Cournot competition with heterogenous firms, welfare and misallocation*

Enrico De Monte† and Bertrand Koebel‡

October 14, 2024

Abstract

This paper characterizes the short- and long-run Cournot equilibrium with heterogeneous firms and stochastic technological change. In our model, firms have different technologies with heterogeneous fixed and variable costs and various degrees of markups. In a framework with homogeneous firms, Mankiw and Whinston (1986) show 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. Using administrative data for French manufacturing firms, we estimate a wide degree of unobserved heterogeneity in both fixed and variable costs. Our simulation results show that markups surprisingly only induce slight inefficiencies in the allocation of output, implying that it is almost compatible with welfare maximisation. Instead, firms' choice to employ heterogeneous and often inefficient technologies turns out to harm substantially welfare and aggregate output.

Keywords: cost function, fixed cost, marginal cost, returns to scale, technological change, misallocation, markups, nonlinear least squares, panel data.

JEL Classification: C33; L11; L60.

*We warmly thank Rabah Amir for continuous and constructive discussions during the process of writing the paper. We also benefited from thorough comments by Flora Bellone, Ivan Ledezma and Bettina Peters. We are also indebted to Ludwig von Auer, Claude d'Aspremont, Rodolphe Dos Santos Ferreira, Serge Garcia, Franọis Laisney, Gauthier Lanot, Anne-Laure Levet, Phu Nguyen-Van, Yasunori Okumura, Hanitra Rakotoarison, Kevin Remmy, Patrick Rey, and the participants of seminars at Barcelona (ESEM 2023), Cergy-Pontoise, Konstanz, Mannheim (ZEW and MaCCI), Nancy, Paris, Strasbourg (PET 2019), Trier, Umeå and Vienna (EARIE 2022) for their helpful comments. Access to confidential data, on which this work is based, has been made possible within a secure environment offered by CASD – Centre d'accès sécurisé aux données (Ref. 10.34724/CASD). We warmly thank Nathalie Picard, FCBA, and ZEW for supporting the access to the data.

†ZEW, Leibniz Centre for European Economic Research, L 7, 1, 68161 Mannheim (Germany). Email: Enrico.DeMonte@zew.de.

‡Université de Strasbourg, CNRS and BETA. Email: koebel@unistra.fr.

1 Introduction

Firms' cost efficiency importantly affects welfare and the standard of living. If firms increase their efficiency in the production process, they are able to produce more without incurring higher costs. In a competitive environment, this decreases output prices, and, consequently, increases consumer welfare. If firms have market power, they usually lack incentives to convert the increase in efficiency into an increase in output and a decrease in prices. This hampers the growth of more efficient firms and leads to a (socially) inefficient allocation of production, and negatively affects consumer welfare. See for instance Berry et al. (2019) and Syverson (2019) for literature reviews and detailed discussions on market power and macroeconomic implications.

To measure firms' cost efficiency and the effect of misallocation implied by market power on welfare, most studies, however, employ cost functions with restrictive functional forms by neglecting fixed costs and unobserved heterogeneity.[1] This has important consequences: such specifications are not able to yield plausible (optimal) output levels and are not suited to investigate issues related to the size distribution of firms and its determinants, and produce biased results and inference. A sound analysis of efficiency and welfare at the firm-level, therefore, requires cost functions with multiple dimensions in unobserved heterogeneity. Moreover, this specification cannot be simply additive, since heterogeneity in the fixed costs vanishes in the derivation of the profit-maximizing condition and is useless for generating 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.

In this paper we propose a novel framework allowing for joint heterogeneities in fixed and variable costs, embedded in the Cournot competition model, where heterogeneous firms interact strategically, choosing their optimal output level given aggregate output, cost and demand parameters. The model not only allows us to investigate the interplay between fixed and variable costs, and firm size, but also to disentangle the effects of technological efficiency, market power, and allocative inefficiency on welfare.

The contribution of the paper to the literature is threefold. First, we empirically implement the theoretical results related to the existence and unicity of the Cournot equilibrium. More precisely, while this literature 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 the 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. 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 theoretical contribution simulating the welfare implications of operating heterogeneous firms at Cournot equilibrium.

Second, using administrative French firm-level data, we contribute to the existing empirical literature by explicitly introducing joint heterogeneity in the fixed cost and in the variable cost of production and study the interplay between both types of heterogeneity. A large part of the literature tackling the issue of productivity and technological change bases its identification strategy on the production function (Ericson and Pakes, 1995). Considering a production function is helpful to estimate productivity, but is

[1] Most studies in the empirical literature rely on efficiency measures derived from production functions (such as translog or Cobb-Douglas) whose corresponding cost functions neglect fixed costs.

not suitable to identify fixed, variable and average costs. The empirical literature on cost functions 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 multidimensional heterogeneity in both the fixed and the variable cost functions. For that purpose, we propose an appropriate identification and estimation strategy of the Cournot model, composed of the inverse demand function addressed to an industry and (nonlinear) cost functions with multidimensional unobserved heterogeneity. More specifically, we estimate in a first step the inverse demand function, applying instrumental variable and fixed effects methods to deal with the simultaneity bias. In a second step, we use the obtained demand parameters to estimate in a nonlinear system-equation approach the firm-level cost and output supply functions. Here, we have to deal with the incidental parameter problem occurring when taking into account unobserved heterogeneity in fixed and variable costs over firms and across time: the number of free parameters to be estimated increases with the number of observations, leading to inconsistent estimates when not appropriately handled. A further factor causing inconsistent estimates if not taken into account arises when heterogeneity is unobserved and neglected while being correlated with firms' decision variables, i.e. the optimal level of output.[2] To solve these problems, we employ a control-function approach in combination with nonlinear least squares allowing us to consistently estimate the cost function parameters and to uncover the distribution of unobserved heterogeneity in fixed and variable costs. Our empirical results confirm the theoretical underpinnings, showing a negative relation between variable and fixed costs: a large share of small firms does not incur any fixed costs, but this share is significantly decreasing in firm size.

Third, we contribute to literature on the measurement of welfare and misallocation. 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 to which extent redistributing output over firms allows an increase in both welfare – by improving allocative efficiency and reducing total costs – and aggregate industry output. Starting from a long-run Cournot equilibrium, we perform simulations to evaluate the welfare loss due to markups, output misallocation and technological inefficiencies.

Measuring misallocation has particularly gained attention in the literature (Hopenhayn, 2014). One reason for this development is the increasing availability of detailed micro data. Baily et al. (1992), for instance, use data from US manufacturing establishments between 1972 and 1988, showing that reallocation from less to more efficient production units accounts for half of aggregate productivity growth. Restuccia and Rogerson (2008) build a general equilibrium model and illustrate that idiosyncratic shocks to producers' decisions importantly affect reallocation of resources and by that total output and productivity. Hsieh and Klenow (2009) find that if production inputs in China and India were allocated as efficiently as in the US, aggregate productivity would increase by 30%–50% and 40%–60%, respectively. Markups, i.e. a firm's ability to open a gap between output price and marginal costs, are considered as an important source of market imperfections, and misallocation. For example, Peters (2020) introduces a Bertrand competition framework, where firms increase markups during the life-cycle of their product(s) by consistently investing in productivity growth. The author then shows that a higher churn intensity - the rate by which new entering firms replace the products of older firms relative to the rate at which firms increase their market power - compresses the markup distribution and reduces the degree of misallocation. Using US data covering the period 1997–2015, Baqaee and Farhi (2020) show in a general equilibrium approach that reallocation from low- to high-markup firms accounts for about 50%

[2] In the production function literature, this is also known as the "transmission bias", see Gandhi et al. (2020) and the cited literature therein.

of aggregate productivity growth (since these firms are also highly efficient). However, the authors also demonstrate that removing firms' markups would further increase aggregate productivity by 15%. Using US manufacturing data, Edmond et al. (2023), find a sizable but much lower effect of firms' markups and implied missallocation on aggregate productivity and welfare. Our paper shares the purpose of that literature but contrasts with its result: in France, the sole removal of price markups has had a hardly visible impact on aggregate output and price. As our simulation shows, the main impact on welfare is obtained by closing firms with negative profit and reallocating their production to more efficient firms.[3]

For the empirical analysis, we use French fiscal firm-level data covering the period from 1994 to 2019 (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 (Ceci-Renaud and Chevalier, 2010).[4] 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.

The reminder of the paper is organized as follows. 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. Sections 8 and 9 discuss the estimation and the simulation results. Section 10 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

[3] See also De Monte (2024), who studies the joint evolution of aggregate productivity and markups and the role of reallocation using similar data on French manufacturing firms.

[4] Various studies showed that size-dependent regulations in France distort labor allocation and so the employment-based firm-size distribution (see, for instance, Garicano et al. (2016); Gourio and Roys (2014)). Our paper distinguishes itself from that literature by aiming to quantify technological differences in fixed and variable costs and how this relates to the output-based firm size distribution and welfare.

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 differences 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 (which 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 $n = 1, ..., N$.

A1 and A2 are common in microeconomics and industrial economics. Assumption A3(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(i) 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. The Cournot equilibrium exists under relatively mild conditions, and we follow Novshek (1985) who showed its 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 the Cournot equilibrium. Amir (1996, Corollary 2.2) used another condition implying the existence of the 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 A3(*ii*) and A4 guarantee that the backward reaction functions are nonincreasing in Y. Given existence, we then characterize the 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.[5] We have the following interesting 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 n's 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 $y_n^{N'} > y_n^N$ for a firm n active at both Nash equilibria.

P1 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 P1 that if we order firms by size (say from the smallest to the biggest), this implies that the same order carries over to the markup and the reverse ordering applies to the marginal cost. P1(*v*) corresponds to what Mankiw and Whinston (1986) refer to as business-stealing: new entries contribute to increase total output but reduce the 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) 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

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

profits.

We rewrite the cost function in order to highlight two key unobserved 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 parameters can 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. *The technological parameters $\gamma_n = (\gamma_n^u, \gamma_n^v)$ are*
(i) stochastic and exogenous to the firm,
(ii) known by firms prior to producing and competing à la Cournot.

A5 ensures that the heterogeneity terms are not a deterministic function of the same explanatory variables as the cost function, and 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 was 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), $\gamma_n^u = \gamma^u$ (Hopenhayn, 1992), γ_n^v *iid* (Jovanovic, 1982), or γ_n^v is independent of γ_n^u (Bresnahan and Reiss, 1991).

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 f is linearly homogeneous in x (which is equivalent to v being linearly homogeneous in y), then $\gamma_n^v = 1/\omega_n$. 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 characterizing different types of firms. 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 costs. 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 $E[\gamma^u] = E[\gamma^v] = 1$.

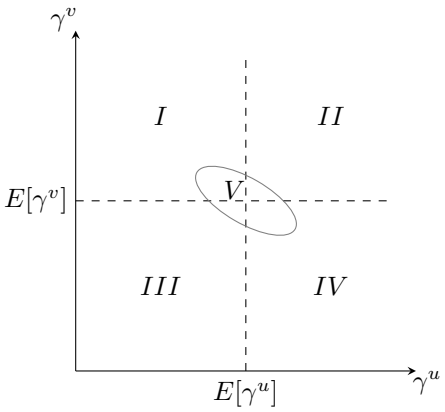

Figure 1: Five technological zones

In the different zones depicted on Figure 1, firms are not only different with respect to their technology, but we also expect to see differences 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's profit decreases with γ_i^v and γ_i^u.
(vi) Firm i's profit increases with γ_j^v.

This result, proven (for completeness) in Appendix A, follows 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) obtain 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 not straightforward 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 (P2(*ii*)) that despite the output levels of all competing firms decreasing after a favorable productivity shock on firm i, the aggregate Cournot output is increasing, too (P2(*iii*)). This means that cost-reducing technological change hurts firms that are not affected by it, they lose market share, 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.

From microeconomic theory, however, we know that the fixed cost is non-decreasing and the variable cost is non-increasing in the level of fixed inputs – see for instance Varian (1992, Chapter 5.1). When the level of fixed input(s) is unobserved, because only information on firms' total capital stock and total labor demand is available, this induces a negative correlation between the fixed and variable cost.

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. A6 implies that, on average, there is a trade-off between technological parameters γ_n^u and γ_n^v, 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 require 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, 1987). 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). The specification given by (2) and (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 multi-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 $\gamma_{2n}^v > 0$, 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 us to obtain an explicit solution for the 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) denotes the backward reaction mapping shown by Novshek (1985). It illustrates that *ceteris paribus*, the higher the variable cost the lower the production level y_n^N (see (P2(*iii*)) 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, such 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, in the sense that:*

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

A8 reduces the dimension of 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, creating a relationship between firms' production levels and their technological characteristics. If we order firms into ascending output levels, 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 only perfect if we control the random term η. With randomness, the order is preserved on average.

The aggregate production Y^N implicitly defined in (19) also depends upon the number N of active firms, and 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 compare the simultaneous game with the one where entry and production choices are sequential. Active firms incur a fixed cost $c_n\left(w_n, 0^+, \gamma_n\right) = u_n\left(w_n\right)$ and inactive firms have $c_n\left(w_n, 0, \gamma_n\right) = 0$.

Active firms expect nonnegative profits and all potential entrants expect nonpositive profits. We introduce the superscript C to characterize the 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 on 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 information sets condition the expectations in (22) and (23) and characterize different LRCE. The distribution of the technological shocks is 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 challenging with this approach to endogenously generate a distribution of firms' sizes close to those usually observed in a given industry.[6]

[6] Even in a setup with homogeneous firms, the Cournot equilibrium can be asymmetric, see for instance Novshek (1984). The corresponding distribution of firm sizes is still very restrictive, however.

4 Welfare and the optimal distribution of production

We now consider the welfare implications of the observed distribution of output, and investigate, following Mankiw and Whinston (1986), the welfare loss at the LRCE. In a setup with identical firms, Mankiw and Whinston have shown that under business stealing (see P1(*v*)), 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 planner controls either entry (but not production) or entry and production. In our situation with heterogeneous firms, the central planner 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 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 contributors to economic activity in W.

4.1 Short-run optimal distribution of production

In the short-run, the planner 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, but a firm can be inactive and produce nothing. 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 are 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 ($y_n^S > 0 \Rightarrow \lambda_n^S = 0$) and (ii) sets $y_m = 0$ for any firm with a marginal cost above the price.

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 greater than or equal to their marginal cost. The next result characterizes the SROW and extends 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 $P\left(Y^C\right) > P\left(Y^S\right)$.
(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: $y_n^C > y_n^S$
(v) Small firms with increasing returns which either produce too little, or should produce nothing.
(vi) A subset of the firms active at LRCE is still producing a positive quantity 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 the SROW. It also defines a big firm as a firm with a level of production at the LRCE such that its marginal cost of production is too low for welfare maximization:

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

and conversely for a small firm. This result is also useful for our empirical purpose of investigating the efficiency of the LRCE (see Section 9). We use P4 to implement the algorithm to compute the SROW and the corresponding reallocation of output over firms at the SROW. Contrary to Mankiw and Whinston (1986), increasing the efficiency of the equilibrium affects firms differently. According to P4(*iii*) and P4(*iv*), it is optimal to reduce the size of smaller firms (with decreasing returns) and increase the size of bigger firms.

Instead of centralizing all production decisions, the central planner 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:

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

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

An interesting consequence of P4 is:

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 the LRCE.*

P5 implies 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, $P\left(Y^N\right) \geq P\left(Y^S\right)$, by P4(*i*) and incites small and inefficient firms to enter the market, while for welfare maximization the planner prefers to increase the production of the technologically more efficient firms. Those big firms, however, do not spontaneously increase their production because they are aware that in order to sell it, the firms have to accept a decrease in price and profits. The proof of P5 is provided in Appendix A, and is both a consequence of the properties of the Hirschman-Herfindahl index, and 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

In the long-run, the planner also has an entrepreneurial duty and selects the production technologies that will be active at Long-Run Optimal Welfare (LROW). The planner 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 the 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 planner 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 planner 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 available. The long-run optimal value satisfies $W^L \geq W^S$, because the planner faces fewer restrictions in (26) in comparison to (25). Solving this problem numerically, by evaluating W over all discrete elements of Γ, is time intensive: 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 is 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:

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

This subset $\Gamma^L \subseteq \Gamma$ can be computed rapidly. At the LROW, the planner can freely choose the technology in order to maximize welfare, so she considers the lower envelope cost function which corresponds to the technological long-run:

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

The cost function c^L is now homogeneous over all firms (and is for instance considered by Mankiw and Whinston (1986)). The long-run technological parameters γ^L are optimal (and vary with w, y in general). In the long-run, the following claims are satisfied:

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 maximizing the central planner problem W^L or decentralized profits wrt (y_n, γ_n), for a given price level which clears the product market with free entry.

By P6(*ii*), at LROW, all firms produce at the minimum of the average cost, which characterizes local CRTS. It is not surprising, given that the planer maximizes welfare with less technological constraints at LROW than in the short-run, that $W^L \geq W^S$. Less common is condition P6(*iv*) which is compatible with the use in the long-run of a technology with positive fixed costs. For a small level of y^L, however, the threshold $u(w)/v(w, y^L)$ in P6(*iv*) can be big, and the planner can choose a technology with no fixed cost, in which case $\gamma^{vL} = e(0)$. When all firms produce the same amount, the Hirschman-Herfindahl index of concentration is $1/N^L$.[7]

[7] If we consider the integer constraint, then further technologies could be used at LROW in order to produce the residual

P6 connects the literature on heterogeneous and homogeneous technologies: at LROW the optimal technological choice is unique, all active firms use the same technology. Under the above assumptions, the distribution of firm sizes degenerates to a mass point at y^L. This degenerate distribution of output is far from the observed density of output, and observed heterogeneity alone is only able to explain a narrow part of the departure between observed and optimal distribution of output. Imperfect competition and unobserved heterogeneity also contribute to explaining this gap, and we will investigate it empirically.

It is not possible to conclude that at LROW $c^L/y^L \leq c_n^S/y_n^S$, because lower average costs are achieved at the price of a higher fixed cost, which is not necessarily efficient at LROW. It is neither true that $Y^L \geq Y^S$, nor that $N^L \leq N^S$ are necessarily satisfied. Regarding the total number of firms active at LROW, the planner closes all firms producing nothing at SROW (and avoids 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 the function describing the relation between variable and fixed cost efficiency, $e(\gamma^u)$, which is an empirical issue.

5 Data and descriptive statistics

We use French fiscal data available at the firm level covering the years 1994 to 2019 (FICUS and FARE data).[8] The data comprises the universe of active firms, but we consider only those belonging to the manufacturing industry.[9] 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 of excluding observations with missing, zero or negative values for sales, labor cost, material cost, and capital cost. We consider all firms with at least one employee, and trim the distribution of profit rates, keeping only observations within the 1% and 99% quantiles. This leaves us with 1,503,299 observations and 172,057 firms. The panel is unbalanced, and on average a given firm is observed for 8.9 years.

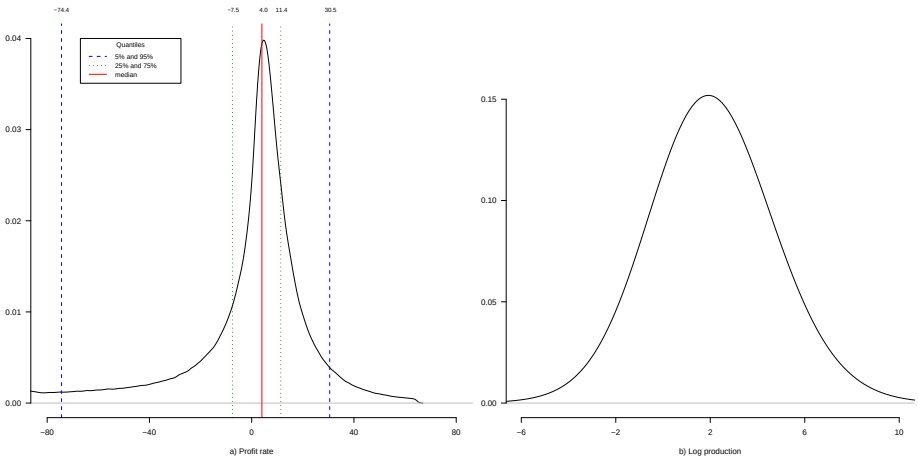

Figure 2: The density of firms' profit rate and log-levels of production

output.

[8] 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, which, in turn, belongs to the current database ESANE.

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

Firm-level 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 in Figure 2 and illustrates that the data are fundamentally heterogeneous. The density of the profit rates (left figure) mimics the distribution of the average cost since by definition $py/c = p/(c/y)$. The density is not normal, but asymmetric, and exhibits a large tail for values below the median (of 4.0%) and a thin tail above. The density of the log-levels of production is illustrated on the right of Figure 2 (over all firms and years).

Some evidence for productive inefficiency is straightforwardly 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. In each industry, 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 –0.077 to –0.039. This negative correlation also means that the average (over firms) of the average cost, is higher than the production share weighted average cost. In other words, bigger firms have a lower average cost than smaller firms.

Table 1: Description of 2-digit industries

Industry[a]	Description	# Firms[b]	# Obs.[c]
11	Beverages	3,404	28,558
13	Manufacturing of textiles	6,695	59,549
14	Manufacturing of wearing apparel	14,378	75,828
15	Manufacturing of leather and related products	2,933	21,842
16	Manufacturing of wood and of products of wood and cork	13,115	114,862
17	Manufacturing of paper and paper products	2,725	29,985
18	Printing and reproduction of recorded media	20,611	174,507
20	Manufacturing of chemicals and chemical products	5,104	49,597
21	Manufacturing of basic pharm. products and pharm. preparations	931	8,661
22	Manufacturing of rubber and plastic products	8,511	90,773
23	Manufacturing of other non-metallic mineral products	11,420	98,991
24	Manufacturing of basic metals	2,098	20,181
25	Manufacturing of fabricated metal products	34,578	352,806
26	Manufacturing of computer, electronic, and optical products	6,982	56,847
27	Manufacturing of electrical equipment	4,901	44,049
28	Manufacturing of machinery and equipment	12,974	115,669
29	Manufacturing of motor vehicles, trailers and semi-trailers	4,003	38,262
30	Manufacturing of other transport equipment	1,831	13,745
31	Manufacturing of furniture	14,863	108,587
	Total	**172,057**	**1,503,299**

[a] Statistical classification of economic activities in the European Community, Rev. 2 (2008)
[b] # Firms describes the number of unique firms (ids) active over the period 1994-2019.
[c] # Obs. describes the total number of observations over period 1994-2019.

5.1 Explained and explanatory 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.[10] 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

[10] The sectoral price data are available in the French national accounts at https://www.insee.fr/fr/statistiques/2832666?sommaire=2832834

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.[11] 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. The total cost of production is defined as $c_{nt} = w_{nt}^{\top} x_{nt}$.

5.2 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 1994–2019 period. At the 4-digit level the number of firms is obtained by dividing the total number of observations available for the year 2015 by 184, the number of 4-digit industries, which yields an average number of 310 active firms.[12] 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 is globally decreasing in firm size. On average, most firms have between 2 to 4 employees, representing a share of about 23% of all firms. Table 2 also informs us about market concentration in a typical 4-digit industry: firms with less than 20 employees represent about 74% of all firms, and produce only 12% of total production, whereas the few firms with 500 employees and more produce about 52% of the aggregate (4-digit) production. These figures not only document that there are few actors with 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 in Table 2.[13]

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	Average cost	Profit rate
1	42	13.55	0.37	0.32	92.1	6.2
2–4	73	23.55	1.78	1.05	94.4	3.9
5–9	66	21.29	3.82	2.17	93.9	3.8
10–19	49	15.81	5.69	3.47	93.4	4.1
20–49	47	15.16	12.49	9.12	92.8	4.1
50–99	15	4.84	9.05	7.06	93.9	3.1
100–199	9	2.90	11.09	9.56	94.4	2.5
200–499	6	1.94	15.16	14.56	94.4	2.0
500+	3	0.97	40.55	52.68	95.8	1.1
Total	310	100.0	100.0	100.0	93.6	4.0

[a] Columns 3 to 5 report averages over all 4-digit industries and years (1994–2019). Shares are given in %. Columns 6 and 7 report the median per-unit cost and median profit rate for each firm size.
[b] Firm sizes are measured by the number of employees.

The last two columns of Table 2 report the median values of the average cost, and profit rate over all years and firms within a specific size class. These descriptive statistics show that the median value of the observed average cost of production is smaller for the smallest firms. This helps to understand why there are so many small firms in France compared to other countries. The highest average cost is achieved for the biggest firms (with 500 employees and more). This descriptive/empirical pattern already invites us to conjecture that there is allocative inefficiency at the long-run Cournot equilibrium where

[11] 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

[12] Edmond et al. (2023) calibrate an oligopoly model based on US manufacturing data at the 4-digit level to study the effect of markups on welfare. Hereby, the total number of firms of an average 4-digit industry is 359, with a large part of small firms, which appears to be similar to the patterns in our data.

[13] See also Table B2 in Appendix B, which is complementary to Table 2, and shows the same statistics but for 2-digit industries.

inefficient firms are too large. For large firms, higher average costs are sustainable due to their ability to price above their marginal cost. Heterogeneity in the unit cost of production implies that it is possible to reduce the total cost of production by reallocating output from firms with high average cost to firms with lower average cost. It does not imply, however, that big firms are inefficient and should be closed and replaced by small firms: their minimum level of the average cost could be below the one of small firms, but they are just lacking incentives to produce at this level in order to preserve their market power. In order to identify this inefficiency, we have to go beyond these stylized facts and investigate firms' average cost curve. We have especially to consider unobserved heterogeneity in the cost of production in order to obtain consistent estimates of the cost and output supply functions, and assess the degree of inefficiency of the economy.

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, which is related 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 26 years as in our firm-level data. For the estimation, two years are lost due to differentiating (and so $T = 24$ years).

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

$$\ln Y_{it} \quad = \quad \alpha_i + \alpha_Y \ln Y_{i,t-1} + \alpha_p \ln P_{it} + \alpha_{IM} \ln P_{it}^{IM} + \epsilon_{it} \tag{29}$$

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 considered in Y_{it}. 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 into 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 differentiating over time:

$$\Delta \ln Y_{it} \quad = \quad \alpha_Y \Delta \ln Y_{i,t-1} + \alpha_p \Delta \ln P_{it} + \alpha_{IM} \Delta \ln P_{it}^{IM} + \eta_{it}, \tag{30}$$

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 labor cost, the price index of intermediate consumption as well as the price index exports and 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 an $(L \times L)$ weighting matrix $\mathbf{W}$, the GMM estimator is defined by minimizing in α:

$$\left(\sum_{i=1}^{I} \sum_{t=1}^{T} \eta_{it} z_{it}^{\top} \right) \mathbf{W} \left(\sum_{i=1}^{I} \sum_{t=1}^{T} z_{it} \eta_{it} \right) = \eta^{\top} \mathbf{Z} \mathbf{W} \mathbf{Z}^{\top} \eta \tag{31}$$

The random terms η_{it} and η_{js} are likely to be correlated, both between industries (which are inter-dependent) 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

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

As there is no possibility of consistently estimating 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:

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

The cardinality of this set is $I(3T-2) + I(I-1)T = 12628$ 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 4.5% 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.[14]

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 high p-value of the over-identification test) supports the validity of our

[14] 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.

instruments. According to the GMM estimation results, the estimated short-run elasticity of demand with respect to price is -0.55 and is statistically significant at the 1% threshold. Domestic products and imports are substitutable with a cross price elasticity of 0.46. 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	0.928	0.040	0.759
	(0.02)	(0.04)	(0.07), [0.03]
α_p	-0.072	-0.747	-0.550
	(0.07)	(0.16)	(0.23), [0.08]
α_{IM}	-0.004	0.606	0.463
	(0.06)	(0.16)	(0.23), [0.07]
OIT	-	-	0.99

Notes: HAC robust standard errors in parenthesis, clustered standard errors in brackets. OIT: p-value of the over-identification test (for the validity of 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, obtained for $Y_{i,t-1} = Y_{it}$. The 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 as -1.82 in the short-run and -0.44 in the long-run. Standard errors are obtained using the delta-method (with the HAC variance matrix).

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

Short-run		Long-run	
$\widehat{\varepsilon}\left(Y^d, p\right)$	$\widehat{\varepsilon}\left(P^d, Y\right)$	$\widehat{\varepsilon}\left(Y^d, p\right)$	$\widehat{\varepsilon}\left(P^d, Y\right)$
-0.550	-1.818	-2.283	-0.438
(0.18)	(0.60)	(0.92)	(0.18)

Standard errors are given in parenthesis and estimated by applying the delta method.

The short-run inverse price elasticity is substantial. The estimate of the long-run elasticity of demand wrt price is somewhat bigger (in absolute value) than the estimate of -1.7 obtained by Koebel and Laisney (2016) for US manufacturing (without controlling for the price of imports, however). With Cournot competition, there is an interesting relationship between the markup and the market share y/Y, parameterized by the inverse demand elasticity:

$$\frac{p}{\partial c/\partial y(w,y)} = \frac{1}{1+\varepsilon\left(P^d, Y\right) y/Y}. \tag{32}$$

Using the estimates of Table 4, we draw the estimated short- and long-run relationship between markup and market-share in Figure 3. Firms in the competitive fringe have a markup of 1. In conformity with P1(*iv*), for which Figure 3 provides an illustration, the markup is monotonically increasing in market share. While in the short-run there is substantial markup, in the long-run this markup falls to the interval $1.10 - 1.15$, which is much smaller. Instead, in the short-run, sluggish adjustment toward market equilibrium price and quantity, according to the dynamic relationship (29) with strong anchoring to the lagged aggregate output level, confers substantial market power and a markup of $1.50 - 2.20$ 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.

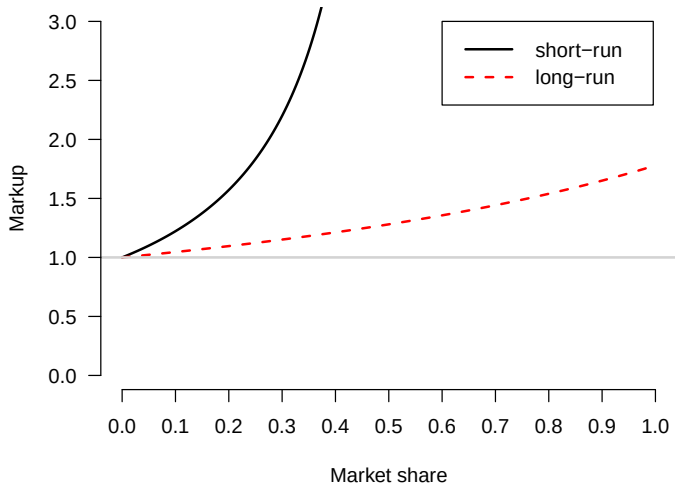

Figure 3: The markup and firms' market share

7 Cost function specification with unobserved heterogeneity

It is well known that unobserved heterogeneity causes estimation biases when it is neglected and correlated with the explanatory variables – see for instance Wooldridge (2010) for a detailed overview of the linear model. Unobserved heterogeneity also raises concerns about the incidental parameters, which, especially in nonlinear models, preclude consistent estimation of parameters and statistics of interest. Martin (2017) and Wooldridge (2019) consider unobserved multiplicative heterogeneity. We also have to deal with the endogeneous output level included as explanatory variable in 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. Further, to allow for variation in firms' technologies, we estimate the cost function for each of the 19 2-digit industries separately.[15]

Unobserved heterogeneity in the fixed and variable costs introduces correlation between their production and the random term. We propose an approach to take this endogeneity into account. Our most general empirical model considers the cost function:

$$c_{nt} = u_{nt}(w_{nt}, t) + v_{1,nt}(w_{nt}, t) y_{nt} + \frac{1}{2} v_{2,nt}(w_{nt}, t) y_{nt}^2 + \eta_{nt}^c. \tag{33}$$

We assume that the random term η_{nt}^c is such that $E[\eta_{nt}^c | w_{nt}, t] = 0$. Regarding unobserved heterogeneity, we assume a somewhat more general specification than the one considered in the theoretical model, and

[15] See Online Appendix B for more details on the estimation procedure.

allow for both multiplicative and additive unobserved heterogeneity. For the variable cost, we consider:

$$v_{j,nt}(w_{nt},t) = \gamma_{nt}^{v_j} v_j(w_{nt},t) + \eta_{nt}^{v_j}, \qquad j=1,2. \tag{34}$$

The fixed cost cannot take negative values, so that we specify:

$$u_{nt}(w_{nt},t) = \max\left\{\gamma_{nt}^{u} u(w_{nt},t) + \eta_{nt}^{u}, 0\right\}. \tag{35}$$

For the sake of identification, we impose:

$$E[\gamma_{nt}^{j}] = 1, \qquad E[\eta_{nt}^{j}] = 0, \qquad j = u, v_1, v_2. \tag{36}$$

Cost heterogeneity $\gamma_{nt} \equiv (\gamma_{nt}^{u}, \gamma_{nt}^{v_1}, \gamma_{nt}^{v_2})$ and $\eta_{nt} \equiv (\eta_{nt}^{u}, \eta_{nt}^{v_1}, \eta_{nt}^{v_2})$ is known by the firm, which uses this information to set its optimal output level in order to equalize marginal revenue and marginal cost:

$$p_t + P'(Y_t) y_{nt} = \gamma_{nt}^{v_1} v_1(w_{nt},t) + \gamma_{nt}^{v_2} v_2(w_{nt},t) y_{nt} + \eta_{nt}^{v_1} + \eta_{nt}^{v_2} y_{nt} + \eta_{nt}^{p} \tag{37}$$

with the random term η_{nt}^{p} such that $E[\eta_{nt}^{p} | w_{nt}, t, \gamma_{nt}^{v}, \eta_{nt}^{v}] = 0$. As γ_{nt}, η_{nt} are unobserved to the econometrician, and because these terms are correlated with output, we need to find suitable control variables to avoid estimation biases. We rely on the assumption:

Assumption 9. *The unobserved technological random terms satisfy (for $j = u, v_1, v_2, c$):*
(i) $E[\gamma_{nt}^{j} | w_{nt}, t, y_{nt}] = E[\gamma_{nt}^{j} | w_{nt}, t, z_{nt}]$, $\quad E[\eta_{nt}^{j} | w_{nt}, t, y_{nt}] = E[\eta_{nt}^{j} | w_{nt}, t, z_{nt}]$
(ii) $E[\gamma_{nt}^{j} | w_{nt}, t, z_{nt}] = E[\gamma_{nt}^{j} | z_{nt}] = \gamma^{j}(z_{nt}) = 1 + (z_{nt} - \overline{z})^{\top} \beta^{j}$
(iii) $E[\eta_{nt}^{j} | w_{nt}, t, z_{nt}] = E[\eta_{nt}^{j} | z_{nt}] = \eta^{j}(z_{nt}) = (z_{nt} - \overline{z})^{\top} \delta^{j}$.

The first two conditions in A9(*i*) imply that the dependence between unobserved heterogeneity terms and y_{nt} can be controlled for by the variables z_{nt}. Similar conditions play a central role in Wooldridge (2019), in the context of correlated random effects. For later use, we also rewrite A9 as:

$$\gamma_{nt}^{j} = \gamma^{j}(z_{nt}) + \zeta_{nt}^{j}, \qquad \eta_{nt}^{j} = \eta^{j}(z_{nt}) + \xi_{nt}^{j}, \tag{38}$$

whose random terms satisfy

$$E[\zeta_{nt}^{j} | w_{nt}, t, y_{nt}] = E[\zeta_{nt}^{j} | w_{nt}, t, z_{nt}] = 0, \tag{39}$$
$$E[\xi_{nt}^{j} | w_{nt}, t, y_{nt}] = E[\xi_{nt}^{j} | w_{nt}, t, z_{nt}] = 0. \tag{40}$$

The last two conditions A9(*ii*) and A9(*iii*) imply that unobserved heterogeneity is mean independent from w_{nt}, t conditionally to z_{nt}. Just as with control functions, conditioning on the variables z_{nt} allows us to control for unobserved correlated heterogeneity. For simplicity, in A9 we restrict the functions γ^{j} and η^{j} to be linear in the parameters and in the control variables z_{nt}. The vector of empirical means $\overline{z}$ is subtracted from z_{nt} to ensure that the unconditional expectations satisfy $E[\gamma_{nt}^{j}] = 1$ and $E[\eta_{nt}^{j}] = 0$. The unobserved $\gamma_{nt}^{j}, \eta_{nt}^{j}$ 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. As these relative efficiency levels are known to the firm, it will produce more when both efficiency indicators are favorable, which makes output y_{nt} endogenous in the expression of the cost function. According to A9, however, these relative efficiency levels depend only upon the control variables z_{nt}. Similar to Olley and Pakes (1996) we consider past investment, the age of the firm, and as recommended by Wooldridge (2019) we include firm-specific averages (correlated with firm specific fixed effects) and the number of

firms' occurrences in the survey, to capture selection effects.[16]

The main empirical implication of A9 is that it allows us to replace the disturbing correlated random terms $\gamma^j_{nt}, \eta^j_{nt}$ in the first-order condition (37), by respectively $\gamma^j(z_{nt}) + \zeta^j_{nt}$ and $\eta^j(z_{nt}) + \xi^j_{nt}$, which comprise the helpful control functions and unproblematic random terms ζ^j_{nt}, ξ^j_{nt}. Using A9, the optimality condition (37) becomes:

$$p_t + P'(Y_t)y_{nt} = \gamma^{v_1}(z_{nt})v_1(w_{nt}, t) + \eta^{v_1}(z_{nt}) + \gamma^{v_2}(z_{nt})v_2(w_{nt}, t)y_{nt} + \eta^{v_2}(z_{nt})y_{nt} + \varepsilon^p_{nt} \tag{41}$$

$$\varepsilon^p_{nt} \equiv \zeta^{v_1}_{nt} + \xi^{v_1}_{nt} + \zeta^{v_2}_{nt}y_{nt} + \xi^{v_2}_{nt}y_{nt} + \eta^p_{nt}. \tag{42}$$

Under A9, the random term ε^p_{nt} satisfies $E[\varepsilon^p_{nt}|w_{nt}, t, z_{nt}] = 0$, but is correlated with y_{nt} through η^p_{nt}. We circumvent the endogeneity of y_{nt} in (41), by solving this optimality condition in y, which gives our output supply function:

$$\begin{aligned} y_{nt} &= y^s(p_t, w_{nt}, t, z_{nt}) + \varepsilon^y_{nt} \\ &= \frac{p_t - \gamma^{v_1}(z_{nt})v_1(w_{nt}, t) - \eta^{v_1}(z_{nt})}{\gamma^{v_2}(z_{nt})v_2(w_{nt}, t) + \eta^{v_2}(z_{nt}) - P'(Y_t)} + \varepsilon^y_{nt}. \end{aligned} \tag{43}$$

The random term ε^y_{nt} is such that

$$E[\varepsilon^y_{nt}|p_t, w_{nt}, t, z_{nt}] = 0. \tag{44}$$

It turns out that under A9, we can consistently estimate the parameters of the output supply function by nonlinear least squares.

The output supply, however, does not allow us to identify the total cost function, because the fixed cost and its heterogeneity distribution over firms (and time) cannot be identified from the expression of y^s. For this reason we append to our model a reformulated cost function. Equation (33) is problematic because, even under A9, the random term η^c_{nt} is likely to be correlated with y_{nt}: random shocks affecting costs lead firms to adjust output. To avoid this difficulty, we substitute y_{nt} by

$$y^s_{nt} + \varepsilon^y_{nt},$$

in the expression of the cost function (for the sake of conciseness, we skip the arguments of the supply function and add a subscript nt to denote the function values). We assume that the random term of the supply function exhibits some heteroskedasticity of the form:

$$\sigma^2_y \equiv E[(\varepsilon^y_{nt})^2|w_{nt}, t, z_{nt}] = \sigma^2_0 + \sigma^2_1 y^s_{nt}, \tag{45}$$

where σ_0, σ_1 denote constant variance parameters, which are squared to ensure that σ^2_y is positive. This allows us to replace in the cost function, the squared production level by

$$y^2_{nt} = (y^s_{nt})^2 + \sigma^2_y + \nu_{nt}, \tag{46}$$

where $E[\nu_{nt}|w_{nt}, t, z_{nt}] = 0$. With these notations, the cost function becomes:

$$\begin{aligned} c_{nt} = \max\{\gamma^u(z_{nt})u(w_{nt}, t) + \eta^u(z_{nt}), 0\} + \gamma^{v_1}(z_{nt})v_1(w_{nt}, t)y^s_{nt} + \eta^{v_1}(z_{nt})y^s_{nt} \\ + \frac{1}{2}\gamma^{v_2}(z_{nt})v_2(w_{nt}, t)((y^s_{nt})^2 + \sigma^2_y) + \frac{1}{2}\eta^{v_2}(z_{nt})((y^s_{nt})^2 + \sigma^2_y) + \eta^c(z_{nt}) + \varepsilon^c_{nt}, \end{aligned} \tag{47}$$

[16] See Appendix B, Table B3, for some descriptive statistics for these variables.

where

$$\varepsilon^c_{nt} \equiv \zeta^u_{nt} u(w_{nt},t) + \xi^u_{nt} + \zeta^{v_1}_{nt} v_1(w_{nt},t) y_{nt} + \xi^{v_1} y_{nt} + \frac{1}{2}\zeta^{v_2} v_2(w_{nt},t) y^2_{nt} + \frac{1}{2}\xi^{v_2} y^2_{nt} \tag{48}$$
$$+ \gamma^{v_1}(z_{nt}) v_1(w_{nt},t)\varepsilon^y_{nt} + \eta^{v_1}(z_{nt})\varepsilon^y_{nt} + \frac{1}{2}\gamma^{v_2}(z_{nt}) v_2(w_{nt},t)\nu_{nt} + \frac{1}{2}\eta^{v_2}(z_{nt})\nu_{nt} + \eta^c_{nt}.$$

As under A9, $E[\varepsilon^c_{nt}|w_{nt},t,z_{nt}] = 0$, we can append equation (47) to (43) and form a system whose parameters can be consistently estimated by nonlinear least squares.

We specify the parametric functional forms for u v_1 and v_2 and consider that they belong to the family of quadratic cost functions:

$$u\left(w,t;\theta^u\right) = \theta^\top_w w + \theta^\top_{wt} wt + \frac{1}{2}\frac{w^\top \Theta_{ww} w}{\zeta^\top w}, \tag{49}$$

$$v_1\left(w,t;\theta_1\right) y = \left(\theta^\top_{1w} w + \theta^\top_{1t} wt + \frac{1}{2}\frac{w^\top \Theta_{1ww} w}{\zeta^\top w}\right) y \tag{50}$$

$$v_2\left(w;\theta_2\right) y^2 = \left(\theta^\top_{2w} w\right) y^2 \tag{51}$$

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 functions of w, we impose that

$$\Theta_{ww} = \Theta^\top_{ww}, \quad \Theta_{1ww} = \Theta^\top_{1ww}, \tag{52}$$

$$\iota^\top \Theta_{ww} = \iota^\top \Theta_{1ww} = 0 \tag{53}$$

where ι denotes a $(J \times 1)$ vector of ones. We use the 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 issue. There is a total of $5J + J(J-1)$ free parameters. In our case, $J = 3$ and there are 21 free θ parameters in the deterministic part of the cost function, and 51 further $\beta, \delta, \sigma^2_y$ parameters behind unobserved (and correlated) heterogeneity.

8 Estimation results

The theoretical model outlined in Sections 2 to 4 corresponds to a specific case of the more general empirical model of Section 7. For this reason, we do not expect the statements of the different propositions to be satisfied at each observation. However, we expect to see the results valid on average, over the years and the population of firms. We discuss these estimates and their relationship with the model below.

8.1 Unobserved heterogeneity

We present our estimates of the unobserved fixed and variable cost efficiency (which corresponds to stochastic technological change). In Section 7, we introduced 6 heterogeneity terms. To shorten the presentation and simplify their interpretation, we aggregate these 6 terms in 2 terms compatible with our theoretical part, namely equations (10) and (16). The aggregate fixed cost heterogeneity is defined as the ratio between individual and the mean fixed cost function evaluated at w_{nt}. Similarly, the variable cost heterogeneity corresponds to the ratio between individual and mean variable cost function values

(obtained for $\gamma_1^v = \gamma_2^v = 1$ and evaluated at w_{nt}, y_{nt}):

$$\widehat{\gamma}_{nt}^u \equiv \frac{\widehat{u}_{nt}}{\widehat{u}(w_{nt})}, \qquad \widehat{\gamma}_{nt}^v \equiv \frac{\widehat{v}_{nt}}{\widehat{v}(w_{nt}, \widehat{y}_{nt})}. \tag{54}$$

Table 5 reports some percentiles of their respective distribution. Fixed cost is found to be zero for most firms (71%), but it is significant for about 29% of the observations. There is considerable heterogeneity about the size of these fixed costs. The distribution of $\widehat{\gamma}^v$ is centered on 1, and with a larger tail on the left of the median than on the right.

Table 5: Distribution of firms' unobserved heterogeneity

	Q10	Q25	Q50	Q75	Q90
$\widehat{\gamma}_{nt}^u$	0.00	0.00	0.00	0.19	3.56
$\widehat{\gamma}_{nt}^v$	−2.55	0.64	1.00	1.11	1.23

Note: Q10 to Q90 report the 10^{th} to the 90^{th} percentile of the respective distribution.

The parameter γ^v represents variable cost heterogeneity. While about 25% of the firms have a variable cost more than 16% below average (for which $\gamma^v = 1$), there are also 25% of the firms with average costs higher than average by 6% or more. This unobserved heterogeneity is estimated to be economically relevant and, according to Proposition 3, we expect it to strongly influence a firm's size.

Figure 4 and 5 show kernel density estimates of the distribution of $\widehat{\gamma}^u$ (on the left) and $\widehat{\gamma}^v$ (on the right).[17] Both densities are single peaked, and show that there is a high probability mass around $\gamma^u = 0$ and around $\gamma^v = 1$.

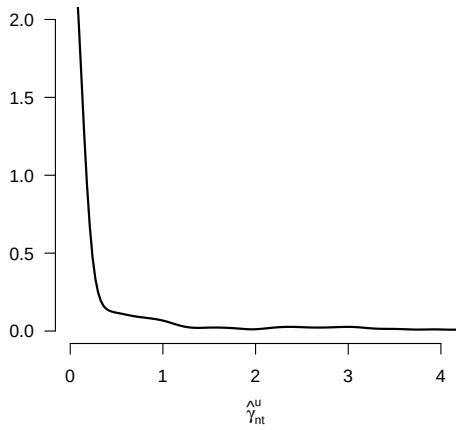

Figure 4: The density of fixed cost heterogeneity $\widehat{\gamma}^u$

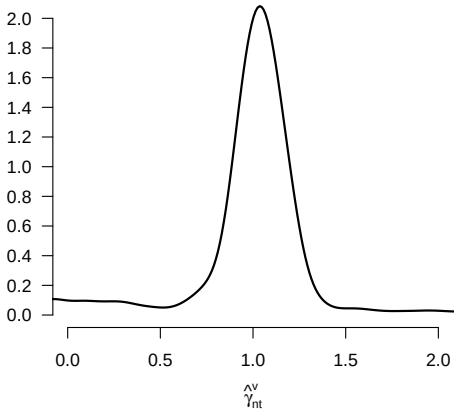

Figure 5: The density of variable cost heterogeneity $\widehat{\gamma}^v$

Table 6 summarizes the percentage of estimates corresponding to 4 possible estimated signs of the linear and quadratic parts of the variable cost function, $v_{1,nt}$ and $v_{2,nt}$, which vary for each observation over the sample. In almost all cases, predicted marginal costs are convex (98% of the observations) with both $v_{1,nt}$ and $v_{2,nt}$ positive. In 0.8% 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. We identify all observations for which the estimated marginal cost was found to be decreasing, and compute

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

their median output level: it was found to be 31% smaller than the median output of firms with increasing marginal cost.

Table 6: Share of observations for different types of heterogeneity in $v_{1,nt}, v_{2,nt}$ in %

	$v_{1n} \leq 0$	$v_{1n} > 0$
$v_{2n} \leq 0$	1.2	0.8
$v_{2n} > 0$	15.3	82.7

Note: Figures are given in %.

Table 7 reports the estimates of unobserved heterogeneity over firm size. The share of firms with no fixed costs and the median value of the fixed cost are given in the first two columns. For small firms, we find that the large majority of firms have no fixed costs: 83% of the firms with one employee produce without any fixed costs, but this rate decreases in firm size: less than 38% of the firms with more than 200 employees have no fixed costs. An economic narrative can be provided: small firms have lower profits (although their profit rate is higher), more credit constraints and a higher probability of bankruptcy, which incite them to invest in technologies with no fixed cost. Bigger firms can afford fixed costs, which are not directly productive (like organisational costs, or cost-reducing research development expenditures), and which allow them to reduce the variable cost of production. The estimates of Table 7 are in line with the theoretical predictions: the fixed cost parameter is larger for bigger firms, and the marginal (and variable) cost parameter is lower for bigger firms, in conformity with P3(i). These findings also highlight the shortcomings of usual specifications for cost functions, such as the Cobb-Douglas or the translog, which exclude, by construction, the occurrence of fixed costs.

Table 7: Fixed and variable costs by firm size

Firm size	$S_{u=0}$	$\widehat{u}_{nt}/c_{nt}$	$\widehat{\gamma}^u_{nt}$	$\widehat{\gamma}^v_{nt}$
1	83.34	0.00	0.00	1.08
2–4	76.79	0.00	0.00	1.05
5–9	76.90	0.00	0.00	1.02
10–19	75.90	0.00	0.00	1.00
20–49	58.76	0.00	0.00	0.95
50–99	43.57	0.10	0.63	0.88
100–199	41.06	0.13	1.98	0.78
200–499	37.55	0.14	5.62	0.70
500+	8.58	0.39	44.08	0.55
Total	70.73	0.00	0.00	1.00

Notes: Firm sizes are measured by the number of employees. $S_{u=0}$ denotes the share of firms with zero fixed cost. Column $\widehat{u}_{nt}/c_{nt}$ reports the median value of the share of fixed cost in total cost. Columns $\widehat{\gamma}^u_{nt}$ and $\widehat{\gamma}^v_{nt}$ report the median value of the estimates $\widehat{\gamma}^u_{nt}$ and $\widehat{\gamma}^v_{nt}$.

8.2 Returns to scale and rate of technological change

The rate of Returns to Scale (RTS) is defined by

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

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):

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

Here we only take into account the direct effect of t on total cost, for constant level of the technological parameters (which also change over time). These statistics depend upon the explanatory variables (both observed and unobserved) and are different for each observation in our sample. Table 8 summarizes the estimates of these elasticities over all observations of the sample. It is interesting to note that about 35% of the estimates correspond to increasing RTS, about 25% to constant RTS (with a cost to output elasticity comprised between 0.92 and 1.11), and 25% to decreasing RTS. The distribution is not symmetric, but positively skewed: the lower percentiles are further away from 1 than the higher percentiles. 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 frequently yields decreasing RTS. 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 both for unobserved heterogeneity and endogeneity of output, our estimates are not affected by these sources of bias.

Table 8: Distribution of firms' returns to scale and rate of technological change

	Q10	Q25	Q50	Q75	Q90
$\partial \ln c / \partial \ln y$	0.59	0.92	1.04	1.11	1.22
$\partial \ln c / \partial t$	–0.26	–0.04	0.00	0.04	0.32

Note: Q10 to Q90 report the 10^{th}, to the 90^{th} percentile of the respective distribution.

The RTC represents deterministic technological change, because the time trend t is not random. The results show a negative RTC for about half of the estimates. The estimates corresponding to the lower and higher quantiles are quite large.

One of the main conclusions of the Cournot model with heterogeneity is that there is an ordering of unobserved heterogeneity and firm size. We investigate this relationship further and report statistics by firm size. Table 9 completes the information given in Tables 8 and reports the quartiles of RTS and RTC by firm size. The median value of RTS is globally diminishing with firm size by about 6%. The share of firms with increasing RTS is smaller among small firms than for bigger firms.

Regarding deterministic technological change, the estimated median value of $\partial \ln c / \partial t$ is stable with firm size. RTC is important for medium-sized firms (column Q25), and represents a cost reduction of about 5% and more by year, *ceteris paribus*. This rate then decreases with firm size (in absolute value), and is close to 5% for the largest firms.

Table 9: Median RTS and RTC statistics by firm size

		c/y			RTS			RTC	
Firm size	Q25	Q50	Q75	Q25	Q50	Q75	Q25	Q50	Q75
1	91.4	100.9	207.7	0.89	1.06	1.20	–0.07	0.00	0.21
2–4	93.5	108.7	261.1	0.90	1.04	1.13	–0.06	0.00	0.12
5–9	92.9	111.0	279.4	0.93	1.04	1.10	–0.03	0.00	0.05
10–19	91.2	108.6	253.1	0.95	1.04	1.10	–0.02	0.00	0.02
20–49	87.5	102.1	205.3	0.95	1.04	1.10	–0.04	0.00	0.02
50–99	86.8	104.1	230.2	0.94	1.03	1.09	–0.04	–0.01	0.01
100–199	83.0	104.4	196.8	0.92	1.02	1.08	–0.04	–0.01	0.01
200–499	81.2	100.4	225.4	0.91	1.02	1.08	–0.03	–0.01	0.01
500+	94.0	126.7	437.8	0.86	1.00	1.07	–0.04	–0.01	0.01
Total	91.0	107.0	245.1	0.92	1.04	1.11	–0.04	0.00	0.04

Notes: Firm sizes are measured by the number of employees. Q25, Q50, and Q75, respectively, denote the lower quartile, the median, the upper quartile of the estimated statistics.

Figure 6 shows the evolution of the median of the unobserved variable cost efficiency ($\widehat{\gamma}^v$, solid line) and the RTS (dashed line), where both measures refer to the left y-axis, as well as the RTC (dotted line),

referring to the right y-axis. The RTC corresponds to the median value (over all firms) of the estimates for $d \ln c/dt$. While it is found to be varying around zero, the effect on total cost is substantial. It should be noted that the reported median values include only the deterministic changes over time. The stochastic changes in $\widehat{\gamma}^v$ correlated with y and t are not included in this estimate of the reported RTC statistic. This stochastic technological change is estimated by $\widehat{\gamma}^v$, which is found to be increasing over 1994–2004 and decreasing from 2010 to 2017. The median value of the RTS varies little over time, between 1.02 and 1.06. Even though the median RTS is close to constant, there is substantial heterogeneity around this value (see tables above).

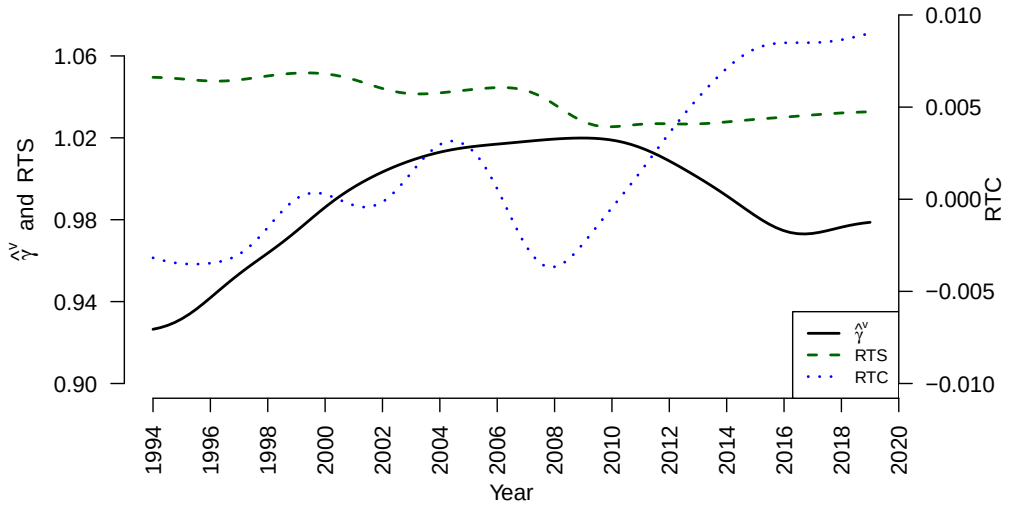

Figure 6: Median evolution of unobserved variable cost efficiency ($\widehat{\gamma}^v$), the rate of Return to Scale (RTS) (both left y-axis), and the Rate of Technological Change (RTC) (right y-axis). The time series are filtered using a kernel-smoother.

9 Welfare implications of market power

In this section we simulate the different policies implied by SROW and LROW. That is, using the estimated parameter values, we now investigate the welfare implications of market power and related technological inefficiencies outlined in Section 4. While the NLS estimates provide decent fit between predicted and observed level of production over all industries (see Appendix C, Table C1), we simulate the policies only for firms belonging to the 6 2-digit industries with the highest model fit, in order to reduce computation time and increase prediction accuracy (these are industries 11, 16, 22, 23, 27, and 31, see Table 1 for a description).[18]

Before exploring the outcome of simulated reallocation(s) of production, we have to investigate whether assumption A6, required for P3, P5, and P6, and conjecturing a decreasing and convex relationship between γ^u and γ^v, is empirically supported. Figure 7 reports the estimated values of the parameters for the firms belonging to the 6 selected industries at the year 2015. We consider a single year in order to reduce computational burden, and avoid dealing with technological change, entry and exit. The orientation of the plots confirms that the estimates are broadly compatible with A6. This evidence gives further support for the statement of the propositions.

[18] See Online Appendix C more more details on the simulation procedures.

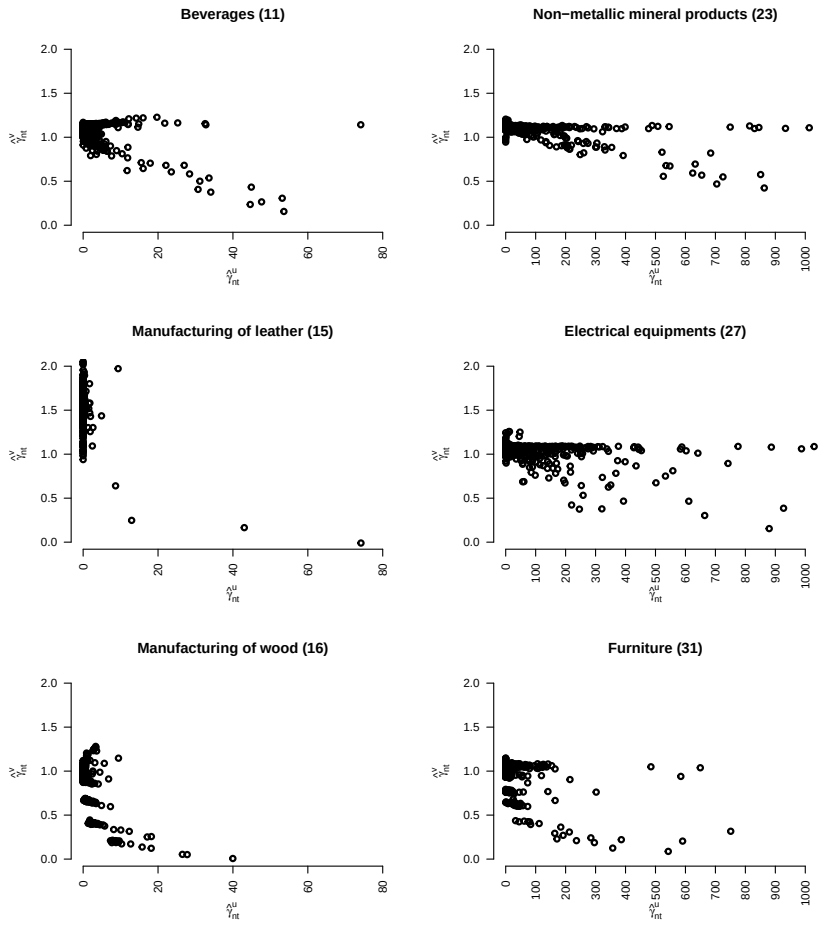

Figure 7: Estimates of unobserved fixed and variable cost efficiency, γ^u and γ^v, for selected industries of the year 2015

We now turn to the simulation. First, we use the parameters' estimates and compute the LRCE, characterized by (22) and (23). The fixed point of the economy is numerically characterized in the second row of Table 10. The simulated LRCE values for aggregate price and quantities are close to the actual data (reported in the first row), which is anticipated, because the LRCE model was estimated to fit the data. There are some differences, however, in the level of welfare, average costs and profit rate. This is partly due to the simulation of the Cournot equilibrium, which imposes a production level equal to zero on firms whose optimal production level would have been negative. About two firms in 77 cases are affected by this corner solution (see Table 10). It turns out that the simulated levels of average cost are somewhat higher, while the simulated profit rate and welfare are lower at the LRCE in comparison to the data. Overall, the gap between the data and the model is rather small and the estimates are plausible.

Second, we simulate the SROW policy, whose results are reported in the third row of Table 10. Here, the simulation consists in setting the market power of all firms to zero, which redistributes individual outputs over firms in order to resolve the inefficiency due to market power. This new and regulated optimum is described in Section 4.1, whose properties are given by P4 and P5. At SROW all firms producing at the LRCE are still active (some firms produce zero output, though), so there is still a technological inefficiency due to too many occurrences of inefficient firms. According to P4, a competitive output reallocation would improve welfare, increase total output, reduce the price, and, under the assumptions of P4, increase the concentration of output over firms. We simulate the SROW separately for each of the 6 industries and find only small empirical support for welfare-reducing market power effects in the short-run. The results reported in the SROW row of Table 10 show that only a small decline in aggregate

price of about 0.7% can be achieved through the suppression of market power, while aggregate quantities stay almost constant (and even slightly decrease). Also, this policy slightly decreases the median profit rate, to –0.1% (from an initial level of 0.0% at the LRCE).

Given the predictions of P4, we initially expected to find output being reallocated from smaller to larger firms (with lower marginal costs). However, the results reported in Table 10 show that industry concentration slightly declines. This is not in contradiction with P4, because the proposition holds when input prices are the same for all firms. In the data, however, average wages tend to be higher in bigger firms, which reduces their profit as well as the welfare gains of allocating supplementary production to those bigger firms. Our main conclusion is that the detrimental welfare effect of imperfect competition is small: If all firms with market power set the price equal to their marginal cost, aggregate production would hardly rise, and the aggregate price would only slightly decrease. This contrasts with the recent literature on market power in the US, where Baqaee and Farhi (2020) and Edmond et al. (2023) find more sizable effects of markups on aggregate productivity and welfare.

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

	Y	P	W	c/y	u/y	π	N	HH	C_3	C_{10}
Data	234.5	99.0	124.3	98.4	-	1.3	79.0	12.0	48.3	74.0
LRCE	239.2	100.3	153.5	101.9	5.3	0.0	77.0	11.3	46.9	73.4
SROW	238.7	99.6	176.9	101.0	5.4	–0.1	77.0	9.1	42.4	65.8
LROW Q75	238.7	99.2	183.2	98.3	0.0	0.7	309.7	-	-	-
LROW Q90	238.8	97.2	191.1	96.7	0.0	0.7	263.9	-	-	-
LROW Q99	239.4	96.8	195.4	95.9	0.0	1.5	88.4	-	-	-

Notes. The table reports median values over all 4-digit industries. The raw values of the variables Y, P, and W, are computed at the 2-digit industry level (over 6 industries). The estimates of c/y (average cost), u/y (average fixed cost), and π are computed for each firm. The statistics N, HH, CR_3 and CR_{10} are computed at the 4-digit aggregation level. N corresponds to the number of active firms. The Hirschman-Herfindahl index and two concentration ratios (resp. for 3 and 10 firms) are denoted by HH, C_3 and C_{10}). At LROW the concentration indices are not computed as they are equal, respectively, to $1/N, 3/N$, and $10/N$.

Figure 8 sheds further light on the narrow price and output gap between LRCE and SROW. It represents the median value of relative wages (given by the dots) over all firms within a given size bin, together with the median value of the variable cost heterogeneity $\widehat{\gamma}_n^v$ (given by the cross).

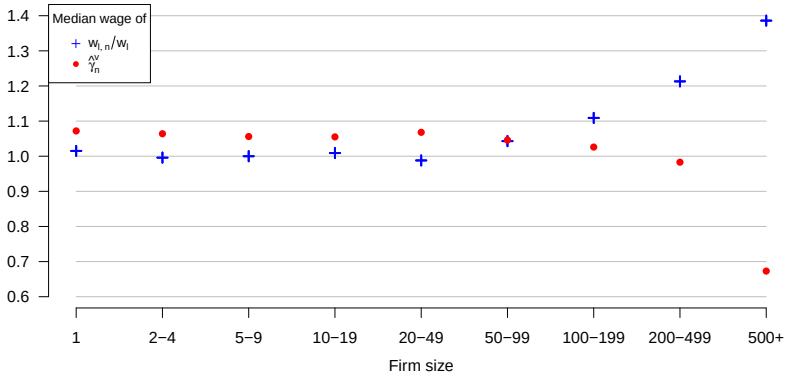

Figure 8: Median values of relative wages and $\widehat{\gamma}_n^v$, by firm size, for the year 2015

This figure helps us to understand why the Cournot equilibrium is empirically very close to the SROW. For constant wages, slight differences in the γ^v parameter yield substantial differences in marginal cost, market power, and welfare. Figure 8 provides empirical evidence for the fact that bigger firms have (in

the median) a substantially lower value of the γ^v parameter, and pay higher wages than smaller firms. This increases the marginal cost of these firms, and reduces the negative welfare effect of market power. More precisely, compared to firms with less than 99 employees, those firms with 100–199, 200–499, and 500+ employees, exhibit a higher median value of wages of 11%, 21%, and 39%, respectively, and a lower value of the median variable cost parameter $\widehat{\gamma}_n^v$ of 2%, 6%, and 36%, respectively.

The last three rows of Table 10 describe LROW simulations which consist in removing firms' market power and replicating for the firm closest to the 0.75 quantile of the profit rate distribution (and similarly for the 0.90 and 0.99 profit quantile). This procedure is inspired by P6(v), and the fact that total profit is a good indicator of technological efficiency. More specifically, for each 4-digit industry belonging to a given 2-digit industry, we identify the firm closest to the 75% (90%, 99%) quantile of the profit rate distribution, which will be replicated. This induces reallocation among firms and adjustments in aggregate price and output (just as described microeconomic textbooks). The iterative replication process stops when the profit is close to zero in one of the 4-digit industries. For realism, we impose the constraint that the weight of each 4-digit industry in the aggregate 2-digit industry is constant.

Our replication exercise shows that the price and the average cost of production decreases when we select and replicate more efficient firms and go from the 0.75 to the 0.99 quantile. At the same time, aggregate production and welfare increase. The average number of active firms in the LROW is close to 310 for the row Q75, and decreases to 264 and 88 when a more efficient firm is replicated. This is a consequence of the fact that more efficient firms are bigger on average. We also note that the median value of the fixed cost is zero for the replicated firms. This shows that there exist efficient technologies with no fixed cost.

10 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. For given input prices, a useful theoretical result allows us to infer the ordering of firms' unobserved cost efficiency from the (observed) ordering of firms' sizes. A further result generalizes Mankiw and Whinston (1986)'s theorem about excess entry at Cournot equilibrium to the case of heterogeneous firms. Once firms' heterogeneity is considered in the analysis, excess entry concerns small and inefficient firms which do not contribute to reducing the market share and market power of bigger and efficient firms.

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. The question of how to best deal with this contradiction from a welfare perspective is an empirical question and should, as we propose, be tackled using both estimations and simulations. The estimation results confirm that unobserved cost heterogeneity is substantial, and affects both the fixed and the variable costs. A main finding presented in the paper, and identified using firm-level data, is the negative correlation between fixed and variable cost heterogeneity. A second result, obtained by using simulations, contrasts with the existing literature on misallocation due to market power. Our simulation results show that removing market power has a negligible impact on aggregate production and price, and triggers little output reallocation from bigger to smaller firms. The third result of our contribution is provided by the simulation which consists in replicating technologically efficient firms, and removing inefficient firms from the market. The effect on output price and welfare is quite substantial, and achieved through the reduction of the average cost of production. The policy implication that we draw from this simulation, is that in France cost-reducing innovations and technological progress are more likely to improve welfare than policies aiming to fight against market power.

Appendix

A Proof of the propositions

Proof of Proposition 1

P1(*i*) and P1(*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 P1(*i*). The inequality also implies that individual market shares are bounded above: $y_n^N/Y < -1/\epsilon(P,Y)$.

P1(*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 P1(*iv*) directly follows from P1(*iii*) and the definition of the price markup $P/(\partial c_n/\partial y)$.
Claim P1(*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 no effect on the Nash equilibrium. These inequalities prove claims (i) to (iv). Claim P2(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

P3(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

$$\begin{aligned}&\frac{\partial c_i}{\partial y}\left(w,y_i^N\right)>\frac{\partial c_j}{\partial y}\left(w,y_j^N\right)\\ \Leftrightarrow\quad &\gamma_i^v\cdot\left(v_1+v_2y_i^N\right)>\gamma_j^v\cdot\left(v_1+v_2y_j^N\right).\end{aligned}\tag{57}$$

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 $\gamma_i^v > \gamma_j^v$.
P3(*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 P3(*i*) and A2 implies that $v_1 \geq 0$. The condition $a + b \geq 0$ corresponds to (57). The implied inequality $a + b/2 \geq 0$ is equivalent to claim P3(*ii*).
P3(*iii*). For $\gamma_i^v > \gamma_j^v$, and same technological shock η, relationship A7 implies that $\gamma_i^u < \gamma_j^u$ and $u_i(w) < u_j(w)$.
P3(*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

P4(*i*). At the LRCE characterized by (3), it turns out that for any active firm,

$$P(Y_{-n}^C + y_n) - \frac{\partial c_n}{\partial y_n}(w_n, y_n) \geq 0. \tag{58}$$

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 planner chooses $\{y_m\}_{m=1}^M$ in order to satisfy $P\left(\sum_{m=1}^M y_m\right) - \partial c_n/\partial y_n\,(w_n, y_n) = 0$ 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 $P\left(Y^S\right) \leq P\left(Y^C\right)$.
P4(*ii*). By definition, W^S maximizes welfare by choosing the optimal level of production over all firms active at the LRCE, hence $W^S \geq W^C$. It follows directly from P4(*i*) and profit maximization, that:

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

P4(*iii*)–P4(*v*). At the aggregate production level $Y^S \geq Y^C$ the firms' production plans have to satisfy:

$$\frac{\partial c_m}{\partial y_m}(w_m, y_m^S) = \frac{\partial c_n}{\partial y_n}(w_n, y_n^S) = P\left(Y^S\right), \tag{59}$$

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 (59). Let us order firms from lower to higher marginal cost, and define "bigger firms" as those having at the LRCE a marginal cost lower than $P\left(Y^S\right)$, and "smaller firms" the other group with $\partial c_n/\partial y_n\,(w_n, y_n) \geq P\left(Y^S\right)$.
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 (59) 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 (59).

- smaller firms with decreasing marginal cost produce more if this allows them 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 (59). If this is not possible, they should stop their activity.

P4(*vi*). In points P4(*iii*)–P4(*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 P4(*vi*) can be sharpened 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 planner 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 the 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 planner 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

P6(i). Under the above assumptions, W is continuous, nondecreasing in γ, and the set of values taken by the welfare function over Γ^L is closed and bounded from below, and so for any given level of y, W admits a maximum over Γ. 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.
P6(ii). From P6(i) it follows that at the LROW point, the planner 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:

$$\mathbf{W}^L(Ny) = \int_0^{Ny} P(s)\, ds - Nc^L(w, y), \tag{60}$$

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 $c^L(w, y^L)/y^L = \partial c^L/\partial y(w, y^L) = P(Y^L)$, and 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.)
P6(iii). Both optimization problems (26) and (25) have the same objective function, but there are fewer constraints in (26), hence $W^L \geq W^S$.
P6(iv). If the inequality holds, then the Kuhn and Tucker complementary slackness condition implies that $\gamma^u = 0$.
P6(v). The claim follows because the first- and second-order conditions to both problems are identical. □

B Data and descriptive statistics

B.1 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 B1 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 772. This is mainly induced by the inclusion in Table B1 of industry 10 and to a smaller extent 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 6.7% (53.0%) of total production, 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 sector.

Table B1: 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	Average cost	Profit rate
0	153	26.06	0.02	3.69	106.50	–6.79
1	59	10.05	0.48	0.35	93.62	4.43
2–4	88	14.99	2.00	1.03	95.66	2.76
5–9	74	12.61	3.99	2.07	94.67	3.17
10–19	52	8.86	5.72	3.36	93.74	3.71
20–49	49	8.35	12.39	8.62	92.74	4.01
50–99	16	2.73	8.88	6.55	93.83	3.08
100–199	9	1.53	10.86	8.82	94.40	2.41
200–499	6	1.02	14.88	13.65	94.41	1.95
500+	3	0.51	40.77	51.38	95.87	1.03
NA	78	13.29	0.00	0.48	100.44	–2.42
Total	587	100.00	99.99	100.00	96.37	1.98

[a] All figures represent averages over all four-digit industries and years (1994–2019). 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.2 Further descriptive statistics

Table B2 shows shares of firms, employees, and production wrt each considered 2-digit industry. The table shows that the manufacturing of metal products (25) represents the biggest industry in terms of the average number of firms and average employment, representing about 23% of all firms and 14% of total employment. Instead, the manufacturing for motor vehicles (29) represents the biggest industry in terms of production, accounting for about 15% of total production. See also De Monte (2024) for more descriptive statistics using similar data.

Table B2: Average statistics by 2-digit manufacturing industry[a]

Industry[b]	# of firms	Share of firms	Share of employees	Share of production	Average cost	Profit rate
11	1098	1.90	1.78	4.01	86.27	3.12
13	2290	3.96	2.93	1.95	91.87	2.68
14	2916	5.04	3.19	1.67	93.76	1.36
15	840	1.45	1.42	0.86	78.09	2.13
16	4417	7.64	2.95	2.03	90.13	4.34
17	1153	1.99	3.37	3.31	95.30	2.96
18	6711	11.61	3.65	1.84	102.94	5.82
20	1907	3.30	7.27	12.61	90.28	0.63
21	333	0.58	3.55	4.35	114.14	1.28
22	3491	6.04	8.52	6.09	94.70	3.49
23	3807	6.59	5.46	5.45	91.08	2.86
24	776	1.34	3.78	5.38	89.92	2.96
25	13569	23.47	13.78	9.31	87.78	6.31
26	2186	3.78	6.41	4.53	157.99	0.42
27	1694	2.93	6.00	5.13	91.85	3.04
28	4448	7.69	8.14	6.93	96.96	1.19
29	1471	2.54	9.87	14.99	95.44	0.12
30	528	0.91	5.35	8.14	96.34	0.40
31	4176	7.22	2.58	1.42	87.21	3.92
Total	57811	100.00	100.00	100.00	93.60	4.00

[a] All figures are based on the cleaned dataset and represent averages over the period 1994–2019. 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 B3 illustrates the distribution of some variables included in z_{nt} to capture unobserved heterogeneity for the estimation of the cost function (see Section 7 and A9). 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 1994–2019 period. Beside the average number and the average share of firms, the table reports the share of investing firms, the investment-to-labor ratio, and 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}, is 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 60%, whereas almost all firms with 500 and more employees report investments in capital (99%). 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 12.3 years (4.9 periods), the largest size group, firms reporting 500 and more employees, are on average 31.4 years old (and observed on average for 14.1 periods). Firms' age, a_{nt}, is calculated as the difference between the current year and the date of creation of the firm. Note that 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 B3: 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	42	13.55	60.08	0.16	12.39	4.87
2–4	73	23.55	69.85	0.07	14.21	7.81
5–9	66	21.29	81.89	0.06	17.45	10.58
10–19	49	15.81	90.20	0.06	20.66	12.31
20–49	47	15.16	94.60	0.06	23.69	12.40
50–99	15	4.84	96.64	0.07	26.49	12.93
100–199	9	2.90	97.66	0.08	27.84	13.32
200–499	6	1.94	98.29	0.10	28.46	13.89
500+	3	0.97	98.69	0.12	31.41	14.16
Total	310	100.01	81.00	0.08	18.55	10.00

[a] All figures are based on the cleaned dataset and represent averages over the period 1994–2019. Shares are given in %.
[b] Firm size is measured by the number of employees.

C Further estimation results

Table C1 summarizes the correlation between fitted and observed values obtained over 19 NLS regressions (for each 2-digit industry). The table shows that the NLS estimates provide decent fits, which is necessary for our simulation of output redistribution from less to more productive firms to make sense. In order to reduce computation time and increase prediction accuracy, we only consider firms belonging to the 6 2-digit industries with the best fit between predicted and observed level of production (given by the industries 11, 16, 22, 23, 27, and 31, see Table 1 for a description).

Table C1: Correlation between observed and predicted values

	$cor_N(c_{nt}, \widehat{c}_{nt})$	$cor_N(y_{nt}, \widehat{y}_{nt})$	$cor_N(mr_{nt}, \widehat{\partial c/\partial y})$
Lower quartile	0.86	0.77	0.99
Median	0.91	0.86	0.99
Upper quartile	0.94	0.91	0.99

The correlations are computed for each of the (19) 2-digit industry separately, using industry-specific parameters' estimates. The table reports the quartiles of these 19 correlations.

References

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

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.

Baily, M. N., Hulten, C., Campbell, D., Bresnahan, T. and Caves, R. E. (1992). Productivity dynamics in manufacturing plants, *Brookings papers on economic activity. Microeconomics* **1992**: 187–267.

Baqaee, D. R. and Farhi, E. (2020). Productivity and misallocation in general equilibrium, *The Quarterly Journal of Economics* **135**(1): 105–163.

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., Gaynor, M. and Morton, F. S. (2019). Do increasing markups matter? Lessons from empirical industrial organization, *Journal of Economic Perspectives* **33**(3): 44–68.

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.

Ceci-Renaud, N. and Chevalier, P.-A. (2010). L'impact des seuils de 10, 20 et 50 salariés sur la taille des entreprises françaises, *Economie et statistique* **437**(1): 29–45.

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. (2024). Productivity, markups, and reallocation: Evidence from French manufacturing firms from 1994-2016, *ZEW Discussion Paper* **24-002**.

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. (1987). Flexible functional forms and global curvature conditions, *Econometrica* **55**(1): 43–68.

Edmond, C., Midrigan, V. and Xu, D. Y. (2023). How costly are markups?, *Journal of Political Economy* **131**(7): 1619–1675.

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.

Gandhi, A., Navarro, S. and Rivers, D. A. (2020). On the identification of gross output production functions, *Journal of Political Economy* **128**(8): 2973–3016.

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.

Gourio, F. and Roys, N. (2014). Size-dependent regulations, firm size distribution, and reallocation, *Quantitative Economics* **5**(2): 377–416.

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.

Hopenhayn, H. A. (2014). Firms, misallocation, and aggregate productivity: A review, *The Annual Review of Economics* **6**(1): 735–770.

Hsieh, C.-T. and Klenow, P. J. (2009). Misallocation and manufacturing tfp in china and india, *The Quarterly journal of economics* **124**(4): 1403–1448.

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* **73**(3): 1345–1368.

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.

Peters, M. (2020). Heterogeneous markups, growth, and endogenous misallocation, *Econometrica* **88**(5): 2037–2073.

Restuccia, D. and Rogerson, R. (2008). Policy distortions and aggregate productivity with heterogeneous establishments, *Review of Economic Dynamics* **11**(4): 707–720.

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.

Syverson, C. (2019). Macroeconomics and market power: Facts, potential explanations, and open questions, *Journal of Economic Perspectives* **33**(3): 23–43.

Varian, H. R. (1992). *Microeconomic Analysis*, Vol. 3, Norton New York.

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.