. But marginal returns to human capital inside the firm are decreasing, and the *ex post workers* have a *low* marginal productivity. Ex post production does not occur in equilibrium, since it is a worse option for any team member, but it is a credible outside option.

Our leading result is that if a region is sparsely populated or if the degree of development in the region is low enough, in the industrial sectors of the region there are multiple equilibria in the level of total sectoral employment. One of this equilibria corresponds to the absence of sectoral activity. On the other hand, if the scale of the region (in terms of total population) is large enough or the level of development is high enough, there is a unique employment level equilibrium in the industrial sectors of the region. In the unique equilibrium there is positive industrial activity. This result is a consequence of the reinforcement effect between total activity and incentives to invest inside the firm. In particular, outside activity is shown to relax an incentive compatibility (IC) constraint to exert effort of the *worker* (not asset owner), without affecting the constraint of the *manager* (asset owner). This asymmetry between the effect of total activity on the two IC constraints is due precisely to the fact that the effect works through the outside opportunities.

From the theoretical model we derive a econometrically tractable equation of the evolution of sectoral employment in regions. We specify a non-linear first-order censored difference equation for sectoral employment. The existence of a unique or multiple equilibria has implications for the parameter values of the difference equation. Our results are strongly consistent with the multiple equilibria hypothesis. The average sector in Spanish regions has three steady states. One stable steady state is the absence of activity (no sectoral employment); the other stable steady state has a positive employment level, greater than the level of the unstable steady state. To the estimated unstable steady state we call the critical mass of the sector. If in a region a shock sets the sectoral employment level below the critical mass, the industry follows a delocation process that ends with the destruction of sectoral employment in the region.

We find moreover that there are significant differences across Spanish regions. The more developed, densely populated regions of Barcelona and Madrid are found to have a stable steady states with greater sectoral size, and smaller critical sizes. In particular for Barcelona we find that there is not a critical mass and therefore there is not the possibility of delocation. For ten sparsely populated and less economically developed regions we find that the critical mass of the average sector ranges from 600 to 900 employees.

These results are confirmed when substitute regional fixed-effects for regions' characteristics in our tobit estimation. This second estimation shows that economic development *or* scale of the population are each sufficient to reduce the probability of industrial delocation.

It is important to emphasize that our empirical findings can be explained by other theories of aggregate increasing returns, different than our theory based on contracting externalities. At this stage we make no attempt to empirically select the best theory to explain the data. We use the model as a frame to yield an econometric specification and to interpret the results. We conclude that our empirical results are consistent with theories of aggregate increasing returns, like the one we propose, but they are inconsistent with models with complete contracts and constant returns.

The rest of the paper is organized as follows. Section 2 lays out the model and the welfare-maximizing allocation. Section 3 introduces our assumptions on transaction costs and derives the decentralized equilibrium. In section 4 the main empirical implications of the model are discussed and the data set is described. Section 5 shows the results from the maximum likelihood estimation. The paper closes with section 6 which includes the final remarks.

2 The Model

There is a continuum of heterogeneous, risk neutral agents in the economy. Agents are indexed by a an efficiency parameter θ, uniformly distributed in $[\underline{\theta}, \overline{\theta}]$, with $\overline{\theta} - \underline{\theta} = 1$. The total mass of agents is N.

Production requires one unit of physical capital, k that costs c_k[1] and human capital, e, that is a discrete variable $(e \in \{\underline{e}, \overline{e}\})$. Effort can either be high $(e = \overline{e})$ or low $(e = \underline{e} \in (0, \overline{e}))$. The cost of exerting high effort for an agent of type θ is $c(\overline{e}, \theta) > c(\underline{e}, \theta) = 0$, with $c_\theta(\overline{e}, \cdot) < 0$, $c_{\theta\theta}(\overline{e}, \cdot) < 0$.

Initially firms are formed by two agents who acquire one unit of physical

[1] Recall that the "units" of physical and human capital are arbitrarily small. c_k is the cost of a marginal unit of physical capital. We assume that the production technology has constant returns to scale (see below).

capital (at cost c_k) and exert effort. Effort is an investment in human capital that is specific to the physical asset and to the team where it originated. This implies that at the production stage an agent is relatively more productive at her initial firm than at a different firm with similar characteristics. In order to model specificities as simply as possible we assume that the production function has two distinct components, depending on whether the agents that produce are part or not of the initial team. The first component is *team production*, y. Call e_θ the human capital of agent θ; then $y\left(e_\theta, e_{\theta'} \mid \theta, \theta'\right)$ is the initial team's output of two agents (θ, θ') that pool their human capital $(e_\theta, e_{\theta'})$ at the production stage.

The second component of the production function is *ex post production.* This is the output of employees $(\theta'' \notin \{\theta, \theta'\})$ who do not belong to the initial team, but who may join the firm at the production stage.

2.1 Team production

Team production y depends on the identity of the agents in the firm at the investment stage $(\theta_i, \theta_j, ...)$ and on the human capital of the set of agents present at the production stage $(e_{\theta_k}, e_{\theta_l}, ...)$: $y = y\left(e_{\theta_k}, e_{\theta_l}, ... \mid \theta_i, \theta_j, ...\right)$. The sets of initial members $(\theta_i, \theta_j, ...)$ and the set of ex post present members $(\theta_k, \theta_l, ...)$ are not necessarily the same. If agent θ is a member of a team but is not present at the production stage, we write $e_\theta = 0$.

We assume that the team production function y is symmetric and that the low effort level is also productive. For all θ, θ':

$$y\left(\underline{e}, \overline{e} \mid \theta, \theta'\right) = y\left(\overline{e}, \underline{e} \mid \theta, \theta'\right) \geq y\left(\overline{e}, 0 \mid \theta, \theta'\right) \geq y\left(\underline{e}, 0 \mid \theta, \theta'\right) > 0$$

where $e_\theta = 0$ indicates that agent θ is not present in the firm at production stage.

On the other hand, if θ'' is not a team member her effort does not affect team production[2] : for all $e_{\theta''}$: $y\left(e_\theta, e_{\theta'}, e_{\theta''} \mid \theta, \theta'\right) = y\left(e_\theta, e_{\theta'} \mid \theta, \theta'\right)$.

We assume that the team-production function y satisfies a strict complementarity condition: the output in one firm when all agents exert high effort

[2] Although it affects ex post production.

is greater than the sum of the outputs when only one exerts high effort and the other exerts low effort[3].

$$y\left(\overline{e},\overline{e}\mid\theta,\theta'\right)>y\left(\overline{e},\underline{e}\mid\theta,\theta'\right)+y\left(\underline{e},\overline{e}\mid\theta,\theta'\right) \tag{1}$$

We assume further that when at least one team member exerts only low effort the firm is not viable:

$$y\left(\overline{e},\underline{e}\mid\theta,\theta'\right)<c\left(\overline{e},\overline{\theta}\right)+c_k \tag{2}$$

2.2 Ex post production

If an agent θ'' that is not a member of the original team (θ,θ') joins the firm at the production stage, the total production of the firm, f, is the sum of team production, y, and ex post production[4]:

$$f\left(e_{\theta},e_{\theta'},e_{\theta''}\mid\theta,\theta'\right)=y\left(e_{\theta},e_{\theta'}\mid\theta,\theta'\right)+\phi e_{\theta''}$$

where $\phi<1$ is interpreted as the fact that agent θ'' is less productive outside her initial team due to human capital specificities.

Introducing the possibility of ex post production implies that if a member θ of a given team (θ,θ') quits, she has a positive productivity in other existing firms. We assume that ex post production is not viable ex ante even for the most efficient agent:

$$\phi e_{\theta''}\leq c\left(e_{\theta''},\overline{\theta}\right) \tag{3}$$

2.3 First-best allocation

Given the mass of agents in the economy N, the planners' problem is to introduce human capital e and physical capital k as long as the marginal firm contributes non-negatively to social surplus. Clearly, under the first-best there is only team production and there is not ex post production.

[3] This assumption is satisfied for instance by the Cobb-Douglas production function $y=\left(e_{\theta}+\beta\right)^{\alpha}\left(e_{\theta'}+\beta\right)^{\alpha}\left(\overline{k}\right)^{1-2\alpha}$ with $\beta>1$.

[4] The number of vacancies for ex post production in each firm could be arbitrarily large and we would yield identical implications.

Call $y^* = y(\overline{e}, \overline{e} \mid \theta, \theta')$ the teams' output when both agents exert effort and remain in the team. The total number of 2-member teams or firms in the economy is x and the number of firms "per capita", $\widetilde{x} \equiv x/N$. Social welfare is:

$$SW(N,x) = N\left(\int_{\left(\overline{\theta}-\frac{2x}{N}\right)}^{\overline{\theta}} \left[\frac{1}{2}(y^* - c_k) - c(\overline{e}, s)\right] ds\right) \equiv N\sigma(\widetilde{x})$$

from which it follows that the optimal number of firms $\widetilde{x}^*$ per capita is independent of N and satisfies:

$$(y^* - c_k) = 2c\left(\overline{e}, \overline{\theta} - 2\widetilde{x}^*\right) \tag{4}$$

Given N, the social welfare is a concave function of the total number of firms introduced, $x : SW_{xx}(N,x) = \frac{4}{N}\frac{\partial c\left(\overline{e}, \overline{\theta}-2\widetilde{x}\right)}{\partial\theta} < 0$, for $\widetilde{x} \in \left[0, \frac{\overline{\theta}-\underline{\theta}}{2}\right]$. Finally, the first-best social welfare and optimal number of firms (x^*) are proportional to the mass of agents in the economy:

$$\begin{aligned} x^* &= N\widetilde{x}^* \\ SW^*(N, x^*) &= N\sigma(\widetilde{x}^*) \end{aligned}$$

which follows from our assumption of constant returns to scale in the team production technology.

3 Incomplete contracts

We now analyze the decentralized outcome when 2-members teams cannot write complete contracts. We follow the incomplete contracting literature in assuming that parties can only contract on a limited number of variables ex ante. In particular we assume that they can only allocate property rights over physical assets at the first stage. They are not able to avoid bargaining for the division of surplus at the production stage. Bargaining introduces the possibility of mutual "hold-up" among team members, which distorts their incentives to exert effort. Our setup is a particular case of Hart and Moore (1990), except that outside opportunities are endogenous in our model.

The timing of the game is as follows:

Introduction of assets. At $t = 1$ firms are formed. Firms are contracts that specify ex ante transactions between team members (θ, θ'), and property rights over identifiable physical assets[5]. Agents face no liquidity constraints, nor imperfect information, when they set up the firm.

Investments. At $t = 2$ agents exert effort. Following the incomplete contracting literature we assume that effort, although observable, is not a verifiable variable.

Bargaining. At $t = 3$ bargaining takes place. With probability $\alpha_m \geq \frac{1}{2}$ the owner makes a take it or leave it offer to the non-owner (that is, asset owners have at least the same bargaining power as non-owners). With probability $\alpha_w = 1 - \alpha_w$ the reverse offer takes place.

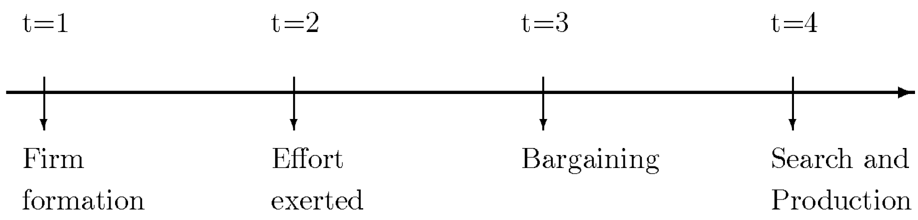
"Search" and production. At $t = 4$ production takes place. If s agents did not reach an agreement at $t = 3$, they search for vacancies in ex post production. The probability for each of these agents to find a vacancy is $\widetilde{\lambda}(s, x) < 1$, where x is the total number of vacancies in ex post production. The matching function[6] satisfies $\widetilde{\lambda}_x(.) > 0$, $\widetilde{\lambda}_{xx}(.) < 0$ and $\widetilde{\lambda}_s(.) < 0$ and for $m > 0 : \widetilde{\lambda}(s, x) = \widetilde{\lambda}(ms, mx)$.

The following definition will be useful: $\lambda(x) \equiv \lim_{\varepsilon \to 0} \widetilde{\lambda}(\varepsilon, x)$. $\lambda(x)$ is the matching function when there x vacancies but there is only "one" worker searching.

The timing is summarized as follows:

[5] Our argument is robust to the alternative assumption that agents do not meet until the production stage and are unable to write contracts before, as in Acemoglu (97).

[6] It is important to note that the function λ can also be interpreted differently than a matching function. It can also be interpreted as reflecting the worker's bargaining power when there are x firms looking for workers and s workers looking for vacancies in a decentralized market.



3.1 Equilibrium allocation

Consider the bargaining stage, $t = 3$. Following Hart and Moore (90), we can rule out joint ownership of the physical asset[7]. We will refer to the owner (or manager) as θ_m, and to the worker as θ_w. For two agents (θ_m, θ_w) that formed a team at $t = 1$, the surplus to be divided when bargaining is given by: $y\left(e_{\theta_w}, e_{\theta_m} \mid \theta_m, \theta_w\right)$. The outside option for the owner when the team breaks down is to produce with the physical asset (that she controls) and obtain $y\left(e_{\theta m}, 0 \mid \theta_m, \theta_w\right)$. For the worker the best option outside the firm is to search for a vacancy and engage in ex post production. This gives: $\phi\widetilde{\lambda}(s,x)e_{\theta_w}$, which leads to:

[7] Maskin and Tirole (97) show that in this context the first best can be achieved if agents can use revelation mechanisms based on subgame perfect implementation. In their words, we focus on this simple institutional set-up on a priori grounds. Yet, it should be said that what is central to our argument is the existence of transaction costs from a hold-up problem, rather than the optimal allocation of property rights per se. We believe that our argument can also be made with similar effect if, as in Grout (84) and Acemoglou (97), agents sink their investments before they meet and before they are able to write contracts. That is, if the market failure is market incompleteness rather than contract incompleteness. In this second setting the particular interaction between the hold-up problem and outside opportunities that we focus on is likely to be similar than in our setting. Yet, we develop on the Hart-Moore framework since we believe it provides a very useful benchmark.

Result 1 *For any allocation of bargaining power, in equilibrium team members in all firms trade at* $t = 3$ $(s = 0)$ *and the relevant matching function at* $t = 4$ *is* $\lambda(x)$

Result 1 [8] implies that as the number of firms in the economy becomes larger, the number of workers' job options outside their initial firm also increases. Since these options are worse than the utility from remaining in the original team, they are never used in equilibrium. This introduces an asymmetry between the growth of options (vacancies for ex post production) as a function of the scale N, and the growth of the number of workers that effectively search for a job ex post, that remains nihil independently of N. Although the options are worse than the team positions, they have value when there is an agency problem at production, because the outside options affect the incentives to invest. The endogenous asymmetry in the search function is what produces increasing aggregate returns to human capital in the decentralized allocation.

Call V_m and V_w the expected payoff from bargaining to the manager and the worker respectively. V_w and V_m are given by:

$$\begin{aligned} V_w(e_{\theta m}, e_{\theta_w}) &= \alpha_w \left[y\left(e_{\theta_w}, e_{\theta_m} \mid \theta_m, \theta_w\right) - y\left(e_{\theta_m}, 0 \mid \theta_m, \theta_w\right)\right] + \alpha_m \lambda(x)\, \phi e_{\theta_w} \\ V_m(e_{\theta_m}, e_{\theta_w}) &= \alpha_m \left[y\left(e_{\theta_w}, e_{\theta_m} \mid \theta_m, \theta_w\right) - \lambda(x)\, \phi e_{\theta_w}\right] + \alpha_w e_{\theta_m} \end{aligned} \tag{5}$$

From expression (5) the number of firms in the economy has distributive effects, as it changes the terms of trade between owners and workers inside the firm. We show that this externality has effects on agents' incentives to provide high effort.

[8] Result 1 is straightforward since for all $(e_{\theta_M}, e_{\theta_M}, x, s)$, from (1) and (2) and since $\lambda(s, x) < 1$, we have :

$$y\left(e_{\theta m}, e_{\theta w} \mid \theta_m, \theta_w\right) > y\left(e_{\theta_m}, 0 \mid \theta_m, \theta_w\right) + \widetilde{\lambda}(s, x)\phi e_{\theta_w}$$

Consider the incentive compatibility constraint of a type-θ owner, given that w exerts e_w. This is given by:

$$V_m(e_w, \overline{e}) - c(\overline{e}, \theta_m) \geq V_m(e_w, \underline{e}) - c(\underline{e}, \theta_m) = V_m(e_w, \underline{e}) \quad \Leftrightarrow$$

$$\begin{aligned} & \alpha_m \left[y\left(\overline{e}, e_{\theta_w} \mid \theta_m, \theta_w\right) - \lambda\left(x\right) \phi e_{\theta_w} \right] + \alpha_w y\left(\overline{e}, 0 \mid \theta_m, \theta_w\right) - c(\overline{e}, \theta_m) \\ \geq \; & \alpha_m \left[y\left(\underline{e}, e_{\theta_w} \mid \theta_m, \theta_w\right) - \lambda\left(x\right) \phi e_{\theta_w} \right] + \alpha_w y\left(\underline{e}, 0 \mid \theta_m, \theta_w\right) \end{aligned}$$

This simplifies to the following incentive compatibility condition:

$$\begin{aligned} c\left(\overline{e}, \theta_m\right) \; \leq \; & \alpha_m \left[y\left(\overline{e}, e_{\theta_w} \mid \theta_m, \theta_w\right) - y\left(\underline{e}, e_{\theta_w} \mid \theta_m, \theta_w\right) \right] \\ & + \alpha_w \left[y\left(\overline{e}, 0 \mid \theta_m, \theta_w\right) - y\left(\underline{e}, 0 \mid \theta_m, \theta_w\right) \right] \end{aligned}$$

The incentive constraint for the type-θ_W worker, when the manager exerts effort e_M is:

$$V_w(e_m, \overline{e}) - c(\overline{e}, \theta_w) \geq V_w(e_m, \underline{e}) - c(\underline{e}, \theta_w) = V_w(e_m, \underline{e}) \Leftrightarrow$$

$$\begin{aligned} & \alpha_w \left[y\left(\overline{e}, e_{\theta_m} \mid \theta_m, \theta_w\right) - y\left(\overline{e}, 0 \mid \theta_m, \theta_w\right) \right] + \alpha_m \lambda\left(x\right) \phi \overline{e} - c(\overline{e}, \theta_m) \\ \geq \; & \alpha_w \left[y\left(\underline{e}, e_{\theta_m} \mid \theta_m, \theta_w\right) - -y\left(\underline{e}, 0 \mid \theta_m, \theta_w\right) \right] + \alpha_m \lambda\left(x\right) \phi \underline{e} \end{aligned}$$

what simplifies to the IC condition for the worker:

$$\alpha_w \left[y\left(e_{\theta_m}, \overline{e} \mid \theta_m, \theta_w\right) - y\left(e_{\theta_m}, \underline{e} \mid \theta_m, \theta_w\right) \right] + \alpha_m \phi \lambda\left(x\right) \left(\overline{e} - \underline{e}\right) \geq c\left(\overline{e}, \theta_w\right)$$

These IC inequalities make clear the effect of market externalities on incentives inside the firm. These effects are asymmetric for the owner and the worker:

Result 2 *The incentive constraint of the manager is not (directly) affected by the total number of firms in the economy. On the other hand, increases in the total number of firms improve the worker's incentive to exert effort.*

In order to obtain the total number of firms in equilibrium, the following definitions will be useful:

Definition 1 *We call the* workers's threshold type, $\widetilde{\theta}_w(x)$, *the minimum value of* θ *such that an agent who is the owner of the asset in a firm will invest when the manager invests.*

$$\begin{aligned} c\left(\overline{e}, \widetilde{\theta}_w(x)\right) &\equiv \alpha_w \left[y\left(\overline{e}, \overline{e} \mid \theta_m, \theta_w\right) - y\left(\overline{e}, \underline{e} \mid \theta_m, \theta_w\right)\right] \\ &\quad + \alpha_m \lambda(x) \phi\left(\overline{e} - \underline{e}\right) \end{aligned} \tag{6}$$

Definition 2 *We call the* manager's threshold type, $\widetilde{\theta}_m$, *the minimum value of* θ *such that an agent that is the owner of the asset in a firm will invest when the worker invests.*

$$\begin{aligned} c\left(\overline{e}, \widetilde{\theta}_m\right) &\equiv \alpha_m \left[y\left(\overline{e}, \overline{e} \mid \theta_m, \theta_w\right) - y\left(\underline{e}, \overline{e} \mid \theta_m, \theta_w\right)\right] \\ &\quad + \alpha_w \left[y\left(\overline{e}, 0 \mid \theta_m, \theta_w\right) - y\left(\underline{e}, 0 \mid \theta_m, \theta_w\right)\right] \end{aligned} \tag{7}$$

These threshold types set an upper bound on the total number of firms. From assumptions (1) and (2) it is clear that $\widetilde{\theta}_m \leq \widetilde{\theta}_w(x)$ for all values of x.

The total number of firms satisfies $x \leq N\frac{1}{2}\left(\overline{\theta} - \widetilde{\theta}_m\right)$, since agents with type $\theta < \widetilde{\theta}_m$ do not exert effort in any case as team members. On the other hand, if in equilibrium $\widetilde{\theta}_w(x) \geq \frac{1}{2}\left(\overline{\theta} + \widetilde{\theta}_m\right)$ the number of firms is given by $N\left(\overline{\theta} - \widetilde{\theta}_w\right)$ since for each individual that is eligible to be a worker there is an agent that is eligible to be a manager, but not vice versa. These conditions together with a fixed point condition on x, yield the total number of firms:

Proposition 1 *The total number of firms in the equilibrium with incomplete contracts,* x^e, *is a solution to:*

$$x^e = N \max\left\{0, \min\left[\left(\overline{\theta} - \widetilde{\theta}_w(x^e)\right), \frac{1}{2}\left(\overline{\theta} - \widetilde{\theta}_m\right)\right]\right\} \equiv \Lambda(x^e) \tag{8}$$

Moreover, all agents with $\theta \in \left[\overline{\theta} - 2x^e, \overline{\theta} - x^e\right]$ *participate in the firm as* managers *(asset owners) and all agents with* $\theta \in \left[\overline{\theta} - x^e, \overline{\theta}\right]$ *participate as* workers *(not-owners).*

Proof: in the Appendix.

It is clear from the comparison of the efficiency condition (4) and the equilibrium condition (8) that x^e will in general be inefficient (x^e is in general suboptimal, since a team will not be formed if it does not generate positive surplus without agency problems).

The fixed point condition in (8) is illustrated in Figure 2:

Insert Figure 2 here

The solution to (8) depends on the position of the mapping $\Lambda(x)$ relative to the identity mapping. It is straightforward to show:

Corollary 1 *From the concavity of* $\Lambda(x)$[9] *at least one solution always exists and the number of solutions is not bigger than three. If there are three equilibria* (x^o, x^{ue}, x^{se}) *we have:* $x^o = 0 < x^{ue} < x^{se}$ *and* (x^o, x^{se}) *are stable equilibria but* (x^{ue}) *is an unstable equilibrium.*

Consider the case where there are three equilibria: $(0 < x^{ue} < x^{se})$. The possibility of multiple equilibria is a direct consequence of incentive externalities: under no renegotiation (first best) there is only one equilibrium in the industry.

Under incomplete contracts, if agents believe that other agents will not enter, then human capital is not a "liquid" asset in the secondary market (ex post production): workers face poor outside opportunities inside the firm and therefore high expropriation. Incentives are poor and agents do not start firms. The initial beliefs are confirmed. If agents believe that there will be many firms in the industry, the converse happens, good incentive conditions are anticipated and agents enter the industry.

[9] $\Lambda''(x) \leq 0$ since:

$$\Lambda''(x) \in \left\{0, -\frac{\alpha_m \lambda''(x)\, \phi \Delta e c_\theta(\overline{e}, \theta(x)) - c_{\theta\theta}(\overline{e}, \theta(x))\, \alpha_m \lambda'(x)\, \phi \Delta e}{\left(c_\theta(\overline{e}, \theta(x))\right)^2} < 0\right\}$$

3.2 Comparative statics

Consider a parameter ψ that is an index of efficiency in the cost function $c(\overline{e},\theta,\psi)$, such that $c\left(\overline{e},\overline{\theta},\psi\right)=\underline{c}(\overline{e})$ and:

$$0<\frac{\partial c(\overline{e},\theta,\psi)}{\partial\theta}<\frac{\partial c(\overline{e},\theta,\psi')}{\partial\theta}\Leftrightarrow\psi>\psi'$$

Let ρ be a parameter of "efficiency" in the matching function $\lambda(x,\rho)$ such that

$$\forall x\in\left[0,\frac{1}{2}N\right],\qquad 1>\lambda(x,\rho)>\lambda(x,\rho')\geq 0\Leftrightarrow\rho>\rho'$$

Definition 3 *We say that the economy has poor incentives if*

$$c\left(\overline{e},\overline{\theta}\right)>\alpha_w\left[y\left(\overline{e},\overline{e}\mid\theta_m,\theta_w\right)-y\left(\overline{e},\underline{e}\mid\theta_m,\theta_w\right)\right]$$

This condition says simply that the most efficient type does not enter the firm as a worker.

Proposition 2 *If the economy has poor incentives then,* $x^{se}=0$ *is a stable solution to (8). Moreover, for* $z\in\{N,\psi,\rho,\phi\}$ *there is a critical value of* $\widehat{z}$ *such that:*

$$\begin{array}{l} x^{se}(z)=x^{o}(z)=x^{ue}(z)=0 \text{ and } z<\widehat{z} \\ x^{se}(z)>x^{ue}(z)>x^{o}(z)=0 \text{ and } z>\widehat{z} \end{array}$$

and $x^{se}(z)>0$ *otherwise.*
Moreover, for increments in z *beyond* $\widehat{z}$ *we have:*

$$\frac{\partial x^{se}}{\partial z}\geq 0 \text{ and } \frac{\partial x^{ue}}{\partial z}\leq 0$$

Proof: in the Appendix.

Proposition 2 shows the existence of increasing returns to scale in population N, in the economy with incomplete contracts, even if the underlying technology is of constant returns to scale.

From proposition 2 there is a discontinuity in the relationship between the development characteristics of the economy and industrial activity. Below a certain development level there is no activity. Beyond that level there are equilibria that support positive industrial employment. In particular, the stable equilibrium implies a greater level of employment in the economy. As the size of the economy increases (or as the economy becomes more efficient) the employment level implied by the stable equilibrium increases, whereas the employment level in the unstable equilibrium decreases.

Since the transaction costs and the inefficiency in the economy are proportional to the critical mass, Result 3 shows a connection between technological progress and transaction costs. Technological progress reduces transaction costs, since it reduces the critical size of the economy required to start production. Moreover, if we interpret this to mean that the density of the economic in a region increases the efficiency of the matching function λ, increases in economic density reduce transaction costs inside the firm and reduce the minimum start-up scale of a sector.

4 Empirical Implications

The leading implication of the previous section is that under incomplete contracts (and therefore aggregate increasing returns) there are in general (if the level of development is sufficiently low) multiple equilibria, with industrial delocation as one of the equilibria. On the other hand, in the economy with complete contracts and constant returns technology, industries have one equilibrium level of employment that is positive. In this section we explore this theoretical implication: it is more likely to find evidence of multiple equilibria and industrial delocation in less developed regions that in more developed regions[10].

We distinguish between regions, $i \in \{1, ..., I\}$, productive sectors, $j \in$

[10] It is important to notice that, if this empirical pattern is true, it can be explained by other theories of aggregate increasing returns different to our theory based on contracting externalities. We make no attempt here of empirically selecting the best theory to explain multiple equilibria. We simply argue that multiplicity of equilibria is inconsistent with the constant returns/complete contracts model and it is not inconsistent with the constant returns (at the firm level)/incomplete contracts setting.

$\{1,...,J\}$ and time periods, $t \in \{0,...,T\}$. y_t^{ij} is the number of employees[11] in sector j, and region i, at time t. A region i at time t is characterized by its population mass N_{it}, and its degree of technological development or efficiency, (ψ_{it}, ϕ_{it}) .

With a straightforward dynamic interpretation to the schedules in (8), given y_t^{ij} the subsequent number of firms y_{t+1}^{ij} is[12]:

$$y_{t+1}^{ij} \equiv \Lambda_i \left(y_t^{ij}, z_{it} \right)$$

Total sectoral next period's employment in region i depends non-linearly on this period's existing sectoral employment and on the characteristics of the region. In order to approximate the mapping Λ we specify the following equation:

$$\widetilde{y}_{t+1}^{ij} = \max \left\{ 0, \left(a_0 + a_1 \widetilde{y}_t^{ij} + a_2 \left(\widetilde{y}_t^{ij} \right)^2 + \sum_{r=1}^{R} b_r \ z_t^i (r) + \widetilde{\varepsilon}_t^{ij} \right) \right\} \qquad (9)$$

The censored structure of (9) follows from (8) and from y_t^{ij} being a stock. We assume the independence and normality of the error term: $\widetilde{\varepsilon}_t^{ij} \sim N(0, \sigma^2)$. The coefficients (a_0, a_1, a_2) capture the non-linearities[13] in Λ. The coefficients $(b_1, .., b_r, .., b_R)$ capture the dependence of Λ on the regional scale and development characteristics:. $\{z_t^i (r)\}_{r=1}^{R}$, as predicted by Proposition 2.

The following table contains necessary conditions for the parameters (a_0, a_1, a_2) in (9) to be consistent with the existence of only one equilibrium and multiple equilibria in the industry:

[11] In the model y is proportional to the number of teams x, equal to $\frac{1}{2}y = x$.

[12] It is strightforward to show that $\Lambda(y_t)$ is the actual number of employees at $t+1$ if the model described in the theoretical section represents the period t economy and there is an arbitrary and discrete number of periods.

[13] We expect $\Lambda()$ to be concave: see footnote 9.

Table 1:
Restrictions on a_o, a_1, a_2.

One equilibrium:	Multiple equilibria:
$a_o > 0$	$a_o < 0$
$a_1 \geq 0$	$a_1 > 1$
$a_2 < 0$	$a_2 < 0$

If there are three equilibria, we call the unstable equilibrium $y^{ue} \in (0, y^{se})$ the *critical mass* of the sector, since if the industry has $y_t < y^{ue}$ sectoral employment converges to zero (or the sector follows a *delocation process*).

From Proposition 2, we expect the coefficients $(b_1, .., b_r, .., b_R)$ related to development effects to be positive. The coefficients in b_r are interpreted as correlations between development and scale characteristics of the region and the increment of sectoral employment. Positive values of b_r shifts upwards the $\Lambda()$ schedule, implying that industry dynamics in large or developed regions have a smaller critical size (unstable steady state x^{ue}) and larger stable steady states (x^{se}) in industry size.

Our empirical strategy is as follows. We construct in the first place proxy variables to implement equation (9) and show descriptive statistics. We estimate the censored endogenous variable (*tobit*) model in (9) by maximum likelihood. We test the unique versus multiple equilibria hypothesis, according to the definition in Table 1.

4.1 Data

We have combined data from a number of sources: the data on the employment stock in sectors, regions and years is computed from the Active Population Survey (*Encuesta de Población Activa*) of the National Institute for Statistics (*INE*), the data on regions' unemployment rate is taken from the *Regional Accounts* of the Spanish Ministry of Economy, the figures for regions' populations are from the Census. Finally the data on regional gross product is from the BBV Foundation's Statistical Sourcebook. More precisely, our proxy variables are constructed as follows:

- $\tilde{y}_t^{ij}$ is defined as the total number (in tens of thousands) of employed individuals in (i, j) in year t, as reported in the Spanish Active Population Survey (*Encuesta de Población Activa*).

- N_{it} is the population size of region i at time t.

- ψ_{it}, ϕ_{it}, are related to the technological efficiency of the economy. We use the gross product per capita statistic and the number of patents issued in region i as proxy variables.

- We introduce in addition a number of control variables in one specification. In particular we use information on the regional unemployment rate and an index of region's specialization (constructed from the employment data).

The following table shows the descriptive statistics of the proxy variables, the controls and some of their cross-products.

Table 2

Descriptive statistics of the proxy variables in equation (9).

Variable	**Mean**	**Std.Dev**	**Min.**	**Max.**
$\frac{\textit{Stock employed}}{10{,}000} \equiv (y_{t+1}^{ij})$[14]	0.19	0.56	0	9.40
$\frac{\textit{Lagged stock employed}}{10.000} \equiv (y_t^{ij})$[15]	0.195	0.56	0	9.58
$\left(\frac{\textit{Lagged stock employed}}{10{,}000}\right)^2 \equiv \left(y_t^{ij}\right)^2$	0.35	3.19	0	91.77
Gross product per capita$_{it}$[16]	1.09	3.83	0.48	2.15
log of population$_{it}$	13.19	0.80	11.45	15.43
Income per capita $*$ log of popul.	14.41	4.43	6.44	30.02
Number of patents issued$_{it}$	39.94	107.00	0	729
Index of specialization$_{it}$	0.16	0.56	0.01	1
Unemployment rate$_{it}$	18.01	1.59	16.21	20.53

[14] Period: [1987-1992].
[15] Period [1986-1991].
[16] In millions peseta.

5 Estimation Results

The main implications of our model are the existence of multiple equilibria in industry dynamics and the negative correlation between the critical size of industrial sectors (measured by the unstable steady state size of the industry) and the scale and development of the region. There is a reinforcement effect where regional development and scale reduce the probability of industrial delocation. The reinforcement effect arises because the larger scale of the economy improves incentives' conditions, directly fostering the formation of firms.

Table 3 shows the result of estimating the censored endogenous variable (*tobit)* model in equation (9) by maximum likelihood, with different subsets of the variables in Table 2. Model 2 estimates the regions' fixed effects. The significant fixed effects coefficients from model 2 are reported in Table 4.

Table 3

Estimation of (9).

Number of observations: 7309

	Model 1	**Model 2**	**Model 3**	**Model 4**
Log likelihood	-1,366	-1,404	-1,399	-1,410
Variable:				
constant:	-0.5298 (0.0190)*	-0.0306 (0.0108)*	-1.0440 (0.1169)*	-1.1761 (0.1215)*
y_{t-1}^{ij}	1.0834 (0.0052)*	1.0561 (0.0064)*	1.0649 (0.0062)*	1.0588 (0.0063)*
$\left(y_{t-1}^{ij}\right)^2$	-0.0139 (0.0009)*	-0.0097 (0.0010)*	-0.0100 (0.0009)*	-0.0100 (0.0009)*
Income per capita			0.7403 (0.1013)	0.9372 (0.1103)*
log(population)			0.0747 (0.0088)*	0.0830 (0.0090)*
Income p.c.$*$ log(pop.)			-0.0554 (0.0076)	-0.0706 (0.0084)*
num. of patents				0.0001 (0.00003)*
unemployment rate				0.0012 (0.0012)
especialization				0.0105 (0.0126)
regional dummies	no	yes	no	no

($*$): significant at the 1% level

In all four models in table 3 we reject the null hypothesis that there is unique industrial steady state, since the intercept term (a_o) is negative and significant, the slope of Λ at the origin (a_1) is greater to one (at the 1% level) and the estimated $\Lambda(y_t)$ mapping is concave in y_t (a_2 is negative and significant). Given the estimated hypothesis, we find that there are three equilibria in the average industry: no activity $(y = 0)$ being an equilibrium and two more equilibria (y^{ue}, y^{se}) that satisfy the estimated equation: $(y^{ue}, y^{se}) = E\widehat{\Lambda}(y^{ue}, y^{se})$. The stable equilibrium is greater than the unstable one $(y^{ue} < y^{se})$.

Are sectoral dynamics different in developed and less developed regions?. Model 2 introduces a regional fixed effect, that measures the difference of the regional intercept with the excluded region's intercept[17]. We find ten regions with an intercept term a_o significantly smaller to that of the excluded region (that was already negative). This means that the Λ schedule shifts downwards for these ten regions and that they have greater critical mass of the industry and a smaller stable steady state. These ten regions are listed in Table 4 below the entry for Barcelona.

Table 4 shows the difference between intercepts and the excluded region's intercept $(a_{oi} - a_{o1})$, the smaller root of the estimated difference equation (the critical mass x^{ue}), the region's population (in level) and the gross product per capita as of 1987. Notice that all ten regions are sparsely populated (the average regions's population is 535,000) and have gross product per capita smaller than national's average of 1.09 (see Table 2). The estimated sectoral critical mass for the sparse regions ranges from 620 employees in Cáceres to 903 in Zamora (to be compared with the absence of a critical mass in the large scale region of Barcelona).

At the other extreme of the spectrum we find that the fixed effect of Barcelona's region is significantly greater than the excluded region (this is also the case of Madrid, but the coefficient is not significantly different to zero). The intercept term for Barcelona is estimated as non-negative. This

[17] The excluded region in the regression is Alava, a relatively industrial and developed region. This explain why most of the etimated fixed efect coefficients in model 2 are negative.

implies that for this region we accept the hypothesis that there is only one steady state and critical mass of the industry is nonexistent.

Table 4

Dummy variables estimated in model 2,
that are significant at the 5%level.

Region	$a_{oi}-a_{o1}$ (std.dev.)	x_i^{ue}	N_i	**product p.c.**
Barcelona	0.005 (0.173)	$\emptyset$	4,629,176	0.9561
Cáceres	−0.0347 (0.0168)	620	422,347	0.6982
Albacete	−0.0348 (0.0159)	620	346,793	0.5867
Soria	−0.0382 (0.0169)	652	97,915	0.8511
Teruel	−0.0414 (0.0169)	682	149,423	0.8459
Almería	−0.0427 (0.0170)	695	446,200	0.7183
Lleida	−0.0438 (0.0168)	705	352,350	1.0793
Avila	−0.0454 (0.0171)	720	182,634	0.6508
Lugo	−0.0461 (0.0170)	727	406,123	0.6204
Cuenca	−0.0588 (0.0173)	847	213,812	0.6023
Zamora	−0.0647 (0.0176)	903	222,240	0.6712

A clearer picture of the relationship between regions' scale and development on the one hand and the size of the industrial critical mass on the other hand, can be drawn from models 3 and 4 in Table 3. The coefficients for Gross Product per capita and log of population are positive and significant. Scale and development alter industry dynamics reducing the critical mass x^{ue}. But absolute scale and development do not complement each other to reduce critical mass, as revealed by the negative (and significant) interaction term "*Income p.c.∗ log(pop.)*" in models 3 and 4. Large scale *or* development are each by itself sufficient to reduce the probability of industrial delocation. That is, regions with sufficiently large population but small gross product per capita have a small critical mass in industries. Conversely, small regions with sufficiently large gross product per capita are likely to avoid a "delocation trap". But regions that are *both* small in size and have small per capita gross product have a large sectoral critical mass. These are the ten regions with the greatest critical size, listed in rows 2 to 12 in Table 4. In these ten regions a negative sectoral shock that sets the level of employment below the critical mass x^{ue} starts a delocation process that can lead to the end of sectoral activities in the region.

In the remaining 39 regions, the intercept term is not statistically different to the excluded region's (*Alava*) intercept term (the constant in Model 2 of Table 3) equal to -0.0306. This intercept term implies a critical size of 210 employees, a middle ground between the case of Barcelona and the ten sparse regions in Table 4. The list of estimated coefficients for these 39 regions is shown in the Appendix.

6 Conclusions

We have derived a theory of increasing returns to human capital that does not arise directly from the assumptions on the productive technology. Our theory starts from a simple observation. As the scale of an economy increases, the number of inefficient re-allocations of factors increases exponentially. Inefficient re-allocations serve as outside options to agents when bargaining for the terms of trade. Although these options do not directly contribute to social welfare, they change agents' incentives to engage in human capital investments. With transaction costs deriving from inefficient ex post bargaining, we find that in the decentralized equilibrium there are generally increasing returns to scale.

As opposed to other theories of agglomeration based on transaction costs derived from ex post bargaining (Acemoglou (97)), our theory does not rely on search externalities. It does not rely either on the idea of distorting the capital-labor ratio from the optimal ratio, so as to make labor "scarce" and force capital to compete for labor. In our setting there are centralized factors' markets ex ante (although there is search -out of the equilibrium path- for secondary markets) and the proportion of factors at production is the efficient one. The source of increasing returns is the fact that the economy produces for free a good that is valuable under transaction costs: the possibility of inefficient re-allocation of factors.

We have shown how to use the theoretical model to restrict the data. In particular, our fixed-point equilibrium condition is used to specify a dynamic equation of the number of new firms as a function of scale and measures of activity, in a region and sector. The theory leads to a censored first-order non-linear difference equation for the level of sectoral employment in regions. Our estimates of the difference equation are consistent with the existence of multiple steady states in sectors and with the existence of a sectoral critical mass. In particular, if a sector suffers a shock that sets total sectoral employment below the critical size, industrial delocation follows and leads to the destruction of sectoral employment in the region.

We find substantial differences across regions. In densely populated or developed regions the average critical size of sectors is smaller (even non-existent in one case) than in sparsely and less developed regions. There is a positive reinforcement effect in development, in the sense that greater development reduces the probability of industrial delocation.

On the other hand we are not able at this stage to empirically distinguish between our theory of agglomerations and other theories.

References

[1] Acemoglu, D. (1996), "A Microfoundation for Increasing Returns in Human Capital Accumulation", *Quarterly Journal of Economics*, **111**, 779-804.

[2] Acemoglu, D. (1997), "Training and Innovation in an Imperfect Labour Market", *Review of Economic Studies*, **64**, 445-464.

[3] Hart, O. and John Moore (1990) "Property Rights and the Nature of the Firm", *Journal of Political Economy*, **98**, 1119-58.

[4] Krugman, P. (1991), "Geography and Trade", *MIT Press*.

[5] Stole, L. and J. Zwiebel, (1996), "Organizational Design and Technology Choice under Intrafirm Bargaining", *American Economic Review*, **86**, 1, 195-222.

Appendix

Proof of Proposition 1:

For two agents $(\theta,\ \theta')$, the condition:

$$(*) \equiv \left[\theta \geq \widetilde{\theta}_w\left(x\right) \text{and} \theta' \geq \widetilde{\theta}_m\right]$$

is sufficient for them forming a firm, from (6) and (7). Condition $(*)$ is necessary for them forming a firm from Proposition 1. So we can restrict ourselves to pairs of agents that satisfy condition $(*)$. In particular, agents with $\theta' \in \left[\widetilde{\theta}_m, \overline{\theta}\right]$ are potential managers only (PMO): they would never exert effort as workers. Agents with $\theta \in \left[\widetilde{\theta}_w, \overline{\theta}\right]$ are potential managers and workers (PMW): they would exert effort both as workers if matched with a manager with $\theta' \in \left[\widetilde{\theta}_m, \overline{\theta}\right]$ and also as managers, as long as matched with a worker with $\theta'' \in \left[\widetilde{\theta}_w, \overline{\theta}\right]$. Since $\widetilde{\theta}_m < \widetilde{\theta}_w$, there are three possible situations: 1) blocked entry. If $\widetilde{\theta}_w > \overline{\theta}$. In this case there are no agents that would exert effort as workers even if managers would exert effort. No managerial firms can enter. 2) scarce workers. If $\overline{\theta} - \widetilde{\theta}_w \leq \widetilde{\theta}_w - \widetilde{\theta}_m$. There are not enough workers for the number of (PMO). The number of managerial firms is set by the number of PMW. 3) scarce managers. If $\overline{\theta} - \widetilde{\theta}_w > \widetilde{\theta}_w - \widetilde{\theta}_m$. PMO are fewer than PMW. Some of the latter can be given property rights and matched with other (necessarily more efficient) PMW. The number of managerial firms is $1/2\left(\overline{\theta} - \widetilde{\theta}_w\right)$, such that all $\theta's$ with $\theta \geq \widetilde{\theta}_w$ are involved in a firm.

■

Proof of Proposition 2:

From condition $c\left(\overline{e}, \overline{\theta}\right) > \alpha_w \left[y\left(\overline{e}, \overline{e} \mid \theta_m, \theta_w\right) - y\left(\overline{e}, \underline{e} \mid \theta_m, \theta_w\right)\right]$ it is clear that the most efficient type $\overline{\theta}$ cannot enter as a worker when $x = 0$. With no workers in the economy there are no firms and $x^e = 0$ is an equilibrium. But $\Lambda\left(x\right)$ is strictly increasing in x : there exist a number of firms x_o such that $\Lambda\left(x\right) = 0$ for $x \leq x_o$ and $\Lambda\left(x\right) > 0$ for $x > x_o$. Since $\Lambda\left(x\right)$ is linear in N the proposition follows.

■

Figure 2