Cournot equilibrium and welfare with heterogeneous firms

Enrico De Monte* and Bertrand Koebel**

February 27, 2023

Abstract

This paper characterizes the short and long-run Cournot equilibrium with heterogeneous firms and stochastic technological change. We consider that firms have different technologies with heterogeneous fixed and variable costs and various degrees of market power in the product market. In a framework with homogeneous firms, Mankiw and Whinston (1986) showed that the long-run Cournot equilibrium may be inefficient due to too many entries. We extend their result to the case of heterogeneous firms and show that higher industrial concentration of production is welfare improving. Empirically, using data for manufacturing firms in France, we found a wide degree of heterogeneity in technologies, and we are able to identify the technology parameters which reproduce the observed distribution of firms size. We characterize the type of fixed and variable cost functions which allow to generate the observed level of cost and profit. We investigate empirically whether profit maximization is compatible with welfare maximisation, and find substantial levels of inefficiency. Imperfect competition, however, contributes modestly to the inefficiency, which is mainly explained by firms' use of heterogeneous and often inefficient technologies.

Keywords: cost function, fixed cost, marginal cost, returns to scale, technological change, nonlinear GMM, panel data.

JEL Classification: C31; L20; L60; M14; O38; Q55.

* ZEW, Leibniz Centre for European Economic Research, L 7, 1, 68161 Mannheim (Germany). Email: Enrico.DeMonte@zew.de.
** Université de Strasbourg, CNRS and BETA, 61 avenue de la Forêt-Noire, 67000 Strasbourg (France). Email: koebel@unistra.fr.

We warmly thank Rabah Amir for continuous and constructive discussions during the writing process of the paper. We also benefited from thorough comments by Flora Bellone, Ivan Ledezma and Bettina Peters. We are also indebted to Claude d'Aspremont, Xi Chen, Rodolphe Dos Santos Ferreira, Serge Garcia, François Laisney, Gauthier Lanot, Anne-Laure Levet, Phu Nguyen-Van, Yasunori Okumura, Hanitra Rakotoarison, Patrick Rey, and the participants of seminars at Cergy-Pontoise, Konstanz, Mannheim, Nancy, Paris, Strasbourg (PET 2019), Trier, Umeå and Vienna (EARIE 2022) for their helpful comments.

1 Introduction

One of the unfortunate consequence of most firm level cost function specifications is their difficulty to yield plausible (optimal) output levels. In this context, empirical studies are often not suited to address issues related to the size distribution of firms and their determinants, and produce biased results and inference. The reasons beyond these shortcomings are related to their restrictive functional forms and the difficulty of considering unobserved heterogeneity adequately. With cost functions, unobserved heterogeneity has multiple dimensions. It cannot be simply additive: Heterogeneity in the fixed costs vanish in the derivation of the profit maximizing condition and is useless to generate heterogeneous firm size. Conversely, heterogeneity in the variable cost function is unable to explain why so many small firms make positive profits while others do not. One objective of this paper is to propose a setup allowing for joint heterogeneities in fixed and variable costs, and enabling to reproduce the observed distribution of firms sizes. Heterogeneous cost functions yield firm and time specific break even points and minimum efficient scale. This in turn characterizes which technologies allow generating positive profits, and identifies those firms which are likely to exit the market as well as potential entrants and survivors. We adopt the Cournot model, with heterogeneous firms interacting strategically and choosing their optimal output level given aggregate output, and further cost and demand parameters.

While the literature on the existence and unicity of Cournot equilibrium often considers industries with identical firms and symmetric equilibrium, there are some interesting exceptions. Novshek (1985) showed that a short-run Cournot equilibrium exists under weak conditions on firms' cost function. Unicity of the short-run Cournot equilibrium with heterogeneous firms was derived by Gaudet and Salant (1991). In the long-run, when firms' entry and exit occurs, Acemoglu and Jensen (2013) and Okumura (2015) proved that existence of the Cournot equilibrium still holds (but is no longer unique in general). We contribute to this literature and amend the homogeneous firm Cournot model and investigate differences in technologies and their interplay with firm size. While our purpose is mainly empirical, we also describe the theoretical implications of heterogeneous technologies at the firm level, both on the short- and the long-run Cournot equilibrium. Interestingly, we show that there is an ordered relationship between firm size (in terms of output) and their type of heterogeneous technology.

It is well known that the short-run Cournot equilibrium is generally not welfare maximizing. Mankiw and Whinston (1986) have shown that even in the long-run, firms' entry and exit do not necessarily contribute to reduce this inefficiency. We extend their result to the case of heterogeneous firms and empirically investigate whether redistributing output over firms allows to increase industry output, reduce total cost and increase efficiency. Especially for France, the stylized facts document that there are many very small firms but a lack of medium sized and large firms. In manufacturing industries, Table **??** illustrates that in comparison to Germany, there is roughly the same number of firms with 0 to 9 employees, but only 54% of the number of small firms (with 10 to 49 employees). This rate decreases to about 35% for larger firms with 50 employees and more. Garicano et al. (2016) attribute the lack of medium sized firms in France to laws specific to firms with 50 employees and more, and which prevent firms' to grow above this threshold. This explanation is not sufficient to describe the lack a medium sized and large firms in France and we investigate whether the low number of firms is related to the nature of the market structure and competition in the manufacturing industries. Starting from a long-run Cournot equilibrium, we study whether total industry output is efficiently allocated over firms, and perform simulations to evaluate the welfare loss due to markups, output misallocation and technological inefficiencies.

We use fiscal data for firms which are available for France for the years 1994 to 2016 (FICUS and FARE data). The data comprises the universe of active firms, but we consider only those belonging to the manufacturing industry. We consider 184 industries at the 4-digit aggregation level, within which firms are assumed to produce an homogeneous output and to compete à la Cournot. Especially for France, the stylized facts document that there are many very small firms but a lack of medium sized and few but influential large firms. In a typical 4-digit industry, 0.5 % of all firms hire about 39 % of the employees working in this industry, and produce 56 % of total industry output. The concentration ratio of the 3 and 10 biggest firms are respectively $C_3 \simeq 53\%$ and $C_{10} \simeq 70\%$. These figures document that there are few actors which must have strong market power, and a large competitive fringe of smaller firms. This seems compatible with the theoretical Cournot model adopted here, allowing for technological differences between firms.

Empirically we have to deal with the incidental parameter problem occurring when taking into

account heterogeneity over firms and across time in fixed and variable costs: new observations carry with them new heterogeneity terms and do not contribute identifying the model in an obvious way. Moreover, when heterogeneity is unobserved but correlated with decision variables (the optimal level of output) least squares estimates are inconsistent. We solve this problem by parameterizing the unobserved heterogeneity, and estimating both the inverse output demand function addressed to an industry, and the firm level output supply, which depends both upon observed market characteristics and unobserved technological parameters. This approach allows to reveal the distribution of unobserved heterogeneity in both the fixed and variable cost. We contribute to the existing empirical literature by introducing explicitly joint heterogeneity in the fixed cost and in the variable cost of production, and studying the interplay between both types of heterogeneity. The existing literature mainly focuses on univariate heterogeneity, either in the variable cost function (Davis, 2006) or in the fixed cost function (Berry, 1992) or in total cost (Esponda and Pouzo, 2019). While these specifications all entail unidimensional heterogeneity in the total cost function, we allow for separate heterogeneity in both the fixed and the variable cost functions. While the theoretical framework for the occurrence of joint heterogeneity and their interdependence is studied by Chen and Koebel (2017), we are not aware of any empirical contributions at the firm level. Another part of the literature tackling the issue of productivity and technological change bases its identification strategy on the production function (Ericson and Pakes, 1995). Adding a firm and time specific effect to the production function, however, imposes strong restrictions on fixed and variable costs. Our cost function based approach allows more flexibility and is compatible with more general specifications of technological heterogeneity.

Section 2 presents the heterogeneous firm setup and describes the short-run Cournot equilibrium. Section 3 characterizes the long-run equilibrium. The theoretical results pertaining to the inefficiency of the Cournot equilibrium are discussed in Section 4, which also describes the welfare maximizing allocation of production over firms. The data and descriptive statistics are presented Section 5. Section 6 and 7 discuss the empirical model along with the estimation strategy and presents the results, and Section 9 concludes.

2 Short-run Cournot equilibrium with heterogeneous quadratic cost functions

Within each industry firms are competing à la Cournot. In the short-run, there are N active firms facing the same inverse demand function

$$p = P(y_n + \sum_{j \neq n}^{N} y_j), \tag{1}$$

where p denotes the output price, y_n the production of firm n and $Y_{-n} \equiv \sum_{j \neq n}^{N} y_j$ the total output of firms' n competitors. We do not introduce subscripts for the industry yet, but it is important to realize that the inverse demand is specific to industry i.

We assume that the total cost function of each firm is the sum of a firm specific fixed cost and a variable cost function:

$$c_n(w_n, y_n) = u_n(w_n) + v_n(w_n, y_n), \tag{2}$$

where the fixed cost of production u_n depends upon input prices w_n but also upon technological choices and constraints which are specific to firm n. The variable cost function v_n satisfies, by definition, the condition $v_n(w_n, 0) = 0$.

Each firm is profit maximizing and chooses its output level according to the first order optimality condition:

$$P(Y) + P'(Y)y_n = \frac{\partial c_n}{\partial y_n}(w_n, y_n) \tag{3}$$

where Y denotes the aggregate output level of the industry.

Note that if the fixed cost function u_n is heterogeneous but the variable cost function v_n is the same over all firms, then (3) implies identical output levels over all firms with the same input prices. Such a model would attribute differences in firm sizes to difference in input prices. Here, heterogeneity in variable costs is helpful to yield optimal individual production levels able to approximate the empirical distribution of firm sizes. The second main advantage of our heterogeneous firm framework, is that it can explain why bigger firms have increasing returns to scale

while smaller firms have decreasing returns. In the homogeneous case with U-shaped average cost functions, returns to scale are increasing for production levels smaller than the efficient scale of production and decreasing for larger production levels. This is not necessarily the case here.

We assume the following regularity conditions (that will be empirically investigated later on):

Assumption 1. The inverse demand function P is nonnegative, continuous, differentiable and decreasing in Y.

Assumption 2. The cost function is continuous in w_n and y_n, nonnegative, differentiable and increasing in w_n and y_n.

Assumption 3. There exist firm-level and aggregate production levels $\overline{y}$ and $\overline{Y}$ such that

(i) the marginal revenue is lower than the marginal cost:

$$P(Y) + P'(Y)\, y < \partial c_n / \partial y_n (w_n, y), \tag{4}$$

for any $y > \overline{y}$ and $Y > \overline{Y}$, and any firm $n = 1, ..., N$;

(ii) the cost function is not too concave:

$$P'(Y) < \partial^2 c_n / \partial y_n^2 (w_n, y), \tag{5}$$

for any $y < \overline{y}$ and $Y < \overline{Y}$, and any firm $n = 1, ..., N$.

Assumptions 1 and 2 are quite common in microeconomics. Assumption 3(i), implies that there is an upper threshold $\overline{y}$ to individual production (because marginal cost is always higher than marginal revenue for $y > \overline{y}$). A3(ii) forbids the occurrence of highly nonconvex cost functions. Condition A3(ii) is common in the literature on Cournot oligopoly, see Amir and Lambson (2000) for instance. Cournot equilibrium exists under relatively mild conditions, we follow Novshek (1985) who showed existence provided that:

Assumption 4. The marginal revenue function satisfies:

$$P^{'}(Y) + y_n P^{''}(Y) \leq 0, \tag{6}$$

for any value of $y_n \leq Y < N\overline{y}$.

A1 and A4 imply that the marginal revenue function is decreasing. A3(ii) and A4 ensure that the profit function is concave, without requiring convexity of the cost function in y. A4 together with the second order condition for profit maximization imply that firms' reaction functions are downward sloping. Gaudet and Salant (1991) have shown that A1-A4 imply the uniqueness of Cournot equilibrium. Amir (1996, Corollary 2.2) used another condition implying the existence of Cournot equilibrium which is not equivalent to A4. A4, however, was found to be more useful for deriving some results below.

We follow Novshek (1984) and consider the backward reaction functions as the solution in $y_n \geq 0$ to the system of N equations (3), for given values of aggregate output Y and input prices w_n:

$$y_n^b(w_n, Y). \tag{7}$$

Assumptions 3(ii) and 4 guarantee that the backward reaction functions are nonincreasing in Y. Given existence, we then characterize Cournot's equilibrium as the solution to the equation

$$Y = \sum_{n=1}^{N} y_n^b(w_n, Y), \tag{8}$$

which guarantees that all firms projections about aggregate output are fulfilled at equilibrium. We denote the equilibrium by Y^N, and $y_n^N = y_n^b(w_n, Y^N)$ and note that these functions depend upon the characteristics of all firms active in the industry.[1] We have the following interesting

[1]The superscript N denotes both Nash equilibrium, and the fact that the number of firms is kept constant (no entry, no exit) here.

implications:

Proposition 1. Under A1-A4, at the Cournot equilibrium with fixed number of firms,
(i) The elasticity of inverse demand $\epsilon(P,Y)$ satisfies $-N < \epsilon(P,Y) < 0$
(ii) Firm's n market share satisfies $y_n^N/Y < -1/\epsilon(P,Y)$
(iii) The value of the marginal cost of production decreases with firm size
(iv) The price markup increases with firm size.
(v) For a subset of $N' < N$ active firms, $Y^{N'} < Y^N$ and $y_n^{N'} > y_n^N$.

Proposition 1 restates several claims that are well known to researchers working in the field of Cournot equilibrium with heterogeneous firms, but often not to be found in textbooks considering mainly homogeneous firms. It follows from Proposition 1, that if we order firms by size (say from the smallest to the biggest), this implies that the same order carry over to the markup and the reverse ordering applies to marginal cost. P1(v) corresponds to what Mankiw and Whinston (1986) refers to as the business-stealing: new entries contribute to increase total output but reduce individual production levels of incumbents. In the context of heterogeneous firms, this result is derived by Acemoglu and Jensen (2013) and Okumura (2015, Lemma 1).

Equality (3) also implies an interesting relationship between firms' profit rate, the inverse demand elasticity and the rate of returns to scale:

$$\frac{py_n^N - c_n}{c_n} = \frac{1}{1+\epsilon\left(P;Y\right)y_n/Y}\epsilon\left(c_n;y_n\right) - 1. \tag{9}$$

Ceteris paribus, the higher the rate of return to scale $1/\epsilon\left(c_n;y_n\right)$, the lower the profit rate; the higher the market share y_n/Y, the higher the profit rate. Equation (9) also implies that for a firm with positive profit there is a lower bound for its market share given by

$$\frac{y_n^N}{Y^N} \geq \frac{\epsilon\left(c_n;y_n\right)-1}{\epsilon\left(P;Y\right)}.$$

Hence, firms with increasing returns to scale must have sufficient market share in order to have positive profits.

We rewrite the cost function in order to highlight the two parameters γ_n^u and γ_n^v which deform the conditional mean functions u and v that are common to all firms:

$$c_n(w_n,y_n) = \gamma_n^u u(w_n) + \gamma_n^v v(w_n,y_n), \tag{10}$$
$$u(w_n) = E[u_n(w_n)|w_n] \tag{11}$$
$$v(w_n,y_n) = E[v_n(w_n,y_n)|w_n,y_n] \tag{12}$$

The definitions of u and v imply that $E[\gamma^u] = E[\gamma^v] = 1$. These heterogeneity parameter can, however, be correlated with w_n, y_n (just as in linear fixed effects models for instance). While actually any cost function (2) can be written this way, we now restrict firm heterogeneity to be stochastic and exogenous:

Assumption 5.
(i) The parameters γ_n^u and γ_n^v are stochastic and their realisations are given to the firm.
(ii) Firms know their technology $\gamma_n = (\gamma_n^u, \gamma_n^v)$ before producing and competing à la Cournot.

A5 ensures that the heterogeneity terms are not a function of further explanatory variables of the cost function, that they are exogenous to the firm, in the sense that they do not (systematically) change with w_n, y_n. This assumption can be justified by the fact that the choice of the technology is made just before the firm first entered the market, and the current value of γ_n^u and γ_n^v are considered as (conditionally) random technological shocks. Note that an increase in γ_n^u or γ_n^v corresponds to a negative technological shock while a decrease in these parameters represents technological progress. More restrictive versions of A5 are found in the literature, assuming either that $\gamma_n^u = 0$ (Jovanovic, 1982), $V[\gamma_n^u] = 0$ (Hopenhayn, 1992), γ_n^v iid (Jovanovic, 1982), γ_n^v independent of γ_n^u (Bresnahan and Reiss, 1991). We aim to stay general in following our purpose of estimating the joint distribution of γ.

The variable cost heterogeneity parameter γ_n^v is related to the additive "total factor productivity" term ω_n often considered in the context of production functions. When $y = \omega_n f(x)$ where

x denotes a vector of inputs, and the production function is linearly homogeneous in x (which is equivalent to v being linearly homogeneous in y) then $\gamma_n^v = 1/\exp(\omega_n)$. Production functions compatible with the bi-dimensional heterogeneity like (10) in the cost function are described by Chen and Koebel (2017).

Figure 1 represents five zones in which different types of firms can be locked in. In zone I, firms exhibit higher than average variable costs and relative low fixed costs. These type of firms can enter or exit the market without bearing high sunk cost. Zone II corresponds to a zone of generalized inefficiency: firms exhibit both higher fixed and variable costs. Firms located in zone III are extremely efficient and able to produce with fixed and variable costs lower than average. Zone IV comprises firms producing with lower than average variable costs and higher fixed costs. In zone V, firms operate with an average technology and are similar to a representative firm characterized by $E[\gamma^u] = E[\gamma^v] = 1$.

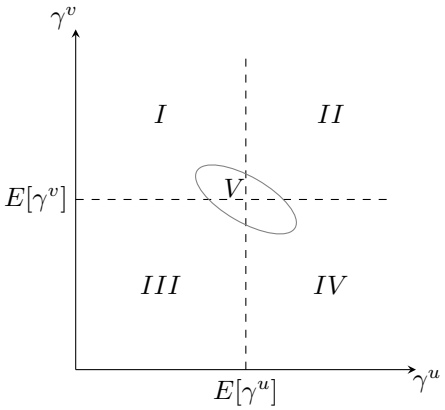

Figure 1: Five technological zones

In each different zone depicted on Figure 1, firms are not only different with respect to their technology, but we also expect to see difference in the levels of the endogenous variables.

Proposition 2. Under A1-A5, at the short-run Cournot equilibrium with fixed number of firms,
(i) firm i individual production level decreases with γ_i^v,
(ii) firm i production level increases with γ_j^v,
(iii) the aggregate equilibrium level of production decreases with γ_i^v,
(iv) individual and aggregate production levels are unaffected by a change in γ_i^u.
(v) firm i profit decreases with γ_i^v and γ_i^u;
(vi) firm i profit increases with γ_j^v.

This result, proven (for completeness) in Appendix A, follows (as usual) from the first and second order optimality conditions and the fact that the marginal cost function is positive. It has been generalized by Acemoglu and Jensen (2013) to cases with multiple equilibria. Related results for input demands have been derived by Koebel and Laisney (2014). For output supply, Février and Linnemer (2004) derive a similar result, but for the case of constant marginal costs. It is intuitive that an increase in firm i's marginal cost (through higher γ_i^v) decreases its output, but quite messy to prove due to firm heterogeneity and the existence of aggregate Cournot effects in the backward reaction functions. According to this result, we expect to see bigger firms located in zone III or IV of Figure 1. It is noteworthy (P2ii) that despite the output level of all competing firms decreases after a favorable productivity shock on i, the aggregate Cournot output is increasing, too (P2iii). This means, cost reducing technological change hurts firms that are not affected by it, which loose market shares, but aggregate production in the industry increases. The increase in market size outweighs the redistributional effect in the market shares.

Assumption A5 does not introduce any restriction about the relationship between γ_n^u and γ_n^v, and we considered in P2 that both variables could be shifted independently the one from the other. We now introduce a form of interrelation between them. The parameter γ_n^v reflects the efficiency of the variable cost function, the lower it is, the better for the firm. Conversely, the parameter γ_n^u

is often considered as an inefficiency, increasing the level fixed cost. However, from microeconomic theory, we know that it is likely that a higher fixed cost usually allows a firm to produce at a lower marginal cost, at least for some range of the output level. See for instance Chen and Koebel (2017) for the theoretical foundations and an empirical investigation. Let us restate this relationship explicitly:

Assumption 6. The variable cost efficiency is a transformation of the fixed cost efficiency:

$$\gamma^v = e(\gamma^u) + \eta, \tag{13}$$

with function e decreasing and strictly convex, and the random term η iid, with an expectation equal to zero, constant variance and uncorrelated with γ^u.

Function e transforms the firm specific fixed cost efficiency γ_n^u into a variable cost efficiency γ_n^v characterizing firm n's production technology. It is identical for all firms (within an industry), because e represents the mean technological frontier between the different types of production possibilities. A6 implies that, on average, technological progress is not transmitted through simultaneous reductions in both cost parameters γ_n^u and γ_n^v, but there is a trade-off characterized by e. A6 has an interesting empirical implication:

$$cov(\gamma_n^u, \gamma_n^v) < 0. \tag{14}$$

This inverse relationship between fixed and variable costs is often neglected in international trade (compare with Melitz (2003) or industrial economics (see for instance Bresnahan and Reiss (1991)), where fixed costs are often considered as a pure inefficiency. We will test whether this assumption or instead our more general version stated in A6 is satisfied or not.

For our empirical investigation, we need still more unobserved heterogeneity than introduced so far, and some more restrictive cost functions. We assume that firms have quadratic cost functions:

Assumption 7. The variable cost function v_n is quadratic in production and exhibits heterogeneity in slope and curvature:

$$v_n(w_n, y_n) = \gamma_{1n}^v v_1(w_n) y_n + \frac{1}{2}\gamma_{2n}^v v_2(w_n) y_n^2, \tag{15}$$

and the heterogeneity terms $\gamma_{1n}^v, \gamma_{2n}^v$ are stochastic and satisfy A5 and $E[\gamma_{1n}^v] = E[\gamma_{2n}^v] = 1$.

The quadratic specification of the cost function stated in A7 is compatible with the criteria of local flexibility of the cost function, which is shown to be important for empirical investigations (Diewert and Wales, 1988). The family of cost functions defined by (2) and (15) is able to approximate a variety of cost functions usually considered in the literature. We introduce three multiplicative firm specific terms $\gamma_n^u, \gamma_{1n}^v$ and γ_{2n}^v to capture heterogeneity over firms, in both the levels of fixed and variable costs and in the slope of the variable and marginal costs. This is more general than the uni-dimensional cost heterogeneity considered by Panzar and Willig (1978). Specification (2), (15) generalizes the heterogeneous fixed cost specification of Spulber (1995) (who sticks to the constant marginal cost assumption). It also extends the heterogeneous (but constant) marginal cost specification of Bergstrom and Varian (1985) and of Salant and Shaffer (1999). While uni-dimensional heterogeneity in marginal cost is useful to allow for unobserved heterogeneity in the level of firms' output, bi-dimensional heterogeneity is important to explain why the growth rate of firms with the same output levels can be different.

The specification of heterogeneity given in (15) is compatible with the former version given in (10) if we define overall variable cost heterogeneity γ^v as a weighted average of $\gamma_{1n}^v, \gamma_{2n}^v$ as

$$\gamma_n^v = \frac{\gamma_{1n}^v v_1(w_n) y_n + \frac{1}{2}\gamma_{2n}^v v_2(w_n) y_n^2}{v(w_n, y_n)}, \tag{16}$$

where the variable cost function v is identical for all firms and defined by evaluating v_n at the mean values $E[\gamma_{1n}^v] = E[\gamma_{2n}^v] = 1$, that is

$$v(w_n, y_n) = v_1(w_n) y_n + \frac{1}{2} v_2(w_n) y_n^2. \tag{17}$$

While the three-dimensional technological heterogeneity in $(\gamma_n^u, \gamma_{1n}^v, \gamma_{2n}^v)$ is important from an empirical viewpoint, the two-dimensional representation of (γ_n^u, γ_n^v) based on (16) is helpful for economic interpretation as well as for drawing (two-dimensional) plots and figures.

For $\gamma_{2n} > 0$, the firm specific average cost function is U-shaped if $u_n > 0$ and $v_{2n} > 0$ and reaches its minimum for production level $\underline{y}_n = \sqrt{2\gamma_n^u u/(\gamma_{2n}^v v_2)}$. The efficient scale of production can therefore be different from one firm to the other (for unobserved technological reasons). The quadratic specification is convenient as it allows to obtain an explicit solution for Cournot's equilibrium in terms of (nonnegative) individual and aggregate production levels:

$$y_n^b(w_n, Y) = \frac{P(Y) - \gamma_{1n}^v v_1(w_n)}{\gamma_{2n}^v v_2(w_n) - P'(Y)}, \tag{18}$$

$$Y^N = \sum_{n=1}^{N} y_n^b(w_n, Y^N). \tag{19}$$

This highlights that the firm level of production at the equilibrium $y_n^N = y_n^b(w_n, Y^N)$ does not only depend upon aggregate output and input prices, but also upon the technological parameters γ_n. Equation (18) also illustrates that *ceteris paribus*, the higher the variable cost the lower the production level y_n^N (see P2iii) if both $\gamma_{1n}^v \geq 0, \gamma_{2n}^v \geq 0$.

Averaging the first order optimality conditions over firms yields

$$P(Y^N) + P'(Y^N)\overline{y}^N = \overline{v}_1 + \frac{1}{N}\sum_{n=1}^{N} v_{2n} y_n^N. \tag{20}$$

The Cournot equilibrium is fully characterized by the average marginal cost. Firms do not need to precisely know the values of (v_{1n}, v_{2n}) of each of their competitors to figure out the Cournot equilibrium: some distributional statistics are sufficient, as the number N of competitors, the sample averages of the marginal cost terms $\overline{v}_1, \overline{v}_2$, and the covariance $cov(v_{2n}, y_n^N)$ between the slopes of the marginal cost and the elementary production levels. Contrary to the case with constant marginal costs, considered by Bergstrom and Varian (1985), the way production and slope characteristics are jointly distributed over firms matters at the equilibrium. This extension also allows firms to respond heterogeneously to exogenous changes in costs and demand.

In order to derive further interesting results, we consider a more restrictive form of heterogeneity characterized by:

Assumption 8. The variable cost heterogeneity is unidimensional the sense that:

$$\gamma_{1n}^v = \gamma_{2n}^v > 0. \tag{21}$$

A8 reduces the dimension on heterogeneity and allows us to focus only on marginal cost heterogeneity instead of having to discuss the first and second derivative of the cost function explicitly. Under A8, γ_n^v defined in (16) is independent of (w, y). The restriction (21) could be weakened and is not necessary for the empirical part of the paper, but it is interesting for giving further intuition on the drivers behind our empirical findings which can hold (by continuity) in cases where A8 is not satisfied.

Proposition 3. Under A1-A8, we consider two firms at Cournot equilibrium, both with similar input prices w and random term η. The Nash equilibrium production levels of firms i and j satisfy $y_i^N < y_j^N$ iff
(i) the biggest firm is more productive: $\gamma_i^v > \gamma_j^v$
(ii) the biggest firm has a lower variable cost for each unit produced: $v_i\left(w, y_i^N\right)/y_i^N > v_j\left(w, y_j^N\right)/y_j^N$
(iii) the biggest firm has higher fixed costs: $\gamma_i^u < \gamma_j^u$ and $u_i(w) < u_j(w)$;
(iv) the biggest firm has a larger efficient scale of production.

P3 implies that when firms are heterogeneous in their technologies, these differences induce them to choose different operating sizes, yielding a relationship between firms' production levels and their technological characteristics. If we order firms along their output level (from the smallest to the biggest), there is equivalently a corresponding ordering of the technological parameters γ^v

and the variable unit cost of production. For the fixed costs and the efficient scale of production, the ordering is not perfect, but subject to random errors in the relationship between fixed and variable costs. On average, however, the order is preserved.

The aggregate production Y^N implicitly defined in (19) also depends upon the number N of active firms, we now study entry and exit and how adjustment in N affects the main results of this section.

3 The long-run Cournot equilibrium

We now characterize a long-run Cournot equilibrium (LRCE) as a short-run Cournot equilibrium in which the number of active firms adjusts to exhaust expected profit opportunities. Firms choose either to enter or exit the market using available information. We denote by $\mathcal{N}$ the set of firms indices which are active, and by $\mathcal{M}$ the set of firms' indices which are inactive. The LRCE corresponds to a game in which firms choose their activity and production levels simultaneously, see Lopez-Cuñat et al. (1999) who also compares the simultaneous game with the one where entry and production choices are sequential. Active firms incur a fixed cost $c_n(w_n, 0^+, \gamma_n) = u_n(w_n)$ and inactive firms have $c_n(w_n, 0, \gamma_n) = 0$.

Active firms expect nonnegative profits and all potential entrants expect nonpositive profits. We introduce the superscript C to characterize long-run Cournot outcomes y_n^C and Y^C. Conditionally on observables, the cost function is subject to randomness due to unknown technological progress at the beginning of the period (see A5). It turns out that aggregate production, individual production, and profits are also random, hence, the entry/exit condition defining the LRCE is given by:

$$E\left[P\left(Y^C\right) y_n^C - c_n\left(w_n, y_n^C\right)\right] \geq 0, \tag{22}$$

$$E\left[P\left(Y^C + y_m\right) y_m - c_m\left(w_m, y_m\right)\right] \leq 0, \tag{23}$$

for any $n \in \mathcal{N}$ and $m \in \mathcal{M}$. The expectation operator E denotes the (rational) expectation with respect to the technological shocks γ_n which are random (and whose distribution is conditional to information available to the firm at the time of decision). We assume that conditions (22) and (23) are satisfied by the data generating process. Acemoglu and Jensen (2013, Theorem 1) or Okumura (2015, Theorem 1) showed that under A1-A4 the LRCE with heterogeneous firms exists. The equilibrium is not unique however: different histories condition the expectations in (22) and (23). The distribution of the technological shocks are conditioned by the firms' specific history: entering firms draw γ_{nt} from a different distribution than firms which have already experienced 20 or 40 years of activity and which have reached some size. We follow Novshek (1984) and Acemoglu and Jensen (2013) and consider that firms cannot change their technology without further cost. Conditionally on observables, differences in the technology over firms (and time) is random (see A5). This is different from Götz (2005), Acemoglu and Jensen (2013) (section 5.4) and Ledezma (2021) who consider that firms can choose their production technology optimally. In this context, only the more efficient technologies are chosen, with the consequence that, at equilibrium, firms tend to be similar in technology and firm size. It would be quite a challenge with this approach to endogenously generate a distribution of firms' sizes close to those usually observed in a given industry.[2]

4 Welfare and LRCE

We now consider the welfare implications of the observed distribution of output and investigate, following Mankiw and Whinston (1986) the welfare loss at LRCE. In a setup with identical firms, Mankiw and Whinston have shown that under business stealing (see P1v), the free entry equilibrium leads too many firms to enter the market in comparison to what is optimal from the welfare viewpoint. This result has been extended by Amir et al. (2014) to a setup where the planer controls either entry (but not production) or entry and production. In our situation with heterogeneous firms, the central planer has to carefully consider technological differences when deciding which firm is allowed to produce and how much. We assume that she knows the technological parameters

[2] We are aware that even in a setup with homogeneous firms, we can end up with asymmetric Cournot equilibrium, see for instance Novshek (1984). The corresponding distribution of firm sizes is very restrictive, however.

γ_n of each firm. The welfare function is similar to the one of Mankiw and Whinston (1986):

$$W\left(\{y_n\}_{n=1}^M, \{\gamma_n\}_{n=1}^M\right) = \int_0^{\sum_{m=1}^M y_m} P(s)\, ds - \sum_{m=1}^M c(w_m, y_m, \gamma_m) \tag{24}$$

Note that all M firms are considered as potential contributor to economic activity in W.

4.1 Short-run optimal welfare / The optimal distribution of production

In the short run, the planer has to decide whether firm m is entitled to produce or not, and how much each firm produces, for given firm level technological choices (there is neither entry nor exit). In this context, the welfare maximizer is able to remove some inefficiencies that are introduced by markups and imperfect competition. Technological characteristics are exogeneous, and the output levels at set such that:

$$W^S \equiv \max_{\{y_n\}_{n=1}^M} \left\{W\left(\{y_n\}_{n=1}^M, \{\gamma_n\}_{n=1}^M\right) : \{y_n \geq 0\}_{n=1}^M\right\}.$$

The Short-Run Optimal Welfare (SROW) is characterized by the first order Kuhn and Tucker necessary conditions for an inner maximum for W:

$$P\left(\sum_{m=1}^M y_m\right) = \frac{\partial c_n}{\partial y_n}(w_n, y_n) - \lambda_n, \qquad y_n \geq 0, \qquad \lambda_n \geq 0, \qquad \lambda_n y_n = 0, \tag{25}$$

for $n = 1, \ldots, M$. The welfare optimizing individual and aggregate productions are denoted by y_n^S and Y^S. It follows that a welfare maximizer:
(i) sets the production level of active firms to equalize price and marginal cost ($y_n^S > 0 \Rightarrow \lambda_n^S = 0$).
(ii) shuts down any firm with a marginal cost above the price: if $\partial c_m/\partial y_m(w_m, y_m) > P(Y_{-m} + y_m)$ for any y_m then $y_m^S = 0$ and $\lambda_m^S = \partial c_m/\partial y_m(w_m, 0) - P(Y_{-m}) > 0$.

A3(ii) ensures that W is concave in y_n at $y_n^S > 0$, and that the above first order conditions are sufficient for y_n^S to maximize W. Condition (25) requires that at the optimum, all active firms produce with the same marginal cost, which contrasts with LRCE at which active firms are characterized by a price above their firms' marginal cost. Some firms active at the LRCE will no longer be active at the SROW: a lower price $P(Y^S) < P(Y^C)$ calls for lower marginal cost by (25), but firms producing less and having little market power, will typically have difficulties to cope with this requirement. It also follows from (25) that at the social optimum, active firms with positive profits exhibit (local) decreasing returns to scale: $Py_{nt}/c > 1 \Leftrightarrow \varepsilon(c; y) > 1$ (and firms with negative profits have increasing returns). We state a result extending the one of Mankiw and Whinston (1986) to a setup with heterogeneous firms.

Proposition 4. Assume A1-A5 and A8. In comparison to the SROW, the LRCE is characterized by
(i) a lower aggregate production and a higher price: $Y^C < Y^S$ and $P(Y^C) > P(Y^S)$;
(ii) Welfare is too low: $W^C \leq W^S$, and profits are too high: $\pi_n^C > \pi_n^S$
(iii) big firms which produce too little, $y_n^C < y_n^S$
(iv) small firms with global decreasing returns which produce too much: $y_n^C > y_n^S$, and some of them should be shut down
(v) small firms with increasing returns which either produce too little, or should be shut down
(vi) a subset of the firms active at LRCE is still active at the SROW: $N^C \geq N^W$.

The proof of P4 (see Appendix A) is constructive in the sense that it characterizes which firm is producing more and which one will be inactive at SROW. It also defines a big firm as a firm with a level of production at LRCE such that its marginal cost of production is too low for welfare maximizing:

$$\frac{\partial c_n}{\partial y}\left(w_n, y_n^C\right) < P\left(Y^S\right),$$

and conversely for a small firms. This result is also useful for our empirical purpose of simulating firms' size distribution at the SROW. We use P4 to code the algorithm calculating the increase in aggregate production and the corresponding reallocation of output over firms at the SROW. Contrary to Mankiw and Whinston (1986), firms are differently affected by the new pricing rule,

however, most results they obtain in the homogeneous firms case carry over to an economy with heterogeneous firms. Instead of centralizing all production decisions, the central planer can equivalently introduce a tax and subvention scheme for inciting firms to produce at the socially optimal level. Comparing the conditions (25) and (3) we see that the aggregate production level of Y^S can be decentralized through the introduction of a sale tax τ specific to each firm and given by:

$$\tau_n(y) = \left| 1 - \frac{P(Y^S)}{P(Y^C_{-n} + y)} \right|.$$

Note that the sale tax rate is decreasing in y at LRCE and takes a value of zero at the SROW. See Guesnerie and Laffont (1978) for related results. An interesting consequence of P4 is the following:

Proposition 5. Under A1-A8, we consider firms with similar input prices w at Cournot equilibrium. Assume that the cost functions are convex. Then $N^S \leq N^C$ and the Hirschman-Herfindahl index of concentration is higher at the SROW than at LRCE.

P5 means that an efficient industrial policy should not try to minimize industry concentration at all costs. Actually, the opposite policy would improve welfare in the case of Cournot competition. A related corollary has been proposed by Salant and Shaffer (1999, Corollary 2), but for a situation where aggregate production stays constant. We generalize their result to the comparison of two situations with different levels of aggregate output since $Y^S \geq Y^N$. The economic intuition behind the result is as follows: for given N the Cournot equilibrium price is too high, $P\left(Y^N\right) \geq P\left(Y^S\right)$, and incites small and inefficient firms to enter the market, while for welfare maximization the planer prefers to increase the production of the technologically more efficient firms. Free entry decreases the long run Cournot prices such that at the LRCE there is no incentives for an efficient and potentially big firm to enter the market. The proof of P5 is provided in Appendix A, and is both a consequence of the properties of the Hirschman-Herfindahl index, than of P4, which states that the SROW is achieved through redistribution of output from the socially inefficient and smaller firms to the efficient and bigger firms. We, however, need to focus on convex technologies in order to exclude the occurrence of P4(v). We also reduce the dimension of heterogeneity sources and assume identical input prices. By continuity in w, P5 still applies if input prices are close enough but not strictly identical for firms n and m.

4.2 Long-run optimal welfare / The optimal technology

In the long-run, the planer also has an entrepreneurial duty and selects the production technologies that will be active at Long-Run Optimal Welfare (LROW). The planer can replicate some production technologies in order to maximize welfare. In a decentralized economy, in contrast, the type of technology is private knowledge of the entrepreneurs. Although there is a financial incentive to adopt most efficient technologies, both the firm size distribution and productivity distribution provide evidence for large differences in technologies.

While at SROW, a firm producing nothing bears the fixed cost u_n, in the LROW, the cost of inactivity is zero (the planer forbids entrance of such a firm). The resulting discontinuity of the cost function at $y_m = 0$ has now to be treated more carefully. A second difficulty is that the planer now has to decide which technologies to activate and to replicate in the long-run. Formally:

$$W^L \equiv \max_{\{y_n,\gamma_n\}_{n=1}^M} \left\{ W\left(\{y_n\}_{n=1}^M, \{\gamma_n\}_{n=1}^M\right) : \{y_n \geq 0\}_{n=1}^M \wedge \{\gamma_n\}_{n=1}^M \in \Gamma \right\}. \tag{26}$$

The technological set $\Gamma \subset \mathbb{R}^2$ denotes the set of all technologies active at LRCE. The long run optimal value satisfies $W^L \geq W^S$. The main difficulty if we solve this problem by evaluating W over all discrete elements of Γ, is that it is time consuming: for a given industry there are M^M ordered arrangements of all elements in Γ. For each arrangement it is necessary to compute the optimal individual and aggregate output levels by solving (25) which is computationally not feasible. Fortunately, a useful property for reducing the set of candidate technologies for optimal welfare are available. Under A1 to A7, the SROW individual output quantities y_n^S are nonincreasing in γ_n^v, and the same applies to the aggregate optimal production Y^S. This implies that all LROW optimal γ parameters belong to the technological frontier, defined as the lower (nonconvex) hull of the technological parameters as:

$$\Gamma_L = \{\gamma_n \in \Gamma : \nexists \gamma_m \in \Gamma \wedge \gamma_m < \gamma_n\}. \tag{27}$$

While the LRCE is an equilibrium with free entry, there not with free technological choice. These choices are constrained by firm specific histories, which explains why at LRCE the cost functions are heterogeneous (it is a long-run equilibrium in the sense that free entry and exit occurs). The SROW is an optimum at which inefficiencies of imperfect competition are removed by the planer, but there is neither entry nor exit and firms specific technologies are given to the planer. This last constraint is removed at the LROW, where the planer can freely choose the same technology for new and more experienced firms. At LRCE and at the SROW, cost functions are heterogeneous, which is compatible both with the literature on existence and uniqueness of Cournot equilibrium and the recent literature on productivity and industrial organization. There is also a long tradition considering the technological long-run, in which the cost function is homogeneous over all firms (see for instance Mankiw and Whinston (1986)). Both traditions can be related using the long-run cost function, defined as:

$$c^L(w, y) = c(w, y, \gamma^L), \tag{28}$$

where the long-run technological parameters γ^L are fixed at their optimal level (which varies with w, y in general). This notation allows to restate some textbook results, which are later useful to assess the gap between actual data and this benchmark, as well as for coding numerical simulations.

Proposition 6. Under A1-A8, we consider firms with similar input prices w, and ignore the integer constraint on N. Then
(i) the LROW exists and is unique;
(ii) at LROW all firms have zero profit and local constant returns to scale;
(iii) $W^L \geq W^S$;
(iv) the fixed cost is zero at LROW if $e'(\gamma^{uL}) < u(w)/v(w, y^L)$;
(v) It is equivalent to maximize the central planer problem W^L or decentralized profits wrt (y_n, γ_n), for given price level which clears the product market with free entry;

P6 connects the literature on heterogeneous and homogeneous technologies: at LROW the optimal technological choice is unique, all active firms use the same technology. The distribution of firm sizes degenerates to a mass point at y^L. This implies that the Hirschman-Herfindahl index of concentration is $1/N^L$. (If we consider the integer constraint, then a further technologies could be used at LROW in order to produce the residual output level.) At LROW, firms produce at the minimum of the average cost functions, which characterizes local CRTS. It is not surprising, given that the planer maximizes welfare with less technological constraints than in the short run, that $W^L \geq W^S$. Less common is (iv) compatible with the use in the long-run of a technology with positive fixed costs. If y^L is small enough, however, the threshold $u(w)/v(w, y^L)$ in P6(iv) decreases, and the planer switches to a technology with no fixed cost, in which case $\gamma^{vL} = e(0)$. See also Chen and Koebel (2017) for further details on such technologies.

It is not possible to conclude that at LROW $c^L/y^L \leq c^S_n/y^S_n$, because lower average cost functions are achieved a the price of a higher fixed cost, which is not necessarily efficient if the increase in total output enabled by the change in technology is too small. It is neither true that $Y^L \geq Y^S$, nor $N^L \leq N^S$ (or the converse) are necessarily satisfied. Regarding the total number of firms active at LROW: the planer closes all firms producing nothing at SROW (and avoid bearing the fixed cost), and replicates the most efficient firm. In the long-run, the number of active firms crucially depends upon the shape of $e(\gamma^u)$, which is an empirical issue. As a corollary, in the same vein as P5 (comparing SROW and LRCE), it is possible that the Hirschman-Herfindahl index of concentration is higher at LROW than at SROW.

In the empirical part of the paper, we do not observe but estimate $\gamma_n = (\gamma^u_n, \gamma^v_{1n}, \gamma^v_{2n})$, and refrain to identify most efficient technologies using (27) to avoid the selection of too efficient technologies. Instead, a more robust procedure (to extreme values and outliers) consists in estimating the technological frontier nonparametrically as the locus of all $\underline{\gamma} \in \mathbb{R}^3$ satisfying:

$$\Pr[\gamma_n \leq \underline{\gamma}] = 0.2. \tag{29}$$

We choose the 0.20^{th} quantile of the distribution of γ_n, because the planer has to achieve a sufficient level of efficiency on the technological frontier to justify its existence. In order both to represent and simplify the solution, we project the 3-dimensional heterogeneity onto the two dimensional

space: using (16) we consider the LR efficiency frontier to be defined as the set of technologies $\{\underline{\gamma}^u, \underline{\gamma}^v\}$ satisfying:

$$\Pr[\gamma_n^u \leq \underline{\gamma}^u, \gamma_n^v \leq \underline{\gamma}^v] = 0.2. \tag{30}$$

When the CDF is strictly monotonic, we can rewrite equivalently the technological frontier (30) as:

$$\gamma_n^v = e(\gamma_n^u), \tag{31}$$

where e is nonincreasing. Substituting this constraint into the long-run social welfare function allows us to characterize the LROW first order conditions, given by (25) and

$$u(w_n) + e'(\gamma_n^u)v(w_n, y_n) = \kappa_n, \qquad \gamma_n^u \geq 0, \qquad \kappa_n \geq 0, \qquad \kappa_n \gamma_n^u = 0, \tag{32}$$

where κ_n denotes the Lagrange multiplier associated to $\gamma_n^u \geq 0$. The planer obtains a specific optimal technological choice for each firm as a function of its observed characteristics. All solutions can be grouped into two sets: either the firm is inactive at LROW in which case $y_n = 0$, and so $\gamma_n^u = 0$, or the firm is active. An inner solution $(y_n > 0, \gamma_n^u > 0)$ is characterized by the equality between the elasticity of variable cost heterogeneity wrt fixed cost heterogeneity and the share of fixed to variable cost:

$$-\frac{\gamma_n^u}{\gamma_n^v} e'(\gamma_n^u) = \frac{\gamma_n^u u(w_n)}{\gamma_n^v v(w_n, y_n)}. \tag{33}$$

Active firms can operate without fixed cost if the fixed cost is too expensive, and the reduction in variable cost not high enough. In this case, $\kappa_n > 0$, and $\gamma_n^v = e(0)$. The social planer shuts down or upscales those inefficient firms whose technological parameters are not located on the efficient frontier. These firms are now able to produce more output than before (just as in P2(i), the individual optimal output levels increase when the γ_n^v decrease). Total output Y^S increases to Y^L, the price decreases from $P^S = P\left(Y^S\right)$ to P^L and the welfare increases.
It is not clear, in general, how the number of firms and the Hirschman-Herfindahl index of concentration varies, this depends upon the slope of e. If the frontier is flat, investment in fixed cost does not reduce the variable cost of production and it is rational for a social planer to run the economy with many small firms with no fixed cost. If instead e decreases steeply, the social planer will close most small firms and upscale some of them. In this case concentration may either increase or decrease. The characterization of the LROW given in this section considerably reduces the computational burden of optimizing W jointly over y_n and γ_n and will be used in the empirical part.

Whether and by how much free entry leads to an excessive number of firms and a too low aggregate production is an empirical question which is studied below. With heterogeneous firms, we expect that free entry to be socially less beneficial than in both perfect and imperfect competitive models. The entry of inefficient firms prevents more efficient firms to reach the optimal scale of production.

5 Data and descriptive statistics

We use French fiscal data available at the firm-level for the years 1994 to 2017 (FICUS and FARE data).[3] The data comprises the universe of active firms, but we consider only those belonging to the manufacturing industry.[4] The observations contain information on firms' balance sheet and income statements, where each firm is identified by a specific identification number, which is constant over time. Table 1 lists the manufacturing sectors considered with the corresponding number of firms and observations.

A basic data cleaning consisted to exclude observations with negative values for sales, labour cost, material cost and capital cost. We also find several extreme values for the profit rates,

[3]FICUS and FARE refer to "fichier de comptabilité unifié dans SUSE" and "fichier approché des résultats d'Esane", respectively. That is, FICUS was part of the French firm-level database SUSE and was replaced in 2008 by FARE that, in turn, belongs to the current database Esane.

[4]We exclude the industry for food processing (10), the manufacture of tobacco products (12), and the manufacture of coke and refined petroleum products (19). The industry 10 is excluded as it comprises the overwhelming part of the total number of firms and should, in our view, be treated separately. The industries 12 and 19 are excluded for the reason of a very low number of observations. See Appendix B for more details.

and decided to consider only observations with a profit rate included between -90% and 500%, which corresponds to 95.2% of the sample. This leaves us with 1,366,608 observations and 158,769 different firms.

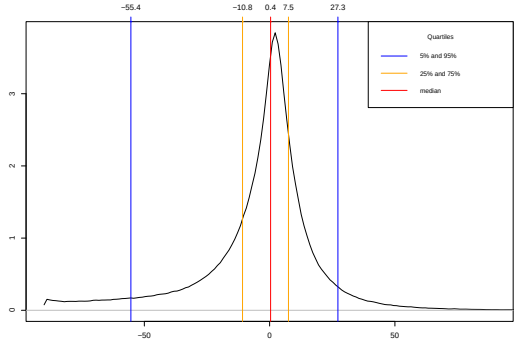

Figure 2: The density of profit rates

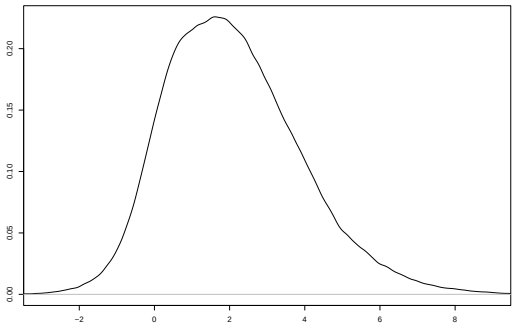

Figure 3: The density of log-levels of production

Profit rates are defined as $(py/c-1)\times 100$, and actually represent pure profit rates, as the user cost of capital is included in the cost of production. The empirical distribution of the observed profit rates cost is given on Figure 2 and illustrates that the data are fundamentally heterogeneous. The density of the profit rates mimics the distribution of the average cost since by definition $py/c-1 = p/(c/y)-1$. The density is not normal, but asymmetric and exhibits a large tail for values below the median (of 0.4%) and a thin tail above. The density of the log-levels of production is drawn on Figure 3 (over all firms and years).

Table 1: Description of two-digit industries

Industry[a]	Description	# Firms[b]	# Obs.[c]
11	Beverages	3,031	26,049
13	Manufacture of tobacco products	7,012	59,299
14	Manufacture of wearing apparel	15,658	82,221
15	Manufacture of leather and related products	3,054	22,220
16	Manufacture of wood and of products of wood	13,220	109,643
17	Manufacture of paper and paper products	2,825	28,447
18	Printing and reproduction of recorded media	21,799	174,024
20	Manufacture of chemicals and chemical products	5,204	47,581
21	Manufacture of basic pharm. products and pharm. preparations	979	8,522
22	Manufacture of rubber and plastic products	8,801	86,595
23	Manufacture of other non-metallic mineral products	11,668	95,613
24	Manufacture of basic metals	2,042	18,767
25	Manufacture of fabricated metal products	34,397	326,264
26	Manufacture of computer, electronic and optical products	7,388	57,119
27	Manufacture of electrical equipment	5,033	42,623
28	Manufacture of machinery and equipment	13,362	111,735
29	Manufacture of motor vehicles, trailers and semi-trailers	4,013	35,857
30	Manufacture of other transport equipment	1,799	12,852
31	Manufacture of furniture	15,355	109,952
	Total	176,640	1,455,383

[a)] Statistical classification of economic activities in the European Community, Rev. 2 (2008)
[b)] # Firms describes the number of firms which were active over the period (it is computed as the total number of different firms identifiers).
[c)] # Obs. describes the total number of observations.

Variables

Firm specific data are mainly nominal values and cover the value of production, total labor costs, the value of intermediate inputs, as well as the capital stock. Firms' nominal production is measured by the sum of firms' sales, stocked production, and production for own use. The value of intermediate inputs is given by firms' expenditures for raw materials and other intermediary goods. As proxy for firms' capital stock we use the amount of tangible assets reported in the

balance sheet. We use industry specific price indices (at a two-digit aggregation level) in order to convert the nominal values in real terms.[5] The wage level is firm specific and is obtained by dividing the labor costs by the number of employees. These calculations yield the firms' total production y_{nt}, and input vector $x_{nt} = (x_{k,nt}, x_{l,nt}, x_{m,nt})^\top$ as well as price indices p_{nt} for output and inputs $w_{nt} = (w_{k,nt}, w_{l,nt}, w_{m,nt})^\top$. In order to calculate the user cost of capital, $w_{k,nt}$, we follow Hall and Jorgenson (1967) and set $w_{k,t} = w_{i,nt}(1+r_t) - w_{i,n,t+1}(1-\delta_{nt})$, with $w_{i,nt}$ denoting the price index for investment (available at the industry level), r_t is the long-run rate of interest and δ_{nt} the annual rate of capital depreciation.[6] Note that, for our purpose, we only keep those firm observations with values larger than zero in capital stock, number of employees, intermediate inputs, and production.

Descriptive statistics

Table 2 shows the average number of firms active in a typical 4-digit industry, as well as the distribution of firm sizes over the 1994-2016 period. In our cleaned sample, over all industries and years, there are about 176,640 firms active in the French manufacturing, which represent 1,455,383 observations. At the 4-digit level the number firms is obtained by dividing the total number of observations by 184×23 (the number of 4-digit industries times the number of years), which yields an average number of 340 active firms. See Appendix B for further details on the data cleaning. The table also reports the average number of firms by different firm size (measured by the number of employees). It shows that the number of firms globally is decreasing in firm size. On average, most firms have between 2 to 4 employees, representing a share of about 24% of all firms. Table 2 also informs about market concentration in a typical 4-digit industry: firms with less than 20 employees represent about 75% of all firms, and produce only 7% of total production, whereas the few firms with 500 employees and more produce about 53.1% of the aggregate (4-digit) production. These figures not only document that there are few actors detaining strong market power, but also that there is a large competitive fringe of smaller firms. In our view, this seems compatible with the theoretical Cournot model adopted here, which allows for unobserved technological differences between firms. This unobserved heterogeneity is important for yielding a size distribution of firms endogenously, and comparable with the observed distribution reported on Table 2.[7]

Table 2: Statistics by firm size in a typical 4-digit manufacturing industry[a]

Firm size[b]	# of firms	Share of firms	Share of employees	Share of production
1	50	14.71	0.40	0.28
2-4	82	24.12	1.86	1.05
5-9	73	21.47	3.93	2.19
10-19	52	15.29	5.67	3.56
20-49	49	14.41	12.29	9.14
50-99	16	4.71	8.83	6.91
100-199	9	2.65	10.76	9.28
200-499	6	1.76	14.83	14.47
500+	3	0.88	41.43	53.11
Total	340	100.00	100.00	100.00

[a] All figures represent averages over all 4-digit industries and years (1994-2016). Shares are given in %.
[b] Firm sizes are measured by the number of employees.

[5] The sectoral price data are available at https://www.insee.fr/fr/statistiques/2832666?sommaire=2832834
[6] The interest rate was provided by the Banque de France at: https://www.banque-france.fr/statistiques/taux-et-cours/taux-indicatifs-des-bons-du-tresor-et-oat. We calculate δ_{nt} at the industry level by considering the ratio between the consumption of fixed capital and fixed capital, see www.insee.fr/fr/statistiques/2383652?sommaire=2383694
[7] See also Table B4 in Appendix B, which is complementary to Table 2, and shows the same statistics but for each 2-digit industries.

6 Inverse output demand estimates

This section studies the output demand addressed to an industry $i = 1, \ldots, I$, and estimates the elasticity of output demand wrt its price. It corresponds to the inverse function of (1). The output price index is available at the two-digit industry level, for $I = 22$ industries, and for the same time range of 24 years as in our firm level data. For the estimation, 2 years are lost due to differencing (and so $T = 22$ years).

We consider the following parametric specification for the output demand to industry i:

$$\ln Y_{it} \quad = \quad \alpha_i + \alpha_Y \ln Y_{i,t-1} + \alpha_p \ln P_{it} + \alpha_{IM} \ln P_{it}^{IM} + \epsilon_{it}, \tag{34}$$

In addition to the (domestic) product price P_{it}, we include as regressor the price index P_{it}^{IM} for the imports of the corresponding goods which are close substitutes to domestic products. Industry fixed effects α_i are included, and, as adjustment of demand to the prices may not be instantaneous but under the influence of the lagged level of aggregate quantities, the variable $\ln Y_{i,t-1}$ is also taken in account. Further variables influencing demand are the economy wide GDP, unemployment rate and demographic variables. All these variables are not industry specific and could be captured by the time dummies (as in Koebel and Laisney (2016)). With only a 484 observations however, we choose not to overparameterize our model and consider the more parsimonious specification with 22 industry specific fixed effects and 3 parameters. The elasticity of demand wrt domestic product price is then given by α_p.

The industry specific effect can be correlated with the explanatory variables and the random term ϵ_{it} is correlated with $\ln P_{it}$ since in the aggregate product price adjusts to shocks. We eliminate the industry specific effect by differencing over time:

$$\Delta \ln Y_{it} \quad = \quad \alpha_Y \Delta \ln Y_{i,t-1} + \alpha_p \Delta \ln P_{it} + \alpha_{IM} \Delta \ln P_{it}^{IM} + \eta_{it}, \tag{35}$$

with $\eta_{it} = \Delta \epsilon_{it}$.

Several variables that shift the output supply (but not directly output demand) can be considered as instruments: they are correlated with $\ln P_{it}$ and uncorrelated with the random term η_{it}, so that $E[\eta_{it} z_{it}] = 0$. The $(L \times 1)$ vector z_{it} of instruments includes industry labour cost, the price of intermediate consumption, of exports and the price index of imports. Lagged values of the endogenous variables are also considered as exogenous. For each period, we include up to 3 lag values of $\ln P_{it}$ and $\ln Y_{i,t-1}$ in the list of instruments. This gives us a total of $L = 130$ instruments. Given a $(L \times L)$ weighting matrix $\mathbf{W}$, the GMM estimator is defined by minimizing in α:

$$\left(\sum_{i=1}^{I} \sum_{t=1}^{T} \eta_{it} z_{it}^{\top} \right) \mathbf{W} \left(\sum_{i=1}^{I} \sum_{t=1}^{T} z_{it} \eta_{it} \right) = \eta^{\top} \mathbf{Z} \mathbf{W} \mathbf{Z}^{\top} \eta \tag{36}$$

The random terms η_{it} and η_{js} are likely to be correlated, both between industries (which are interdependent) in a given year, and within a given industry over two consecutive time periods. So we use two-ways clustering and allow for heteroscedasticity, for contemporaneous dependence between residuals of different industries, and for temporal dependence within a given industry and consecutive time periods. See for instance Cameron and Miller (2015) for details about multi-ways clustering and Cameron et al. (2011) for a detailed discussion in the context of GMM. More formally, we assume that

$$\begin{array}{l} E\left[\eta_{is}\eta_{it}\right] = \sigma_{iist} \text{ for } |s-t| \leq 1, \\ E\left[\eta_{it}\eta_{jt}\right] = \sigma_{ijtt}, \\ E\left[\eta_{is}\eta_{jt}\right] = \sigma_{ijst} = 0, \text{ for } i = j \text{ and } |s-t| \geq 2 \text{ and for } i \neq j \text{ and } |s-t| \geq 1. \end{array}$$

As there is no possibility to consistently estimate these parameters, we are instead looking to consistently estimate the variance matrix $V[\widehat{\alpha}]$ of dimension $K \times K$. It is convenient to define the set $\mathcal{S}$ of indices of the dependent random terms:

$$\mathcal{S} = \left\{ i, j, s, t : (i = j, |s-t| \leq 1) \vee (i \neq j, s = t) \right\}.$$

The cardinality of this set is $I(3T-2) + I(I-1)T = 11572$ and increases with I and T. The GMM weighting matrix is estimated in a first step (using IV estimates $\widehat{\eta}_{it}$) by the inverse of

$$\widehat{\mathbf{B}} = \sum_{i=1}^{I} \sum_{j=1}^{I} \sum_{s=1}^{T} \sum_{t=1}^{T} z_{is} z_{jt}^{\top} \widehat{\eta}_{is} \widehat{\eta}_{jt} \mathbf{1}_{[i,j,s,t \in \mathcal{S}]},$$

where the dummy variable $\mathbf{1}_{[i,j,s,t\in\mathcal{S}]} = 1$ if the indices are included in the set $\mathcal{S}$ and 0 otherwise. An alternative (and easier to code) version of matrix $\widehat{\mathbf{B}}$ is:

$$\widehat{\mathbf{B}} = \mathbf{Z}^\top(\widehat{\eta}\widehat{\eta}^\top \circ \mathbf{S})\mathbf{Z},$$

where the $IT \times IT$ selection matrix $\mathbf{S}$ has an entry (h, j) equal to one if the random terms η_h and η_j are correlated, and zero otherwise. In our case, only about 5% of the elements of $\mathbf{S}$ are nonzero. The Hadamard (term by term) multiplication is denoted by $\circ$. One difficulty comes from the fact that $\widehat{\mathbf{B}}$ is not necessarily positive definite. The same applies to our estimated parameters' variance matrix:

$$V[\widehat{\alpha}] = (\mathbf{X}^\top\mathbf{Z}\widehat{\mathbf{B}}^{-1}\mathbf{Z}^\top\mathbf{X})^{-1},$$

where the matrices $\mathbf{X}$ and $\mathbf{Z}$ are respectively of dimension $(IT \times K)$ and $(IT \times J)$ with the number of instruments not smaller than the number of regressors $L \geq K$. We follow Cameron et al. (2011) and impose positive definiteness on the parameters variance matrix by setting negative eigenvalues to zero in the eigendecomposition.[8]

Table 3 reports the estimated values of the parameters along with their standard deviations. The estimates of the fixed-effects and first difference specifications of the output demands are given for the purpose of comparison in columns 1 and 2. Our preferred specification relies on GMM and the corresponding estimated parameter values are included in the range of the Fixed effects (FE) and the first difference (FD) estimates. The test for overidentification does not reject the validity of our instruments. Tests for the occurrence of autocorrelation in the η_{it} of order two and higher lead to rejecting this hypothesis. This rejection (together with the the high p-value of the over-identification test) supports the use of endogenous variables as instruments (with a lag of two periods and more). According to the GMM estimation results, the estimated short-run elasticity of demand with respect to price is -0.64 and is statistically significant at the 1% threshold. Domestic products and imports are substitutable with a cross price elasticity of 0.49. The coefficient of lagged output is estimated at 0.76 and found to be significant. This introduces a gap between short- and long-run price elasticities. The clustered standard errors are substantially smaller than the HAC-robust standard errors, probably because additional independence over spaced time periods is assumed when clustering.

Table 3: Output demand estimates

	FE	FD	FD-GMM
α_Y	0.92 (0.02)	0.05 (0.05)	0.76 (0.06), [0.03]
α_P	-0.12 (0.07)	-0.67 (0.17)	-0.64 (0.18), [0.08]
α_{IM}	0.04 (0.07)	0.55 (0.16)	0.49 (0.18), [0.07]
OIT	-	-	0.99

Notes. HAC robust standard errors are given in parenthesis, clustered standard errors are in brackets. OIT: p-value of the over-identification test, for the validity of the 130 orthogonality conditions.

These estimates are useful to calculate the inverse demand elasticity which is central in our model, and also for computing the long-run elasticities, characterized by $Y_{i,t-1} = Y_{it}$. These corresponding estimates are provided in Table 4. The inverse demand elasticity is obtained by $\varepsilon\left(P^d, Y\right) = 1/\varepsilon\left(Y^d, p\right)$ and is estimated to -1.56 in the short-run and -0.37 in the long-run. Standard errors are obtained using the delta-method (with the HAC variance matrix).

[8] We actually compare different methods for imposing positive definiteness, by either restricting matrix $\mathbf{S}$, $\mathbf{B}$, $\widehat{\eta}\widehat{\eta}^\top \circ \mathbf{S}$ or $V[\widehat{\alpha}]$ to be positive definite, the results were different but in all cases, the diagonal terms of the restricted variance matrix were much lower than the HAC variance matrix.

Table 4: Industry short- and long-run elasticities of output demand

	Short-run		Long-run	
	$\varepsilon\left(Y^d, p\right)$	$\varepsilon\left(P^d, Y\right)$	$\varepsilon\left(Y^d, p\right)$	$\varepsilon\left(P^d, Y\right)$
Estimate	-0.64	-1.56	-2.67	-0.37
s.e.	0.18	0.44	0.87	0.12

The short-run inverse price elasticity is substantial. With Cournot competition, there is an interesting relationship between the markup and the market share y/Y, parameterized by the inverse demand elasticity:

$$\frac{p}{\partial c/\partial y(w,y)} = \frac{1}{1+\varepsilon\left(P^d, Y\right) y/Y}. \tag{37}$$

Using the estimates of Table 4, we draw the estimated short- and long-run relationship between markup and market-share on Figure 4. Firms in the competitive fringe have a markup of 1. In conformity with point (iv) of Proposition 1, for which Figure 4 provides an illustration, the markup is monotonically increasing in market share. While in the short-run there is substantial markup for a firm having a market share of 20 to 30%, in the long-run this markup falls to the interval 1.08 - 1.12, which is quite small. However, in the short-run, sluggish adjustment toward market equilibrium price and quantity, according to the dynamic relationship (34) with strong anchoring to the lagged aggregate output level, confers substantial market power and a markup of 1.45 - 1.88 to the few firms with the biggest market share.

Our estimate of the inverse demand elasticity satisfies A1 and is also broadly compatible with A4. Indeed, when the inverse demand elasticity ϵ is constant,

$$P'(Y) + y_h P''(Y) = \epsilon \frac{P(Y)}{Y}\left[1 + (\epsilon - 1)\frac{y_h}{Y}\right],$$

which is negative for any individual market share satisfying $y_h/Y \leq 1/(1-\epsilon)$. Our estimate of this upper bound is a market-share of 39.1% in the short-run, and 73.0% in the long run. It turns out that the inequalities are respectively satisfied by 98.6% and 100% of the observations.

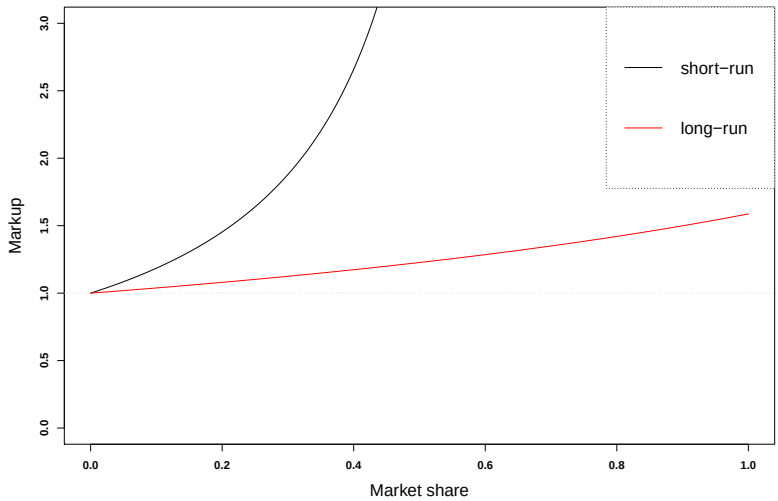

Figure 4: The markup and firms' market share

We also use these parameter estimates to build the instrumental variable $\widehat{Y}_{nt}^d$. This predicted demand to the market is correlated with individual output y_{nt} and orthogonal to the individual supply random term.

7 Cost function estimation with heterogeneity in fixed and variable costs

It is well known that unobserved heterogeneity causes estimation biases when it is neglected and correlated with the explanatory variables, see for instance Gouriéroux and Peaucelle (1990) or

Wooldridge (2010) for a detailed overview of the linear model. Unobserved heterogeneity also rises concerns about the incidental parameters, precluding consistent estimation of parameters and statistics of interest. Martin (2017) and Wooldridge (2019) consider unobserved multiplicative heterogeneity. When additive and multiplicative unobserved heterogeneity appears in the econometric specification, as is the case with our cost function, some specificities have to be considered. Then, the statistics of interest can be consistently estimated under some assumptions which are outlined below.

7.1 Empirical specification of the cost function

Given the quite long time dimension of our data, we now include a deterministic time trend, t, as a further argument of the cost function.

The first type of unobserved heterogeneity is specific to the production technologies and the cost functions characterizing a given industry. We deal with this difficulty, by estimating the cost specifications over all firms belonging to a given 2-digit industry (there are 19 different 2-digit manufacturing industries). Within a given industry, a further type of unobserved heterogeneity in the fixed and variable costs characterizes firms, and introduces correlation between their production and the random term. We propose a method for dealing with this endogeneity problem and avoiding estimation bias. As heterogeneity is unobserved it is subsumed in the additive random term and the cost function satisfies:

$$c_{nt} = u(w_{nt}, t) + v(w_{nt}, t, y_{nt}) + \epsilon_{nt} \tag{38}$$

$$\epsilon_{nt} \equiv u_{nt}(w_{nt}, t) - u(w_{nt}, t) + v_{nt}(w_{nt}, t, y_{nt}) - v(w_{nt}, t, y_{nt}) + \eta^c_{nt}. \tag{39}$$

We assume that the random term η^c_{nt} is such that $E[\eta^c_{nt}|w_{nt}, t, y_{nt}] = 0$. Its variance can exhibit heteroscedasticity and correlation. We use the reparameterization $u_{nt} = \gamma^u_{nt} u$ and $v_{j,nt} = \gamma^{v_j}_{nt} v_j$. For the sake of identification, we impose

$$E[\gamma^j_{nt}] = 1, \qquad j = u, v_1, v_2. \tag{40}$$

Cost heterogeneity is known by the firm, but unobserved by the econometrician. If we were able to control for unobservable heterogeneity, the condition $E[\epsilon_{nt}|w_{nt}, t, y_{nt}, \gamma_{nt}] = 0$ would be useful for parameter estimation. However, $E[\epsilon_{nt}|w_{nt}, t, y_{nt}] \neq 0$ due to the fact that ϵ_{nt} includes the unobserved heterogeneity terms of the fixed and variable cost function and the optimal optimal production level is decreasing in $\gamma^{v_j}_{nt}$ by (18) and Proposition 2. So $E[\epsilon_{nt} y_{nt}] \neq 0$ in (38) in general. Moreover, a firm can choose a high level of fixed cost if it allows to decrease its variable cost for (indirectly) achieving a higher production; in this case $E[\epsilon_{nt} y_{nt}] \leq 0$.

We are interested in identifying the cost functions u and v_1, v_2 which are common to all firms and time periods as well as the deforming weights $\gamma_{nt} = (\gamma^u_{nt}, \gamma^{v_1}_{nt}, \gamma^{v_2}_{nt})$. As there are three times more γ_{nt} parameters than observations, we will not be able to estimate them, but we will be able to approximate their joint and marginal distributions. Firm and time specific heterogeneity is interesting in order to account for technological differences between firms and over time.

We specify the parametric forms for u and v. We consider that u and v belong to the family of quadratic cost functions:

$$u\left(w, t; \theta^u\right) = \theta^\top_w w + \theta^\top_{wt} wt + \frac{1}{2} \frac{w^\top \Theta_{ww} w}{\zeta^\top w}, \tag{41}$$

$$v_1\left(w, t; \theta_1\right) y = \left(\theta^\top_{1w} w + \theta^\top_{1t} wt + \frac{1}{2} \frac{w^\top \Theta_{1ww} w}{\zeta^\top w}\right) y \tag{42}$$

$$v_2\left(w; \theta_2\right) y^2 = \left(\theta^\top_{2w} w\right) y^2 \tag{43}$$

The vectors of parameters θ_w, θ_{wt}, θ_{1w}, θ_{1t} and θ_{2w} have dimension $(J \times 1)$, whereas the symmetric matrices Θ_{ww} and Θ_{1ww} are $(J \times J)$. In order to identify the terms in the linear and quadratic fonctions of w, we impose that

$$\Theta_{ww} = \Theta^\top_{ww}, \quad \Theta_{1ww} = \Theta^\top_{1ww}, \tag{44}$$

$$\iota^\top \Theta_{ww} = \iota^\top \Theta_{1ww} = 0 \tag{45}$$

where ι denotes a $(J \times 1)$ vector of ones. We use the a Laspeyres price index $\zeta^\top w$ for normalization in order to impose linear homogeneity in w on the cost function. Both fixed and variable cost

functions are flexible, in the sense that they provide a second order approximation to an arbitrary fixed and variable cost function; see Chen and Koebel (2017), on this point. There is a total of $5J + J(J-1)$ free parameters. In our case, $J = 3$ and there are 21 free parameters in the cost function.

7.2 Unobserved heterogeneity

We still need to specify how unobserved heterogeneity can be identified and estimated. The main reason for which we have to take care about unobserved heterogeneity, is that neglecting it produces biased estimates. We are also economically interested in the distribution, because according to sections 2-4, it explains an important part of observed heterogeneity in firms size, size distribution, returns to scale, welfare loss. The cost function with unobserved heterogeneity:

$$c_{nt} = \gamma_{nt}^{u} u(w_{nt}, t; \theta^{u}) + \gamma_{nt}^{v_1} v_1(w_{nt}, t; \theta^{v_1}) y_{nt} + \frac{1}{2}\gamma_{nt}^{v_2} v_2(w_{nt}, t; \theta^{v_2}) y_{nt}^2 + \eta_{nt}^{c} \tag{46}$$

Let $w_n \equiv \{w_{ns}\}_{s\in\mathcal{T}_n}$, $y_n \equiv \{y_{ns}\}_{s\in\mathcal{T}_n}$ and $\mathcal{T}_n$ represents the set of all time indices for which firm n is observed. This definition allows us to interpret the functions u, v_j as the fixed and variable cost function of a representative firm, characterized by $\gamma_{nt}^{u} = \gamma_{nt}^{v_1} = \gamma_{nt}^{v_2} = 1$. We assume that, given γ, the additive random term satisfies strict exogeneity:

$$E(\eta_{nt}^{c} | w_n, t, y_n, \gamma_{nt}) = 0. \tag{47}$$

When cost heterogeneity is known by the firm, but unobserved by the econometrician, the firm knows $\gamma_{nt}^{u}, \gamma_{nt}^{v_1}, \gamma_{nt}^{v_2}$ when deciding about its output level, which is set to equalize marginal revenue and marginal cost:

$$p_t \left(1 + \varepsilon \frac{y_{nt}}{Y_t}\right) = \gamma_{nt}^{v_1} v_1(w_{nt}, t; \theta^{v_1}) + \gamma_{nt}^{v_2} v_2(w_{nt}, t; \theta^{v_2}) y_{nt} + \eta_{nt}^{p} \tag{48}$$

with the random term η_{nt}^{p} such that $E(\eta_{nt}^{p} | w_n, t, y_n, \gamma_{nt}^{v}) = 0$.

This first order condition (**??**) can be solved in y_{nt} to yield the optimal output supply function:

$$y_{nt} = \frac{p_t - \gamma_{nt}^{v_1} v_1(w_{nt}, t; \theta^{v_1})}{\gamma_{nt}^{v_2} v_2(w_{nt}, t; \theta^{v_2}) - \varepsilon \frac{p_t}{Y_t}} + \eta_{nt}^{y}. \tag{49}$$

The random term η_{nt}^{y} is related to η_{nt}^{p} and satisfies

$$E(\eta_{nt}^{y} | w_n, t, y_n, \gamma_{nt}^{v}) = 0. \tag{50}$$

The main difficulty we are confronted with in this section, is that the γ_{nt}^{j} are unobserved and correlated with w_{nt}, y_{nt}, and in our empirical part, we cannot control for it as we do in (47) and (50). This prevents consistent estimation of the parameters of interest when simply ignoring unobserved heterogeneity.

We try to capture unobserved heterogeneity, and follow a proxy variable approach similar to Olley and Pakes (1996) and Levinsohn and Petrin (2003) in context of production functions. For this purpose, we rely on plausible assumptions to identify the values taken by these functions. We begin to note that unobserved γ_{nt}^{j} values capture the relative state of firm n's technology at time t in comparison to a reference technology (denoted by u and v_j) that is identical for all firms and time periods. These relative efficiency levels may depend upon input prices and the production level, on unobserved firm specific effects, time specific affect, lagged efficiency level achieved at $t-1$. As these relative efficiency levels are known to the firm, it will invest more intensively and produce more when both efficiency indicators are good. Like Olley and Pakes (1996) we consider past investment, the age of the firm, and as recommended by Wooldridge (2019) we consider the number of firms' occurrences in the survey, to capture selection effects.[9] Let us gather all these variables into the vector z_{nt}, and consider the following version of a (conditional) strict exogeneity assumption:

[9] See Appendix B, Table B5, for some descriptive statistics for these variables.

Assumption 9. Conditionally to z_{nt} the random terms satisfy:

$$E[\eta_{nt}^c|w_n, t, y_n, z_{nt}] = 0, \tag{51}$$
$$E[\eta_{nt}^y|w_n, t, y_n, z_{nt}] = 0, \tag{52}$$
$$E[\gamma_{nt}^j|w_n, t, y_n, z_{nt}] = \gamma^j(z_{nt}), \qquad j = u, v_1, v_2 \tag{53}$$

The first two conditions of Assumption 9 (A9) correspond to strict exogeneity of the additive random terms conditionally to z_{nt}. The vector z_{nt} includes variables which are correlated with unobserved heterogeneity and uncorrelated with the random terms $\eta_{nt} = (\eta_{nt}^c, \eta_{nt}^p)^\top$. This allows us to replace the condition $E[\eta_{nt}^c|w_n, y_n, \gamma_{nt}] = 0$, which is not useful when the γ_{nt} are unknown, by $E[\eta_{nt}^c|w_n, t, y_n, \gamma_{nt}] = 0$, provided that the instruments z_{nt} are sufficiently comprehensive. Then the conditions in A9 are a good proxy and informative about the data generating process. The flexibility of A9 allows us to nest different models and to test the validity of different specifications or sets of instruments.

Applying A9 to our parametric model, implies that we can replace γ_{nt}^j by $\gamma^j(z_{nt})$ in the expression of the cost and marginal cost function (46) and (48), as well as in the output supply (49).

Several estimation strategies can be followed. With Cobb-Douglas type production functions, a semi-parametric two-stage approach is often adopted (see Ackerberg et al. (2015) for references). The first stage consists in a nonparametric estimation of the technology. In a second stage, the parameters of interest are identified. In contrast, we rely on a full (but quite flexible) parametric specification. In our context, many variables are included in z_{nt} and the curse of dimensionality prevents us from using a nonparametric setup. Another advantage of a parametric specification, is that it is computationally less burdensome in face of a large number of observations. We can also quite easily estimate the system of equations while imposing cross equations identifying restrictions. These advantages may outweight issues related to misspecifications of the functional form for γ. The third advantage of a parametric specification is that it allows for correlated random γ terms when including, as advocated by Wooldridge (2019), firm (and time) specific means into z_{nt}. The last advantage of our parametric approach, and also highlighted by Wooldridge (2009), is that a single estimation step is sufficient to provide most statistics of interest.

For simplicity, we specify the γ^j in A9 as linear functions in the parameters and in the explanatory variables for $j = u, v_1, v_2$:

$$E[\gamma_{nt}^j|w_n, y_n, z_{nt}] = 1 + (z_{nt} - \overline{z})^\top \beta^j, \tag{54}$$

where the constant vector of empirical means $\overline{z}$ is subtracted to ensure that the unconditional expectations satisfy $E[\gamma_{nt}^j] = 1$.

8 Estimation results

8.1 Returns to scale and rate of technological change

This subsection evaluates the rate of Returns to Scale (RTS) defined by

$$\frac{\partial \ln c}{\partial \ln y}(w, t, y), \tag{55}$$

over our observations. When the estimated statistic is lower than one, the observation exhibits increasing RTS, while RTS are constant or decreasing when the statistic is equal to or greater than one. The cost function also comprises a time trend as argument, and allows us to compute estimates for the Rate of Technological Change (RTC):

$$\frac{\partial \ln c}{\partial t}(w, t, y). \tag{56}$$

These statistics depend upon the explanatory variables (both observed and unobserved) and are different for each observation in our sample.

Table 5 summarizes the elasticity of total cost with respect to output, which corresponds to our measure of the rate of return to scale. While the estimated values depend somewhat on the model specification, the broad conclusions are the same over all models: there is evidence for a

variety of rate of returns: about 40% of the observations exhibit increasing returns to scale, 25% have almost constant returns to scale, while about 35 % of the observations have decreasing RTS. Our baseline Model 1 (without correlated unobserved heterogeneity) is already compatible with some heterogeneity in RTS over observations, and the distribution of the rates of RTS are broadly compatible with those obtained for the more general Models 2 to 4.

Table 5: Distribution of firms' returns to scale

	Model 1	Model 2	Model 3	Model 4
P25	0.89	0.86	0.83	0.83
P50	1.04	1.01	0.98	0.98
P75	1.16	1.11	1.09	1.09
MAD	0.13	0.12	0.12	0.12

Note: P25, P50, and P75 report the 25^{th}, the 50^{th}, and the 75^{th} percentile of the respective distribution. MAD denotes the Median Absolute Deviation.

Estimates for increasing returns are quite common for cost functions, and this result contrasts with the estimates usually found with a production function approach which often make a case for decreasing returns to scale. See for instance Diewert and Fox (2008) for a discussion. These contradictory empirical results are often attributed to the endogeneity of the production level in the cost function, which is expected to be correlated with unobserved heterogeneity. As our approach controls for unobserved heterogeneity, we expect no endogeneity bias to occur in our estimates. The finding that increasing RTS do not disappear in Model 4, despite the strong statistical rejection of Model 1 (see Table **??**), supports the hypothesis according to which increasing RTS are not due to endogeneity of output.

The quartiles of the estimates for the RTC are given in Table 6. The median value of the RTC is negative, and indicates that for given values of (w, y), costs tend to decrease over time, by a median value of 0.71% (Model 1) or 0.24% (Model 4). This measure of technological change, however, varies quite importantly over the 4 specifications considered. This quite important difference between Model 1, neglecting unobserved heterogeneity, and Model 4, which is the most flexible specification, is not surprising. Indeed, Model 2-4 allow for correlated technological progress (mediated through changes in γ^u, γ^v), while Model 1 only considers the deterministic and exogenous time trend as source of technological change. Overall, we conclude that about 50% of technological change is endogeneous and reallocates output over firms.

Table 6: Distribution of firms' rate of technological change

	Model 1	Model 2	Model 3	Model 4
P25	-3.29	-1.81	-1.36	-1.77
P50	-0.71	-0.01	-0.08	-0.24
P75	0.67	1.49	0.77	0.89
MAD	1.80	1.66	1.06	1.33

Note: P25, P50, and P75 report the 25^{th}, the 50^{th}, and the 75^{th} percentile of the respective distribution. MAD denotes the Median Absolute Deviation.

For about 40% of the estimated total cost tend to increase over time, this means that many firms have to compensate this positive trend by lower values of (γ^u, γ^v) if they want to keep their cost efficiency unchanged or improved.

8.2 Unobserved heterogeneity

We first provide some insights in the distribution of estimated values of the unobserved fixed and variable cost efficiency, $\widehat{\gamma}^u_{nt}$ and $\widehat{\gamma}^v_{nt}$. Table 7 presents the quartiles of their respective distribution and allows comparing different specifications of unobserved heterogeneity. As already discussed, Model 1 does not take any unobserved heterogeneity into account, which is equivalent to $\widehat{\gamma}^u_{nt} = \widehat{\gamma}^v_{nt} = 1$, for all n, t. Comparing the other models, we find a wide degree of unobserved heterogeneity especially in firms' fixed cost parameter. Considering Panel A, it can be seen that distribution of

$\widehat{\gamma}^u_{nt}$ changes somewhat by increasing the number of z-variables contained the function of $\widehat{\gamma}^u$ and $\widehat{\gamma}^v$. Instead, Panel B shows that the distribution of $\widehat{\gamma}^v$ is much more stable over the different models, and highly concentrated around one, which is also indicated by the small MAD.

Table 7: Distribution of $\widehat{\gamma}^u_{nt}$ and $\widehat{\gamma}^v_{nt}$

	Panel A: Distribution of $\widehat{\gamma}^u$			
	Model 1	Model 2	Model 3	Model 4
P25	1.00	-0.33	-0.31	-0.33
P50	1.00	0.31	0.30	0.26
P75	1.00	1.20	1.07	1.06
MAD	0.00	0.75	0.68	0.68
	Panel B: Distribution of $\widehat{\gamma}^v$			
	Model 1	Model 2	Model 3	Model 4
P25	1.00	0.94	0.90	0.90
P50	1.00	0.98	0.99	0.99
P75	1.00	1.01	1.08	1.07
MAD	0.00	0.04	0.09	0.09

Note: P25, P50, and P75 report the 25^{th}, the 50^{th}, and the 75^{th} percentile of the respective distribution. MAD denotes the Median Absolute Deviation.

The parameters respectively represent fixed and variable cost unobserved heterogeneity, so that we can conclude from these figures that about 60% of all firms operate with virtually zero or very small fixed cost. The other firms have a positive fixed cost, and there is considerable heterogeneity about the size of these fixed cost. The parameter γ^v represents variable cost heterogeneity. We conclude from panel B, that while about 25% of the firms have a variable cost of 10% below average (for which $\gamma^v = 1$), there are also 25% of the firms with average costs higher than average by 7% or more. This unobserved heterogeneity is economically relevant, and strongly influences firms' size, according to Proposition 3.

As the Fisher test (Table **??**) supports the specification of Model 4, we report below only results based on that model. Table 8 summarizes the percentage of estimates corresponding to 4 possible families of cost functions according to the estimated values of v_{1nt} and v_{2nt}. In almost all cases, predicted marginal cost are positive and convex (94.8% of the observations). In 5.2% of the cases, we find evidence for decreasing marginal cost. Such a result is only economically sustainable if firms are able to charge a markup over their marginal cost. Table 8 gives an crude overview of the joint distribution of the v_{1n}, v_{2n} values.

Table 8: Share of observations for different type of heterogeneity in v_{1n}, v_{2n} in %

	$v_{1n} \leq 0$	$v_{1n} > 0$
$v_{2n} \leq 0$	0.0	5.2
$v_{2n} > 0$	0.2	94.6

Note: .

Figure 5 shows kernel density estimates of the distribution of $\widehat{\gamma}^u$ (on the left) and $\widehat{\gamma}^v$ (on the right).[10] Both densities are single peaked, and show that there is a high probability mass around $\gamma^u = 0$ and around $\gamma^v = 1$. For completeness, we also report on Figure 7 and 6 the joint density of $\widehat{\gamma}^u$ and $\widehat{\gamma}^v$ as well as the corresponding contour plot. From this figure, there is a priori no strong dependence between both random variables, and the existence of a relationship like (13) is not supported by our estimates.

[10]The densities are estimated using a second-order Gaussian kernel and likelihood cross-validation to obtain optimal bandwidths.

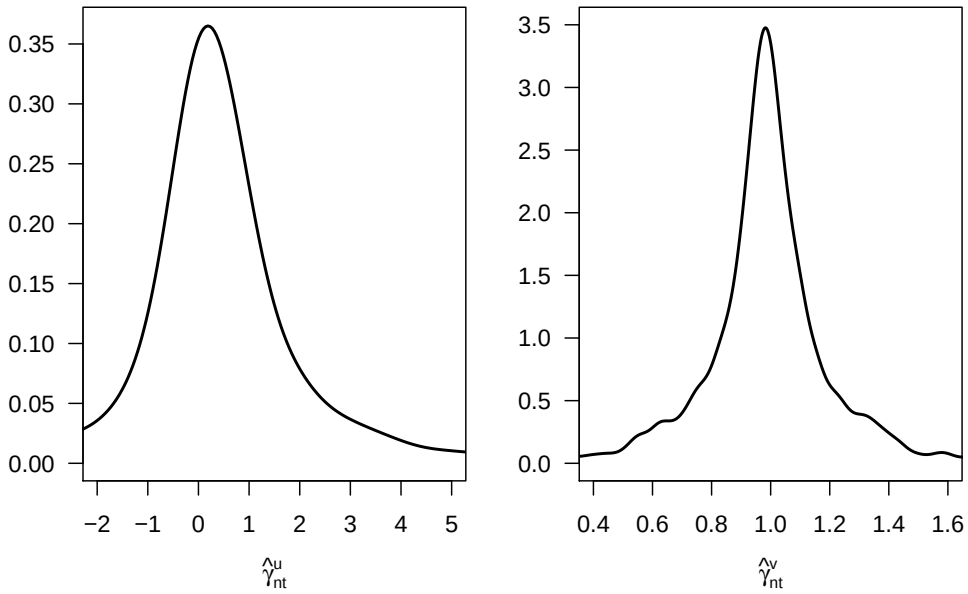

Figure 5: Kernel density estimate of unobserved fixed and variable cost parameters, $f_u(\widehat{\gamma}^u)$ and $f_v(\widehat{\gamma}^v)$

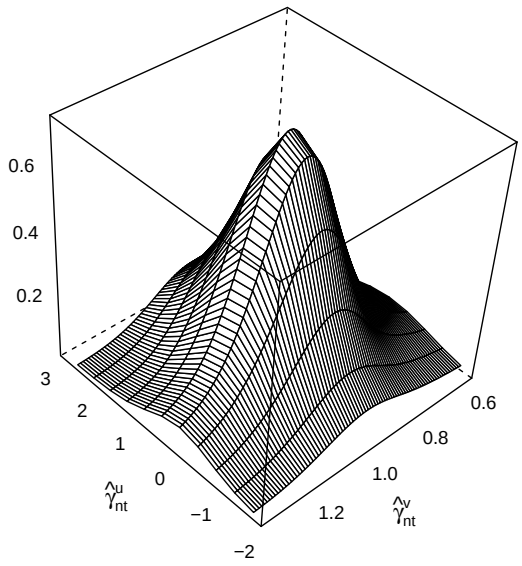

Figure 6: Joint density estimate of unobserved cost parameters $f(\widehat{\gamma}^u, \widehat{\gamma}^v)$

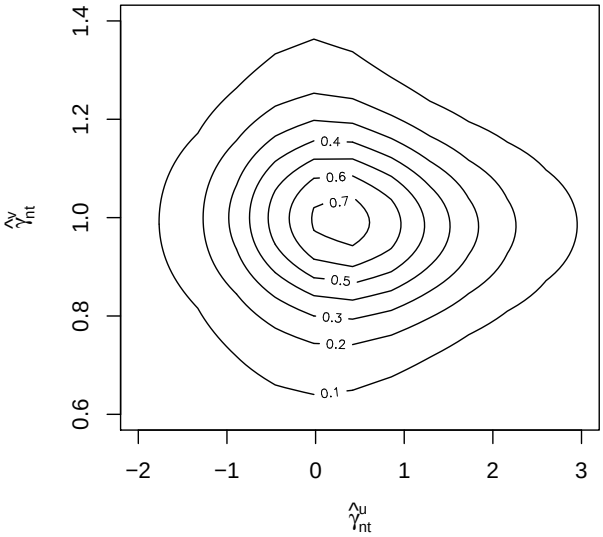

Figure 7: Contour plot of the joint density of unobserved cost parameters $f(\widehat{\gamma}^u, \widehat{\gamma}^v)$

8.2.1 Heterogeneity by firm size and years

One of the main conclusion of the Cournot model is that there is an ordering of unobserved heterogeneity and firm size. We investigate these relationship further and report the above statistics by firm size.

Table 9 completes the information given in Tables 5 and 6 and reports, among other, the median value of fixed and variable cost, together with RTS and RTC over firm size. Surprisingly, we find that the fixed costs represent 80% of total cost for firms with one employee, and this falls to 23% for firms with 2 to 4 employees. This value is below 5% for most sizes, but suddenly increases to 8 % for the biggest firms with 500 employees and more. The median value is γ^u is globally increasing with firm size, while this of γ^v is almost constant, close to 1, but falls to 0.88 for the biggest firms in our sample. This means that these firms are more efficient than average, in conformity with Proposition 3(i). These findings also highlight the shortcomings of usual specifications for cost functions, as the Cobb-Douglas or the translog, which exclude by construction the occurrence of fixed costs.

Table 9 shows that increasing returns are mainly prevalent for big firms in the upper tail of the size distribution. Some small firms also exhibit increasing RTS, in relation with higher than average fixed costs, and the difficulty to achieve a positive profit. For all small and medium size classes the median RTS is close to 1 (constant RTS). For the largest firms, however, we find the strongest median RTS with a value of 0.91, which is related to their market power and conform to Proposition 1(iii). Regarding technological change, the estimated median value of $\partial \ln c / \partial t$ is almost monotonically decreasing with firm size. For the smallest firms, the RTC is very important and represents a cost reduction of 0.72% by year, ceteris paribus. This rate rapidly decreases with firm size (in absolute value), and is close to 0 for the largest firms. This empirical evidence strongly supports Arrow's view about the virtue of competition for innovation, against Schumpeter's argument. (We are aware though that cost reduction is only one aspect of innovation.)

Table 9: Median statistics by firm size[a,b]

Firm size	$\frac{\widehat{\gamma}^u_{nt}\hat{u}}{c_{nt}}$	$\hat{\gamma}^u_{nt}$	$\hat{\gamma}^v_{nt}$	$\frac{\partial \ln c}{\partial \ln y}$	y_{nt}/Y_t	Markup	$\frac{\partial \ln c}{\partial t}$	$cor(c_{nt}, \widehat{c}_{nt})$	$cor(mr_{nt}, \widehat{mc}_{nt})$
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
1	0.80	0.18	0.99	1.00	0.00	1.00	-0.72	0.77	0.52
2-4	0.23	0.17	0.98	0.97	0.01	1.00	-0.53	0.75	0.63
5-9	0.08	0.24	0.98	0.97	0.02	1.00	-0.34	0.86	0.68
10-19	0.04	0.31	0.98	0.98	0.04	1.00	-0.13	0.90	0.69
20-49	0.03	0.41	1.01	0.99	0.10	1.00	-0.06	0.98	0.71
50-99	0.02	0.62	1.02	0.98	0.26	1.00	-0.05	0.94	0.72
100-199	0.02	0.84	1.03	0.97	0.59	1.01	-0.04	0.92	0.66
200-499	0.03	1.25	1.02	0.95	1.35	1.02	-0.08	0.93	0.59
500+	0.08	2.30	0.88	0.91	4.82	1.08	0.03	1.00	0.33
Total	0.08	0.26	0.99	0.98	0.02	1.00	-0.24	1.00	0.63

[a] Firm sizes are measured by the number of employees.

[b] Column (1) reports the share of fixed costs over total costs, (4) reports returns to scale, (5) reports 4-digit market shares, (7) reports the rate of technological progress, (8) reports the correlation between firms' observed costs and the fitted values from the cost regression, and (9) reports the correlation between firms' (computed) marginal revenues and the fitted values of the marginal cost.

The last two columns of Table 9 give an indication of the fit obtained by our model, for both regressions and different firm sizes. While our cost function fits the cost data quite well for all size groups, the marginal cost function is farther away from the marginal revenue function, especially for the smallest and biggest firm sizes.

We also illustrate how some key estimates change over the entire sample period from 1994 to 2016. In particular, Figure 8 shows the evolution of the median of $\widehat{\gamma}^u$, which fluctuates around 0.25 over the period. Instead, the evolution of the median value of $\widehat{\gamma}^v$, depicted on Figure 8, reveals a clearly decreasing pattern in a quite narrow range, from about 1.04 in 1994 to 0.94 in 2016. This implies that, at the median, firms produce with a lower variable cost over time. The decrease is not continuous, however, and γ^v remains almost constant around 0.95 from 2008 onwards.

Figure 10 depicts the evolution of returns to scale over time and illustrate that this value varies little over time and remains close to 1. Regarding technological change, Figure 11 reports the median value of the RTC, i.e. the change in costs wrt time, for constant γ^u, γ^v. For most periods, we estimate a negative median RTC, indicating that firms generally become more cost efficient over time. However, we also see that the median RTC slows down from 2008 forward and stabilizes to a value around 0 in 2012 and after.

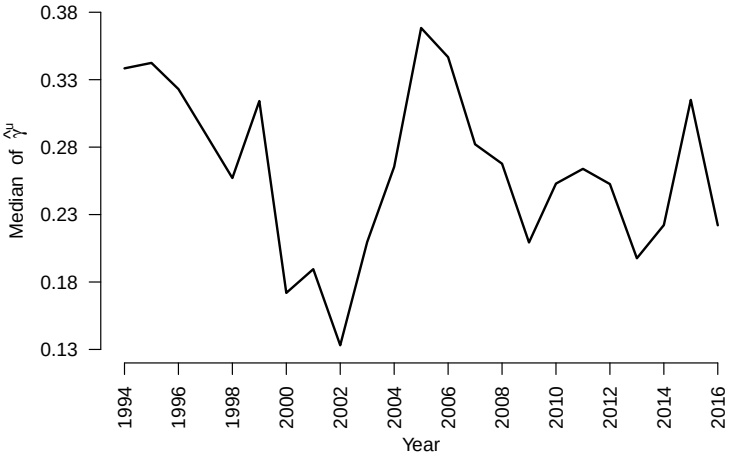

Figure 8: Median evolution of unobserved fixed cost efficiency $\widehat{\gamma}^u$

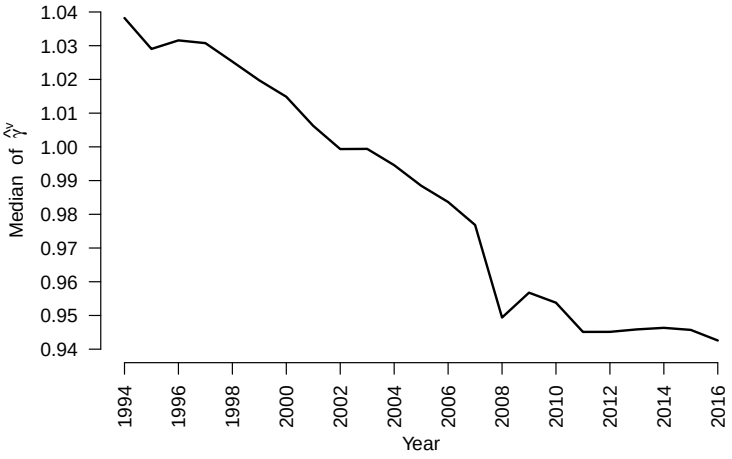

Figure 9: Median evolution of unobserved fixed cost efficiency $\widehat{\gamma}^v$

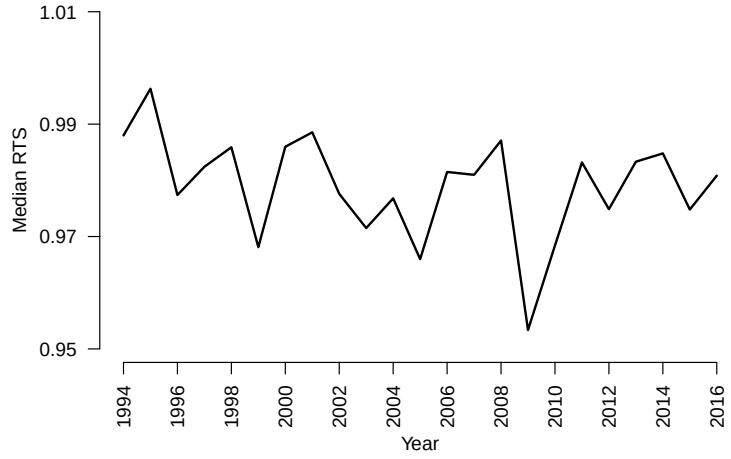

Figure 10: Median value of the rate of returns to scale over time

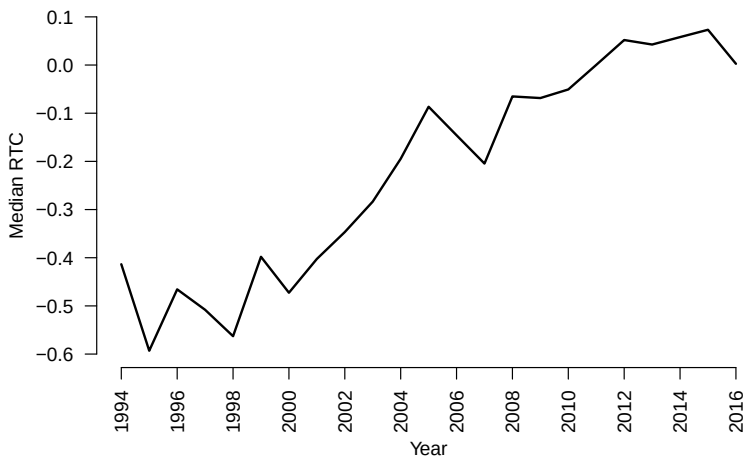

Figure 11: Median value of the rate of technological change over time

8.2.2 Firm size distribution

One of our objectives was to propose a theoretical and empirical model able to cope with unobserved heterogeneity and able to endogenously reproduce the distribution of output levels over firms. Econometric models neglecting unobserved heterogeneity fail in this respect. Additive unobserved heterogeneity in the cost function only will miss the point, because this type of heterogeneity disappears in the marginal cost function. Hence our specification with bivariate joint heterogeneity in fixed and variable (or marginal) cost. We now evaluate our econometric approach by comparing the actual distribution of firms' output levels with the one endogenously predicted by our model. We predict the optimal production level $\widehat{y}_{nt}^{C}$ for each firm (at Cournot equilibrium) using (18), and report the corresponding density on Figure 12. We also consider the Cournot model without unobserved heterogeneity (Model 1) and compute firms' optimal output level $\widehat{y}_{nt}^{C,sym}$ by (18) after setting $\gamma^u = \gamma^v = 1$. It is convenient to represent the density for the logarithm of the output level to avoid having a large support with paucity of observations when output is measured in level.

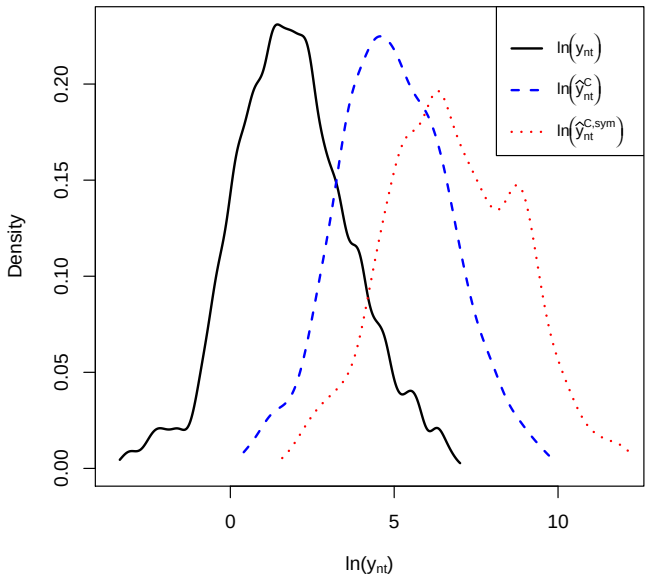

Figure 12: Firms' size distributions, observed and predicted

Figure 12 is informative both about the strengths and shortcomings of our approach. All three

distributions have a similar shape but quite a different support. Our model including unobserved heterogeneity is closer to the observed log-output distribution than the model neglecting it (or considering it as random and uncorrelated with output). The three densities reach a peak at respectively $\ln y = 1$, $\ln y = 5$ and $\ln y = 7$, which represents a sizable gap between observed and predicted production values. The main reason for this discrepancy is that our model targets the cost and the marginal cost function, but not the production level of our firms. Our objective is to extend our model to include this additional criteria into the econometric framework, either by using a moment fitting or simulated maximum likelihood approach. A further reason for the lack of fit between the log-output distributions is that we have included only two unobserved heterogeneity terms, which respectively affect the fixed cost and the first derivative of the cost function. A third unobserved heterogeneity term is actually needed to allow for heterogeneous second order derivatives of the cost function wrt y, and which determines the optimal (and heterogeneous) firm size. Such extensions are part of our future research agenda.

8.3 Production reallocation

Some evidence for productive inefficiency is straightly available from the descriptive statistics. There is a huge heterogeneity in the level of average cost c_{nt}/y_{nt} over firms, which is compatible both with technological heterogeneity and inefficient output allocation over firms. About 10% of the firms have an average cost which is three times the median average cost in manufacturing. Output reallocation spontaneously occurs, but at a very slow pace. Over the 184 four digit industries, the inter-quartile range of $cor(c_{nt}/y_{nt}, y_{nt}/Y_{4t})$ goes from -0.047 to -0.017. Moreover, the average (over firms) average cost, is always much higher than the production share weighted average cost, which means that their is a negative correlation between average cost and market share y_{nt}/Y_{4t}.

Starting from the observed output allocation over firms (see Fig. 12), which we assume to correspond to a long-run Cournot equilibrium, we compute the values of the aggregate output, price, welfare, Hirschman-Herfindahl index, concentration ratio and number of firms. Table 10 reports the median value of these numbers over all 4-digit industries. We then redistribute individual outputs over firms in order to resolve the inefficiency due to market power. This new and regulated optimum correspond to the SROW described in Section 4.1, whose properties are given by P4 and P5. At the SROW all firms producing at SRCE are still active (some firms produce zero output though), so there is still a technological inefficiency due to too many occurrences of inefficient firms. We then allow the social planer to pick up those firms allowed to produce at LROW and exit other firms according to Section 4.2. The numerical results of these simulations are summarized in Table xy.

Table 10: Welfare and output distribution at LRCE, SROW and LROW

	Y	P	C	Π	W	HH	C_{10}	$1-C_{50\%}$	N
LRCE	100	100							
SROW	110	90							
LROW 2	120	80							
LROW 1	130	70							

In order to built a bridge with the homogeneous firm setup of Mankiw and Whinston, we define the firm level average cost at Cournot equilibrium as:

$$c^C \equiv \frac{1}{N^C} \sum_{m=1}^{N^C} c\left(w_m, y_m^C, \gamma_m\right). \tag{57}$$

This allows to recast our heterogeneous firms Cournot equilibrium into a virtual homogeneous firms Cournot equilibrium with the same total number of firms and the same total production level. Using the formulas for the welfare level at Cournot equilibrium and at the optimal point, we decompose the total inefficiency of Cournot competition into three terms. The first term on the RHS of (58) reflects inefficiency due to too low output and too high price at Cournot equilibrium. The second term corresponds to the inefficient big number of firms and the last term is due to the inefficient way of producing at too high costs. This last term is zero if firms are homogeneous.

$$W^S - W^C = \int_{Y^C}^{Y^S} P(s)\,ds + (N^C - N^S)c^C + \left(N^S c^C - \sum_{m \in \mathcal{N}^S} c(w_m, y_m, \gamma_m) \right) \quad (58)$$

Our estimates show that at the welfare maximizing point, output redistribution would improve welfare by about xx%. A quite important part of welfare gain is achieved by increasing output and reducing its price (xy%). A reduction in the number of firms, due to an homogeneous optimization of the production scale represents about xz% of welfare gain. The total cost of production can also be reduced through redistribution of output to more cost efficient firms, and this represents about xx% of the welfare gain.

9 Conclusion

This paper investigates Cournot competition with heterogeneous firms, and highlights the regularities emerging in this context between firm size, market shares, marginal cost and market power. While greater firm size is a good indicator of cost efficiency, it is at the same time an indicator of welfare inefficiency due to market power. Whereas most competition policies limit the overall inefficiency by constraining firms' admissible market shares, we emphasize that an alternative policy could be to promote firms' cost efficiency while limiting their market power.

Our empirical contribution consists to develop an estimation strategy able to identify the distribution of two multiplicative correlated random terms: one affecting the fixed cost and one associated with the variable cost of production. In this context, standard estimation procedures yield inconsistent estimates. We extended a technique available for the estimation of one additive productivity term occurring with a production function to two multiplicative terms affecting the cost function.

Our empirical results highlight the importance of both observed and unobserved heterogeneity for explaining firms' cost and marginal revenues. Fixed costs are often very small, but found to be significant for the smallest and largest firm sizes, which may have policy implications, both for increasing the survival probability of small firms, than for fighting inefficiencies (or market power) of bigger firms. Unobserved heterogeneity in variable costs give a competitive advantage to bigger firms by lowering their variable cost function (ceteris paribus). However, we also estimate that this type of cost efficiency is compensated by lack of technological improvement over time for bigger firms.

One important theoretical result is the generalization of Mankiw and Whinston (1986)'s theorem about excess entry at Cournot equilibrium to the case of heterogeneous firms. It would be interesting in a further paper to evaluate quantitatively the size of the inefficiencies due to too many small firms producing with fixed cost and high variable cost, and to evaluate the welfare gains of redistributing their production to bigger firms producing with lower marginal cost. For this purpose, Proposition 4 would be helpful to characterize the different configurations, and guide us for writing the computer code for redistributing market shares. We could then compute the optimal degree of concentration together with the optimal number of firms active in each market. Before to be able to tackle this issue, however, we have to amend our model and estimation approach towards still more flexibility, so that our models' predictions still improve and catch more stylized facts of the industrial structure.

A Proof of the propositions

Proof of Proposition 1.
(i) and (ii). By A1 it follows that $\epsilon(P,Y) \equiv P'(Y)\,Y/P(Y) < 0$. By A2, at equilibrium $P(Y) + P'(Y)\,y_n^N > 0$ hence $P(Y)\left(1+\epsilon(P,Y)\,y_n^N/Y\right) > 0$. Summing these inequalities over N gives (i). The inequality also implies that individual market shares are bounded above: $y_n^N/Y < -1/\epsilon(P,Y)$.
(iii) From the first order condition $\partial c_n/\partial y = P(Y)(1+\epsilon(P,Y)\,y_n^N/Y)$ it turns out that at Cournot equilibrium

$$y_i^N > y_j^N \Leftrightarrow \frac{\partial c_i}{\partial y}\left(w_i, y_i^N\right) < \frac{\partial c_j}{\partial y}\left(w_j, y_j^N\right).$$

Claim (iv) directly follows from (iii) and the definition of the price markup $P/(\partial c_n/\partial y)$.
Claim (v) corresponds to Okumura (2015, Lemma 1). □

Proof of Proposition 2.
Input prices could be heterogeneous over firms, but without affecting the result, so we use notation w instead of w_n. The Cournot equilibrium is characterized by N individual production levels $y_n^N\left(w, \{\gamma_n^v\}_{n=1}^N\right)$ and $Y^N\left(w, \{\gamma_n^v\}_{n=1}^N\right)$ such that the first and second order optimality conditions are satisfied. We find it convenient to omit the arguments $\left(w, \{\gamma_n^v\}_{n=1}^N\right)$ of Y^N and y_n^N in the equations below. At Cournot equilibrium, individual and aggregate output levels satisfy:

$$\begin{aligned} P\left(Y^N\right) + P'\left(Y^N\right) y_i^N &= \gamma_i^v \frac{\partial v}{\partial y}\left(w, y_i^N\right) \\ Y^N &= \sum_{n=1}^N y_n^N \end{aligned}$$

Differentiating the first order optimality condition with respect to γ_i^v for two different firms, i and n, gives

$$\begin{aligned} \left(P'\left(Y^N\right) + P''\left(Y^N\right) y_i^N\right)\frac{\partial Y^N}{\partial \gamma_i^v} + P'\left(Y^N\right)\frac{\partial y_i^N}{\partial \gamma_i^v} &= \frac{\partial v}{\partial y}\left(w, y_i^N\right) + \gamma_i^v \frac{\partial v^2}{\partial y^2}\left(w, y_i\right)\frac{\partial y_i^N}{\partial \gamma_i^v} \\ \left(P'\left(Y^N\right) + P''\left(Y^N\right) y_n^N\right)\frac{\partial Y^N}{\partial \gamma_i^v} + P'\left(Y^N\right)\frac{\partial y_n^N}{\partial \gamma_i^v} &= \gamma_n^v \frac{\partial v^2}{\partial y^2}\left(w, y_n^N\right)\frac{\partial y_n^N}{\partial \gamma_i^v}. \end{aligned}$$

Let us define

$$a_n^N \equiv \left[P'\left(Y^N\right) - \gamma_n^v \frac{\partial v^2}{\partial y^2}\left(w, y_n^N\right)\right]^{-1},$$

which is negative by A3(ii), and write

$$\begin{aligned} \frac{\partial y_i^N}{\partial \gamma_i^v} &= a_i^N \cdot \left(\frac{\partial v}{\partial y}\left(w, y_i^N\right) - \left(P'\left(Y^N\right) + P''\left(Y^N\right) y_i^N\right)\frac{\partial Y^N}{\partial \gamma_i^v}\right) \\ \frac{\partial y_n^N}{\partial \gamma_i^v} &= -a_n^N \cdot \left(P'\left(Y^N\right) + P''\left(Y^N\right) y_n^N\right)\frac{\partial Y^N}{\partial \gamma_i^v} \end{aligned}$$

If we sum all partial effects $\partial y_n^N/\partial \gamma_i^v$ over all $n = 1$ to N this gives

$$\begin{aligned} \frac{\partial Y^N}{\partial \gamma_i^v} &= -\sum_{n=1}^N a_n^N \cdot \left(\left(P'\left(Y^N\right) + P''\left(Y^N\right) y_n^N\right)\frac{\partial Y^N}{\partial \gamma_i^v}\right) + a_i^N \frac{\partial v}{\partial y}\left(w, y_i^N\right) \\ \Rightarrow \quad \frac{\partial Y^N}{\partial \gamma_i^v} &= \frac{a_i^N}{1 + \sum_{n=1}^N \left(P'\left(Y^N\right) + P''\left(Y^N\right) y_n^N\right) a_n^N}\frac{\partial v}{\partial y}\left(w, y_i^N\right). \end{aligned}$$

Then A1 guarantees that $\partial v/\partial y\left(w, y_i^N\right) \geq 0$, by A3 $a_i^N < 0$, and A4 implies that the denominator is positive, so

$$\frac{\partial Y^N}{\partial \gamma_i^v} \leq 0.$$

Replacing this term in the individual output supply reaction, shows that for $n \neq i$,

$$\frac{\partial y_n^N}{\partial \gamma_i^v} \geq 0$$

so that necessarily

$$\frac{\partial y_i^N}{\partial \gamma_i^v} \leq 0.$$

We also see, that a marginal change in the fixed cost parameter γ_i^u, holding the parameter γ_i^v constant, has not effect on the Nash equilibrium. Claim (v) follows from the definition of the profit function

$$\pi_i^N\left(w, \{\gamma_n^v\}_{n=1}^N\right) = P\left(Y^N\right) y_i^N\left(w, \{\gamma_n^v\}_{n=1}^N\right) - \gamma_i^u u\left(w\right) - \gamma_i^v v\left(w, y_i^N\left(w, \{\gamma_n^v\}_{n=1}^N\right)\right)$$

which is impacted by a change in γ_i^u and γ_i^v as follow

$$\begin{aligned}
\frac{\pi_i^N}{\partial \gamma_i^u}\left(w, \{\gamma_n^v\}_{n=1}^N\right) &= -u\left(w\right) \leq 0 \\
\frac{\pi_i^N}{\partial \gamma_i^v}\left(w, \{\gamma_n^v\}_{n=1}^N\right) &= P\left(Y^N\right)\frac{\partial y_i^N}{\partial \gamma_i^v} + P'\left(Y^N\right) y_i^N \frac{\partial Y^N}{\partial \gamma_i^v} - v_i - \gamma_i^v \frac{\partial v}{\partial y_i}\frac{\partial y_i^N}{\partial \gamma_i^v} \\
&= P'\left(Y^N\right) y_i^N \frac{\partial Y_{-i}^N}{\partial \gamma_i^v} - v_i < 0,
\end{aligned}$$

where the last simplification is obtained by using firm's i first order condition for optimality. Similarly:

$$\frac{\pi_i^N}{\partial \gamma_j^v}\left(w, \{\gamma_j^v\}_{j=1}^N\right) = P'\left(Y^N\right) y_i^N \frac{\partial Y_{-i}^N}{\partial \gamma_j^v} \geq 0.$$

□

Proof of Proposition 3.
(i) As input prices are identical for both firms we skip w from most of our notations and write for instance v_1 instead of $v_1\left(w\right)$. When the cost functions are quadratic, marginal costs are linear, and for $y_i^N < y_j^N$ at Nash equilibrium, we also have

$$\begin{aligned}
&\frac{\partial c_i}{\partial y}\left(w, y_i^N\right) > \frac{\partial c_j}{\partial y}\left(w, y_j^N\right) \\
\Leftrightarrow\quad &\gamma_i^v \cdot \left(v_1 + v_2 y_i^N\right) > \gamma_j^v \cdot \left(v_1 + v_2 y_j^N\right).
\end{aligned} \tag{59}$$

By convexity, $v_2 \geq 0$, we use the fact that $\gamma_i^v > 0, \gamma_j^v > 0$ and $y_j^N > y_i^N$, to conclude that this inequality is equivalent to $\gamma_i^v > \gamma_j^v$.
(ii) We use the fact that for two numbers $a \geq 0$ and b such that $a + b \geq 0$, we also have $a + b/2 \geq 0$. We identify

$$\begin{aligned}
a &\equiv \left(\gamma_i^v - \gamma_j^v\right) v_1 \\
b &\equiv v_2 \cdot \left(\gamma_i^v y_i^N - \gamma_j^v y_j^N\right)
\end{aligned}$$

The term a is nonnegative by (i) and A2 implies that $v_1 \geq 0$. The condition $a + b \geq 0$ corresponds to (59). The implied inequality $a + b/2 \geq 0$ is equivalent to claim (ii).
(iii) For $\gamma_i^v > \gamma_j^v$, and same technological shock η, relationship A7 implies that $\gamma_i^u < \gamma_j^u$ and $u_i\left(w\right) < u_j\left(w\right)$.
(iv) From $\gamma_i^v > \gamma_j^v > 0$ and A7 with $\eta_i = \eta_j$ we have $\gamma_i^u < \gamma_j^u$ and so

$$\frac{\gamma_i^u}{\gamma_i^v} < \frac{\gamma_j^u}{\gamma_j^v} \Leftrightarrow \left(\frac{2\gamma_i^u u}{\gamma_i^v v_2}\right)^{1/2} < \left(\frac{2\gamma_j^u u}{\gamma_j^v v_2}\right)^{1/2}.$$

□

Proof of Proposition 4.
(i) At the LRCE characterized by (3), it turns out that for any active firm,

$$P(Y_{-n}^C + y_n) - \frac{\partial c_n}{\partial y_n}(w_n, y_n) \geq 0. \tag{60}$$

By A1 and A3(ii) this function is decreasing in y_n at the LRCE for any active firm. At SROW, for maximizing W, the social planer chooses $\{y_m\}_{m=1}^M$ in order to satisfy $P\left(\sum_{m=1}^M y_m\right) - \partial c_n/\partial y_n\left(w_n, y_n\right) = 0$ for any active firm, which requires that $\sum_{m=1}^M y_m^S \geq \sum_{m=1}^M y_m^C$. Equivalently, by A1, we have $P\left(Y^S\right) \leq P\left(Y^C\right)$.
(ii) By definition, W^S maximizes welfare by choosing the optimal level of production over all firms active at LRCE, hence $W^S \geq W^C$. It follows directly from (i) and profit maximization, that:

$$\pi_n^S = P(Y^S)y_n^S - c_n\left(w_n, y_n^S\right) < P(Y^C)y_n^S - c_n\left(w_n, y_n^S\right) \leq P(Y^C)y_n^C - c_n\left(w_n, y_n^C\right) = \pi_n^C.$$

(iii)-(v) At the aggregate production level $Y^S \geq Y^C$ the firms' production plans have to satisfy:

$$\frac{\partial c_m}{\partial y_m}(w_m, y_m^S) = \frac{\partial c_n}{\partial y_n}(w_n, y_n^S) = P\left(Y^S\right), \tag{61}$$

for active firms. At the LRCE, firms' marginal costs are related by:

$$\frac{\partial c_n}{\partial y_n}(w_n, y_n^C) = P'(Y^C)\left(y_n^C - y_m^C\right) + \frac{\partial c_m}{\partial y_m}(w_m, y_m^C),$$

so that bigger firms have lower marginal cost at the LRCE (just as in P1). This equation also shows how each firm n has to adjust y_n^C in order to achieve y_n^S satisfying (61). Let us order firms from lower to higher marginal cost, and define "bigger firms" as those having at LRCE a marginal cost lower than $P\left(Y^S\right)$, and "smaller firms" the other group with $\partial c_n/\partial y_n\left(w_n, y_n\right) \geq P\left(Y^S\right)$. Starting from the LRCE, the social planer requires that:

- bigger firms produce more output: $y_n^S > y_n^C$. Bigger firms with lower but increasing marginal costs, increase their production up to the point where (61) is satisfied (A3 ensures that such a point exists). Bigger firms with decreasing marginal cost at y_n^C cannot have globally decreasing marginal cost by A3, so their production can be increased to met (61).
- smaller firms with decreasing marginal cost produce more if this allows to sufficiently decrease their marginal cost and reach $P\left(Y^S\right)$. If this is not possible, they are shut down.
- smaller firms with increasing marginal costs have to produce less and reduce their marginal cost in order to satisfy (61). If this is not possible, they should stop their activity.

(vi) In points (iii)-(v) we have identified either firms which should continue to produce at SROW, or firms which should be shut down. So that $N^C \geq N^S$. □

Proof of Proposition 5. We use the fact that the Hirschman-Herfindahl index of concentration $H\left(Y, \sum_{n=1}^N y_n^2\right)$ is nonincreasing in N and increasing when individual outputs are redistributed from smaller to bigger firms. Under decreasing returns to scale, point P4(v) vanishes, and point (vi) can be sharpen to $N^S \leq N^C$. Let us define $\kappa \equiv Y^S/Y^C \geq 1$ and starting from LRCE, let us scale all individual output levels up to κy_n^C. This leaves the value of Hirschman-Herfindahl index unchanged as

$$H\left(Y^C, \sum_{n=1}^{N^C}\left(y_n^C\right)^2\right) = \sum_{n=1}^{N^C}\left(\frac{y_n^C}{Y^C}\right)^2 = \sum_{n=1}^{N^C}\left(\frac{\kappa y_n^C}{Y^S}\right)^2 = H\left(Y^S, \sum_{n=1}^{N^C}\left(\kappa y_n^C\right)^2\right).$$

Individual firms have now seen their production arbitrarily scaled up by κy_n^C, so that aggregate production is equal to Y^S. However, in order to produce Y^S optimally, such as characterized in P4, the social planer still has to redistribute the individual output levels κy_n^C while keeping the aggregate level fixed at Y^S. We will show that this is achieved by redistributing output from smaller to bigger firms, which increases the value taken by H at SROW. We know that at LRCE

$$\frac{\partial c_n}{\partial y}(w, y_n^C) = P'(Y^C)\left(y_n^C - y_m^C\right) + \frac{\partial c_m}{\partial y}(w, y_m^C)$$

and so $y_n^C \geq y_m^C$ iff $\partial c_n/\partial y(w, y_n^C) \leq \partial c_m/\partial y(w, y_m^C)$ as in P1. By A7, A8 and convexity, using also P3(i), we have for any value of y

$$0 \leq \frac{\partial^2 c_n}{\partial y^2}(w, y_n) = \gamma_n^v v_2(w) < \gamma_m^v v_2(w) = \frac{\partial^2 c_m}{\partial y^2}(w, y_m).$$

This inequality implies that marginal costs increase more strongly in small firms; so that if we inflate all individual outputs by multiplication with $\kappa \geq 1$ then,

$$\frac{\partial c_n}{\partial y}(w, \kappa y_n^C) \leq \frac{\partial c_m}{\partial y}(w, \kappa y_m^C),$$

which means that bigger firms have still lower marginal costs at $\left\{\kappa y_n^C\right\}_{n=1}^M$ than smaller firms. The social planer wants to implement the equality:

$$\frac{\partial c_n}{\partial y}(w, y_n^S) = P(Y^S)$$

which she can achieve from individual production levels $\left\{\kappa y_n^C\right\}_{n=1}^M$, by increasing further the output of the bigger firms (with lowest marginal cost), and decreasing the output of the smaller firms characterized by

$$\frac{\partial c_m}{\partial y}(w_m, \kappa y_n^C) > P(Y^S).$$

This redistribution of constant aggregate output from small to bigger firms increases the value of H achieved at SROW. □

Proof of proposition 6.
(i) Under the above assumptions, W is continuous and the set of values taken by the welfare function over Γ^L is closed and bounded from above, and so it admits a maximum. The maximum of W on Γ is reached on $\Gamma^L \subseteq \Gamma$. The points on the technological frontier satisfy $\gamma^v = e(\gamma^u)$, a function which under A6 is strictly convex. For any (w, y) function W has straight line isoquants in (γ^u, γ^v), and so reaches a unique maximum in (γ^u, γ^v) on the technological set.
(ii) From (i) it follows that at the LROW point, that the planer adopts the same technology γ^L for all active firms, and so all firms produce the same quantity $y = Y/N$. Under this constraint, the welfare function (24) becomes:

$$\mathbf{W}^L(Ny) = \int_0^{Ny} P(s)\, ds - Nc^L(w, y), \tag{62}$$

with c^L defined in (28). Differentiation wrt y and N then yield the first order conditions for a maximum, which states the zero profit condition, and the equality between price and average cost. Together they imply that $c^L(w, y^L)/y^L = \partial c^L/\partial y(w, y^L) = P(Y^L)$, returns to scale are constant locally. (If N is restricted to be an integer, then this condition is approximately valid for small values of y in comparison to Y.)
(iii) Both optimization problems (26) and (25) have the same objective function, but there are fewer constraints in (26), hence $W^L \geq W^S$.
(iv) If the inequality holds, then the Kuhn and Tucker complementary slackness condition implies that $\gamma^u = 0$.
(v) The claim follows because the first and second order conditions to both problems are identical. □

B Further information on the data and descriptive statistics

B.1 Merging of the datasets FICUS and FARE

For the analysis we merge the two fiscal firm-level data sets FICUS and FARE, covering the periods from 1994 to 2007, and 2008 to 2016, respectively. Both in FICUS and FARE firms are classified by a 4-digit sector nomenclature "NAF" (nomenclature d'activité française). However, from 2008 onward, the FARE sectoral nomenclature changed: new sectors appeared (some FICUS sectors

were split), some FICUS sectors disappeared (were merged into a FARE sector). In FICUS, the nomenclature was organized according to "NAF 1", while in FARE the nomenclature is organized according to "NAF 2". In this study we construct a single data set, 1994 - 2016, by extending the sector nomenclature NAF 2 throughout the whole period. That is, we assign the current 4-digit sector nomenclature NAF 2 retrospectively to all firms observed in FICUS. For firms that are observed both in FICUS and FARE or only in FARE their 4-digit sector according to NAF 2 is known. However, for firms that have exited the market before 2008 we do not know to which NAF 2 4-digit sector they would have belonged to if they had continued their activity. To also classify these firms by the NAF 2 4-digit nomenclature we use the following methodology. We first only look at firms that are observed in both data sets FICUS and FARE. From these observations we build a transition matrix where each row represents a 4-digit sector according to NAF 1 and each column represents a 4-digit sector according to NAF 2. Each cell of the transition matrix contains the number of firms transiting from a specific 4-digit sector in FICUS (NAF 1) to the new 4-digit sector in FARE (NAF 2). Table B1 shows an exemplifying transition matrix, where we chose the NAF 1 4-digit sectors 201A - 205C, i.e. the manufacture of wood and products of wood. For instance it can be seen that there are 2060 firms observed that were classified in FICUS in 201A (first row) and in FARE in the sector 1610 (third column), while there are only 46 observations that were classified in 201A and in FICUS in 0220 (first column). From these observed transition frequencies we then calculate the transition probabilities by simply dividing each element of the matrix by the sum of its corresponding row. That is, the NAF 1 - NAF 2 transition probabilities are calculated by

$$p_{IJ} = \frac{\sum_{n \in I,J}^{N_J} \mathbf{1}_{[n \in I \text{ and } n \in J]}}{\sum_{n \in I}^{N_I} \mathbf{1}_{[n \in I]}}, \tag{63}$$

where n is a firm observed in both FICUS and FARE, I and J are specific 4-digit sectors according to NAF 1 and NAF 2, respectively. $\mathbf{1}$ is an index variable equal to 1 if the condition in parenthesis is fulfilled. Table B2 contains the transition probabilities according to the observed transitions Table B1. It can be seen that those 4-digit transitions between FICUS and FARE that were more frequently observed obtain accordingly higher probabilities. In a second step, firms only observed in FICUS belonging to a specific NAF 1 4-digit sector, are assigned to a NAF 2 (at the 4-digit level), by a random draw with transition probabilities given the row of Table B2.

Table B1: FICUS - FARE: Observed transition frequencies

	NAF 2																		
NAF 1	0220	1392	1610	1621	1622	1623	1624	1629	2223	2512	3101	3109	3319	4329	4332	4391	4399	5610	9524	Total
201A	46	0	2060	5	6	22	35	12	0	0	0	7	0	0	25	24	9	5	0	2256
201B	0	0	498	0	0	0	0	0	0	0	0	0	0	17	4	36	24	0	0	579
202Z	0	0	0	108	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	112
203Z	0	7	33	0	15	1880	8	8	41	26	0	41	0	6	1005	386	34	0	0	3490
204Z	0	0	17	0	0	4	857	6	0	0	0	0	35	0	6	0	0	0	0	925
205A	4	16	10	4	0	21	5	1215	0	0	12	317	0	0	87	0	4	10	156	1861
205C	0	0	0	0	0	0	0	86	0	0	0	0	0	0	0	0	0	0	0	86

Table B2: FICUS - FARE: Transitions probabilities

	NAF 2																		
NAF 1	0220	1392	1610	1621	1622	1623	1624	1629	2223	2512	3101	3109	3319	4329	4332	4391	4399	5610	9524	Total
201A	0.02	0.00	0.91	0.00	0.00	0.01	0.02	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.00	0.00	0.00	1.00
201B	0.00	0.00	0.86	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.01	0.06	0.04	0.00	0.00	1.00
202Z	0.00	0.00	0.00	0.96	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.00	0.00	0.00	0.00	1.00
203Z	0.00	0.00	0.01	0.00	0.00	0.54	0.00	0.00	0.01	0.01	0.00	0.01	0.00	0.00	0.29	0.11	0.01	0.00	0.00	1.00
204Z	0.00	0.00	0.02	0.00	0.00	0.00	0.93	0.01	0.00	0.00	0.00	0.00	0.04	0.00	0.01	0.00	0.00	0.00	0.00	1.00
205A	0.00	0.01	0.01	0.00	0.00	0.01	0.00	0.65	0.00	0.00	0.01	0.17	0.00	0.00	0.05	0.00	0.00	0.01	0.08	1.00
205C	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

B.2 Data cleaning

As mentioned in the main text, the industry for food processing (10), the manufacture of tobacco products (12), and the manufacture of coke and refined petroleum products (19) are excluded from the treated sample. Further, we only keep observations reporting values larger than zero in capital stock (tangible assets), number of employees, materials, and production. Table B3 illustrates summary statistics of a typical four-digit industry if no data cleaning at all was made. The table shows that, compared to the case with data cleaning (Table 2), the average number of firms is more than doubled, given by 765. This is mainly induced by the inclusion in Table B3 of industry 10 and to a smaller extend by keeping firms reporting zero and missing values in the number of employees. However, the table also shows that firms with less than 10 (500 or more) employees account for about 6.2% (53.9%), which is very close to the figures presented based on the cleaned sample. Hence, our sample generally matches the main characteristics of the French manufacturing.

Table B3: Average statistics of a typical four-digit manufacturing industry without data cleaning[a]

Firm size[b]	# of firms	Share of firms	Share of employees	Share of production
0	156	20.39	0.04	2.77
1	96	12.55	0.72	0.34
2-4	161	21.05	3.33	1.15
5-9	110	14.38	5.36	1.99
10-19	60	7.84	6.02	3.01
20-49	52	6.80	12.03	7.94
50-99	16	2.09	8.56	6.20
100-199	10	1.31	10.55	8.58
200-499	6	0.78	14.66	13.61
500+	3	0.39	38.71	53.91
NA	95	12.42	0.00	0.48
Total	765	100.00	100.00	100.00

[a] All figures represent averages over all four-digit industries and years (1994-2016). Shares are given in %.

[b] Firm sizes are measured by the number of employees. The group NA represents those firms with missing values in the number of employees.

B.3 Further descriptive statistics

Table B4 shows shares of firms, employees, and production w.r.t. each considered two-digit industry. The table shows that the manufacture of metal products (25) represents the biggest industry in terms of the average number of firms and average employment, representing about 22.4% of all firms and 13.4% of total employment. Instead, the manufacturing for motor vehicles represents the biggest industry in terms of production, accounting for about 14.6% of total production. See also De Monte (2021) for more descriptive statistics using the same data, with a particular attention on firm dynamics (entry and exit).

Table B4: Average statistics by 2-digit manufacturing industry[a]

Industry[b]	# of firms	Share of firms	Share of employees	Share of production
11	1,132	1.79	1.75	4.04
13	2,578	4.07	2.93	1.94
14	3,574	5.65	3.36	1.76
15	966	1.53	1.39	0.84
16	4,767	7.53	2.91	1.94
17	1,236	1.95	3.33	3.32
18	7,566	11.96	3.77	1.90
20	2,068	3.27	7.49	13.52
21	370	0.58	3.62	4.56
22	3,765	5.95	8.40	6.01
23	4,157	6.57	5.50	4.87
24	815	1.29	3.84	5.41
25	14,185	22.42	13.40	9.14
26	2,483	3.92	6.60	4.49
27	1,853	2.93	5.93	5.16
28	4,858	7.68	7.93	6.78
29	1,559	2.46	10.15	14.58
30	558	0.88	5.06	8.30
31	4,780	7.55	2.63	1.44
Total	63,270	100.00	100.00	100.00

[a] All figures are based on the cleaned dataset and represent averages over the period 1994-2016. Shares are given in %.

[b] 11-beverages, 13-textiles, 14-wearing apparel, 15-leather/related products, 16-wood/products of wood and cork, 17-paper/paper products, 18-printing/reproduction of recorded media, 20-chemicals/chemical products, 21-pharmaceutical products/preparations, 22-rubber/plastic products, 23-other non-metallic mineral products, 24-basic metals, 25-fabricated metal products, 26-computer, electronic, and optical products, 27-electrical equipment, 28-machinery and equipment, 29-motor vehicles/(semi-) trailers, 30-other transport equipment, 31-furniture.

Table B5 illustrates the distribution of some variables included in z_{nt} to capture unobserved heterogeneity for the estimation of the cost function (Section 7.2). As in the descriptive statistics section, the table reports averages in a typical 4-digit industry, as well as the distribution of firm sizes over the 1994-2016 period. Beside the average number and the average share of firms, the table reports the share of investing firms, the investment-to-labor ratio, the average firm age as well as the average number of observed periods (denoted by T_n in the main text). Note that firms' investment, i_{nt}, are given by expenditures in intangible assets, reported in the balance sheets, deflated by the corresponding 2-digit investment price index. Unfortunately, firms' investments are not observed for the specific year 2008. We replace the largest part of these missing values by computing $i_{n2008} = K_{n2009} - (1 - \delta_{2008})K_{n2008}$, where K_{nt} represents firms' intangible assets from the balance sheet, deflated by a corresponding 2-digit price index, and δ_t denotes the capital depreciation rate, likewise calculated at the 2-digit level. It can be seen that the share of investing firms is increasing in firm size, where the share of investing firms with only one employee is given by 57.6 %, whereas almost all firms with 500 and more employees report investments in capital (99.1 %). Regarding the investment-to-labor ratio there seems to be two clusters: one with an investment level of about 6000€ (or 0.06) per worker and another cluster with average investment around 10000€. Considering firms' average age and average number of observed periods, it can be seen that, as expected, both variables are increasing in firm size. That is, while the average age (number of observed periods) of firms with only one employee is given by 12.4 years (5 periods), the largest size group, firms reporting 500 and more employees, are on average 29.1 years old (and observed on average for 12.7 periods). Firms' age, a_{nt}, is calculated as the difference between the current year and the date of creation of the firm. So, firms' age does not necessarily correspond to the number of observed periods as especially small firms often show temporal inactivity and/or drop out of the sample because of missing values. Both variables should represent good proxies to

capture unobserved heterogeneity.

Table B5: Further average statistics by 4-digit manufacturing industry[a]

Firm size[b]	# of firms	Share of firms	Share of investing firms	Investment-to-labor ratio	Firm age	# of obs. periods
1	50	14.71	57.63	0.11	12.37	5.04
2-4	82	24.12	68.63	0.07	13.83	7.48
5-9	73	21.47	81.95	0.06	16.79	9.51
10-19	52	15.29	90.91	0.06	19.73	10.91
20-49	49	14.41	95.42	0.06	22.98	11.56
50-99	16	4.71	97.35	0.06	25.83	11.96
100-199	9	2.65	98.14	0.08	27.14	12.29
200-499	6	1.76	98.83	0.10	27.65	12.83
500+	3	0.88	99.14	0.12	29.13	12.68
Total	340	100.00	80.07	0.07	17.77	9.15

[a] All figures are based on the cleaned dataset and represent averages over the period 1994-2016. Shares are given in %.
[b] Firm size is measured by the number of employees.

References

Acemoglu, D. and Jensen, M. (2013). Aggregate comparative statics, *Games and Economic Behavior* **81**: 27–49.

Ackerberg, D. A., Caves, K. and Frazer, G. (2015). Identification properties of recent production function estimators, *Econometrica* **83**: 2411–2451.

Amir, R. (1996). Cournot oligopoly and the theory of supermodular games, *Games and Economic Behavior* **15**: 132–148.

Amir, R., De Castro, L. and Koutsougeras, L. (2014). Free entry versus socially optimal entry, *Journal of Economic Theory* **154**: 112–125.

Amir, R. and Lambson, V. E. (2000). On the effects of entry in Cournot markets, *The Review of Economic Studies* **67**: 235–254.

Bergstrom, T. C. and Varian, H. R. (1985). When are Nash equilibria independent of the distribution of agents' characteristics?, *The Review of Economic Studies* **52**: 715–718.

Berry, S. T. (1992). Estimation of a model of entry in the airline industry, *Econometrica* **60**: 889–917.

Bresnahan, T. F. and Reiss, P. C. (1991). Entry and competition in concentrated markets, *Journal of Political Economy* **99**: 977–1009.

Cameron, A. C., Gelbach, J. B. and Miller, D. L. (2011). Robust inference with multiway clustering, *Journal of Business & Economic Statistics* **29**: 238–249.

Cameron, A. C. and Miller, D. L. (2015). A practitioner's guide to cluster-robust inference, *Journal of Human Resources* **50**: 317–372.

Chen, X. and Koebel, B. M. (2017). Fixed cost, variable cost, markups and returns to scale, *Annals of Economics and Statistics* **127**: 61–94.

Davis, P. (2006). Estimation of quantity games in the presence of indivisibilities and heterogeneous firms, *Journal of Econometrics* **134**: 187–214.

De Monte, E. (2021). Productivity, markups, entry, and exit: Evidence from French manufacturing firms from 1994-2016, *BETA Working Paper, Université de Strasbourg* .

Diewert, W. E. and Fox, K. J. (2008). On the estimation of returns to scale, technical progress and monopolistic markups, *Journal of Econometrics* **145**: 174–193.

Diewert, W. E. and Wales, T. J. (1988). A normalized quadratic semiflexible functional form, *Journal of Econometrics* **37**: 327–342.

Ericson, R. and Pakes, A. (1995). Markov-perfect industry dynamics: A framework for empirical work, *The Review of Economic Studies* **62**: 53–82.

Esponda, I. and Pouzo, D. (2019). The industry supply function and the long-run competitive equilibrium with heterogeneous firms, *Journal of Economic Theory* **184**: 104946.

Février, P. and Linnemer, L. (2004). Idiosyncratic shocks in an asymmetric Cournot oligopoly, *International Journal of Industrial Organization* **22**: 835–848.

Garicano, L., Lelarge, C. and Van Reenen, J. (2016). Firm size distortions and the productivity distribution: Evidence from France, *American Economic Review* **106**: 3439–79.

Gaudet, G. and Salant, S. W. (1991). Uniqueness of Cournot equilibrium: new results from old methods, *The Review of Economic Studies* **58**: 399–404.

Götz, G. (2005). Market size, technology choice, and the existence of free-entry Cournot equilibrium, *Journal of Institutional and Theoretical Economics* **161**: 503–521.

Gouriéroux, C. and Peaucelle, I. (1990). Hétérogénéité I. Étude des biais d'estimation dans le cas linéaire, *Annales d'Économie et de Statistique* **17**: 163–183.

Guesnerie, R. and Laffont, J.-J. (1978). Taxing price makers, *Journal of Economic Theory* **19**: 423–455.

Hall, R. E. and Jorgenson, D. W. (1967). Tax policy and investment behavior, *The American Economic Review* **57**: 391–414.

Hopenhayn, H. A. (1992). Entry, exit, and firm dynamics in long run equilibrium, *Econometrica* **60**: 1127–1150.

Jovanovic, B. (1982). Selection and the evolution of industry, *Econometrica* **50**: 649–670.

Koebel, B. and Laisney, F. (2014). Aggregation with Cournot competition: The Le Chatelier Samuelson principle, *Annals of Economics and Statistics* **115/116**: 343–360.

Koebel, B. and Laisney, F. (2016). Aggregation with Cournot competition: An empirical investigation, *Annals of Economics and Statistics* **121–122**: 91–119.

Ledezma, I. (2021). Product-market integration with endogenous firm heterogeneity, *Oxford Economic Papers* .

Levinsohn, J. and Petrin, A. (2003). Estimating production functions using inputs to control for unobservables, *Review of Economic Studies* **70**: 317–341.

Lopez-Cuñat, J. M. et al. (1999). One-stage and two-stage entry Cournot equilibria, *Investigaciones Economicas* **23**: 115–28.

Mankiw, N. G. and Whinston, M. D. (1986). Free entry and social inefficiency, *The RAND Journal of Economics* **17**: 48–58.

Martin, R. S. (2017). Estimation of average marginal effects in multiplicative unobserved effects panel models, *Economics Letters* **160**: 16–19.

Melitz, M. J. (2003). The impact of trade on intra-industry reallocations and aggregate industry productivity, *Econometrica* **71**: 1695–1725.

Novshek, W. (1984). Finding all n-firm Cournot equilibria, *International Economic Review* **25**: 61–70.

Novshek, W. (1985). On the existence of Cournot equilibrium, *The Review of Economic Studies* **52**: 85–98.

Okumura, Y. (2015). Existence of free entry equilibrium in aggregative games with asymmetric agents, *Economics Letters* **127**: 14–16.

Olley, G. S. and Pakes, A. (1996). The dynamics of productivity in the telecommunications equipment industry, *Econometrica* **64**: 1263–1297.

Panzar, J. C. and Willig, R. D. (1978). On the comparative statics of a competitive industry with inframarginal firms, *The American Economic Review* **68**: 474–478.

Salant, S. W. and Shaffer, G. (1999). Unequal treatment of identical agents in Cournot equilibrium, *American Economic Review* **89**: 585–604.

Spulber, D. F. (1995). Bertrand competition when rivals' costs are unknown, *The Journal of Industrial Economics* **43**: 1–11.

Wooldridge, J. M. (2009). On estimating firm-level production functions using proxy variables to control for unobservables, *Economics Letters* **104**: 112–114.

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data*, MIT press.

Wooldridge, J. M. (2019). Correlated random effects models with unbalanced panels, *Journal of Econometrics* **211**: 137–150.