The Great Dispersion in Energy Productivity Across Firms

Josh De Lyon[1,2] and Antoine Dechezleprêtre[1,2]

[1] OECD Directorate for Science, Technology and Innovation

[2] Centre for Economic Performance, London School of Economics and Political Science

Abstract

This paper explores the differences between firms' energy productivity – value added per joule of energy – and its relationship with economic performance based on representative data from Chile, Croatia, France, Indonesia, Lithuania, Netherlands, Portugal, and Sweden. It documents wide dispersion in energy productivity across firms within narrowly defined industries, which significantly exceeds differences in labour productivity. On average, the 90th percentile firm is 21 times more energy productive than the 10th percentile firm in the same industry, compared with five times for labour productivity. Even comparing firms producing the same 8-digit product or of similar size, there is huge dispersion in energy productivity. There is no sign that this dispersion is narrowing down over time. Addressing this dispersion could have huge implications: raising the energy productivity of the least productive firms to the energy productivity of the 25th percentile firm in their industry would reduce aggregate energy and emissions by almost half for the same level of output. Machine learning prediction models and linear regression show that the most important factor in explaining within-industry differences between firms' energy productivity is labour productivity, suggesting a link between economic and energy efficiency. Carbon pricing, stronger competition, and access to finance also explain dispersion, suggesting a role for policies.

Acknowledgements

This project has benefited from helpful comments from Chiara Criscuolo, Guy Lalanne and Francesco Manaresi. Gabriela Garcia Sandoval and Xavier Koch provided excellent research assistance and Isabelle Desnoyers-James provided invaluable help with data collection and meta data. The project benefited greatly from work of previous OECD micro-distributed data projects, and especially from the help of Rudy Verlhac and Martin Reinhard. The authors are grateful to Ulrika Stavlöt, Michael Polder, Joram Vuik, the Croatian Bureau of Statistics, the Chilean Institute of National Statistics, and to Statistics Lithuania for running the module on confidential micro data or for helpful discussions relating to data. The authors are also grateful to Delegates of the OECD Committee on Industry, Innovation, and Entrepreneurship (CIIE) and Working Party on Industry Analysis (WPIA), especially discussions by Paul Cusson, Danny Leung, Mika Maliranta, and Michael Polder. The project benefited from helpful comments and support with data from Dan Andrews, Jack Blundell, Arlan Brucal, Miguel Cardenas Rodriguez, Steven Cassimon, Thomas Chalaux, Hélder Costa, Hélène Dernis, Luis Diaz, and Florian Mante, as well as seminar participants at LSE Grantham Research Centre, Max Planck Institute, NBER, OECD and University of Gothenburg. Access to some confidential data on France used in this work has been made possible within a secure environment offered by CASD – Centre d'accès sécurisé aux données (Ref. 10.34724/CASD). This project is partly financed by the European Commission under Work Programme Horizon Europe – 2021-2022 Horizontal support expenditure for Horizon Europe and Euratom Programmes under grant no. LC-01965484. The authors declare to have no conflict of interest to disclose regarding the research on this paper.

1 Introduction

Energy represents a relatively modest share of firms' total costs – around 5% on average – but is a crucial input for industrial production. The recent surge in energy prices – which tripled from 2020 to 2022 and remain at double their 2020 level in 2024 – has underscored the vulnerability of firms to energy costs, as highlighted by the number of measures put in place by governments to shield companies from the impact of rising prices. Improving energy productivity – the amount of output produced per unit of energy consumed – has therefore emerged as a key lever for strengthening firms' competitiveness while simultaneously reducing their exposure to energy price shocks.

Improving energy productivity is also crucial to achieve the net zero transition while maintaining economic growth. If not accompanied by higher energy productivity, higher labour productivity – a key policy objective to improve firms' competitiveness – would raises emissions. How to accelerate labour productivity and energy productivity growth at the same time therefore has important policy implications. The industrial sector, which accounts for 23% of greenhouse gas emissions in the OECD (Aulie et al., 2023[1]), has become more carbon efficient in recent years: in 2018, the ratio of real output to CO_2 emissions was 56% higher than in 2000 on average across the manufacturing sector of OECD Member countries. This pattern holds even after accounting for carbon embodied in intermediate inputs and sectoral structural change.[1]

Policy has played a role in this clean-up of manufacturing. For example, there is robust evidence suggesting that the European Union (EU) Emissions Trading System (ETS) reduced the emissions and increased energy and carbon productivity of regulated firms without affecting their competitiveness.[2] Whether these results will hold under much higher carbon prices is unknown, and understanding the factors that influence firms' energy and labour productivity is crucial to implement policies that will ensure the transition to the low-carbon economy is compatible with strong economic growth.

This paper provides an extensive empirical analysis of firms' energy productivity – defined as value added per unit of energy – based on representative firm-level data from Chile, Croatia, France, Indonesia, Lithuania, Netherlands, Portugal, and Sweden. It documents the dispersion in energy productivity across firms, explores the potential causes and consequences of this dispersion and analyses the relationship between energy productivity and measures of firm performance. Building on the OECD MultiProd project (Berlingieri et al., 2017[5]), typical measures of economic performance such as labour productivity are computed alongside energy productivity (constructed thanks to detailed data on energy use) and carbon productivity (calculated as value added per tonne of energy-related CO_2). Thorough procedures of data cleaning, harmonisation, and checks are developed to ensure that data are comparable across countries, with the final datasets covering 1995 to 2021, depending on the country.

A recent literature has documented huge differences in efficiency across firms, especially in terms of their labour and multi-factor productivity (see for example Andrews, Criscuolo and Gal (2019[2])), which result in a drag on aggregate productivity growth. Yet, while there is extensive evidence on the extent and

[1] Authors' calculations using OECD data (Yamano and Guilhoto, 2020[42]). Also see Shapiro and Walker (2018[37]) and Levinson (2015[38]) for a discussion of the manufacturing cleanup in the United States.

[2] See for example Dechezleprêtre, Nachtigal, and Venmans (2023[39]), Colmer et al. (2020[40]), or Petrick and Wagner (2014[41]).

determinants of the differences between firms' economic performance, much less is known about differences in their environmental performance. A handful of papers suggest there is also significant dispersion across firms in their environmental performance (Wagner et al. (2020[3]); Lyubich, Shapiro, and Walker (2018[4])), but they do not investigate the potential benefits of addressing this dispersion, how to improve energy productivity across the distribution of firms, or the potential relationship and implications for economic performance. It could, for instance, be related to similar factors as dispersion in labour productivity, as well as under-pricing of carbon externalities and energy comprising a relatively small share of total costs. Understanding this firm heterogeneity in environmental performance is key to help design policies to improve the efficiency of firms across the board, from the most efficient firms to the least efficient.

The data reveal huge dispersion in energy productivity across firms within narrowly defined industries. On average, the 90th percentile firm is 21 times more energy productive than the 10th percentile firm. There is far more dispersion in energy productivity than in labour productivity in all countries included in the sample: on average, the 90th percentile firm is five times more labour productive than the 10th percentile firm. Therefore, there is four times more dispersion in energy productivity than in labour productivity.

The dispersion in energy productivity is not due to firms in the same industry producing different products. Using data on firms' sales by 8-digit products for France, there is still large dispersion in energy productivity when comparing firms producing the same specific products, and this dispersion is still magnitudes greater than dispersion in labour productivity. Comparing firms producing the same 8-digit product, the extent of dispersion is almost identical to dispersion within 4-digit industries computed on the same sample. The dispersion also persists, and is even magnitudes greater, when measured as quantity produced per energy instead of value added per energy.

There is no evidence that the rate of catch-up for low energy productivity firms has improved over time. Regression analysis relating within-firm growth in energy productivity to their distance from the frontier shows that there has been no improvement, or even a slowdown, in the rate of catch up over time in almost all countries in the sample. Furthermore, the average distance between the 10th and 90th percentiles of energy productivity has been stable or even increasing in most countries in the sample.

What are the most important firms' characteristics predicting their energy productivity and what are the directions of these relationships with energy productivity? To answer the first question, a machine learning prediction model using gradient boosting methods is estimated, followed by ex-post analysis of the model predictions (using SHAP analysis). To answer the second question, cross-sectional firm-level regressions are estimated. These relationships are examined comparing firms in the same industry, country, and year. The results show that the most important factor in predicting differences in firms' energy productivity is labour productivity, conditional on other factors. The relationship is not simply mechanical: when instead investigating which factors are most important in predicting labour productivity, average wage is the most important. The relationship also holds looking at within-firm changes over time, controlling for growth and initial characteristics. This relationship between energy and labour productivity suggests that there may be synergies between improving aggregate labour productivity – crucial for achieving growth and wage increases – while also reducing energy use and carbon emissions from the industrial sector. Another important finding is that larger firms tend to be *less* energy productive, in stark contrast to the wealth of theory and evidence documenting the positive relationship between firm size and economic productivity (a finding that is also confirmed in this data). Firms with higher energy productivity also tend to be less capital intensive and have newer vintages of capital.

Addressing dispersion in energy productivity could have important aggregate implications. A counterfactual exercise is conducted, whereby the energy productivity of the least productive firms is improved to at least that of the 25th percentile firm in their industry. By improving the energy productivity of the least productive firms to the 25th percentile level, energy use and carbon emissions could be reduced by nearly half to achieve the same total level of output, aggregating across the eight countries. These results are large due to the empirical observation that less energy productive firms tend to be larger and the unproductive firms

are very far behind the others. In addition to this, regression analysis shows that industries with more dispersion in energy productivity tend to have worse aggregate energy productivity, and that as industries become more dispersed, they also tend to become relatively more energy intensive. Together, these results highlight the potential gains from improving the performance of lagging firms and from improving allocative efficiency.

The characteristics of industries that are more dispersed in energy productivity are examined using both the machine learning prediction model and linear regression. There is a strong relationship between dispersion in energy productivity and dispersion in labour productivity: industries with greater dispersion in labour productivity are also more dispersed in energy productivity, and this also holds in changes over time. It is also the most important factor in predicting dispersion in energy productivity. This suggests that the factors that have been proposed to explain dispersion in labour productivity, such as a lack of competitive pressure, limited technology diffusion, or barriers to entry and scale-up, may also explain dispersion in energy productivity. In line with this, there is a positive relationship between dispersion in energy productivity and the concentration of employment among firms – a proxy of a lack of competition. Similarly, there is a strong negative correlation between restrictiveness to foreign direct investment (FDI) and dispersion in energy productivity. The results also show that the average energy price faced by a sector is amongst the most important factors in explaining dispersion: increases in energy prices are associated with a relative decline in dispersion. This suggests that carbon pricing policies covering all firms across a sector – rather than a subset of firms, as is the case in the EU ETS for example – may be effective in incentivising unproductive firms to adopt more energy efficient production processes. Finally, industries with more cashflow relative to their capital stock are relatively less dispersed, suggesting that providing access to finance to facilitate technological upgrading is important to promote catch up.

This evidence suggests that there is a significant untapped potential of boosting competitiveness, improving environmental performance and reducing industrial emissions by addressing the energy productivity of lagging firms. Furthermore, the strong positive relationships between economic and environmental performance of firms suggest that these improvements will not necessarily come at the expense of economic performance.

These findings have important implications for policy. Policies should incentivise catch up in energy productivity of the least productive firms. This includes carbon pricing schemes that do not disproportionately favour larger and less efficient producers but also goes well beyond policies relating to energy pricing. Policies that improve technology upgrading and promote allocative efficiency – such as those relating to business dynamism, competition, and access to finance – can also help to improve aggregate energy productivity. These types of policies are often proposed to boost labour productivity and growth. Therefore, structural policies to improve technology diffusion between firms and promote healthy competition could benefit economic performance whilst simultaneously reducing firms' exposure to energy shocks and helping to achieve the net zero transition.

This paper contributes to two branches of the literature. First, there is a small literature documenting the extent of heterogeneity between firms in their environmental performance. Most closely related to the present paper is Wagner et al. (2020[3]) who document the distributions of CO_2 per worker for seven countries using representative data for manufacturing sector plants. They show that there is high dispersion in CO_2 per job within 2-digit industries to a similar degree across countries, but that this dispersion has declined over time. Also in a cross-country setting, Capelle et al. (2023[6]) document extensive within-industry dispersion in environmental performance using self-reported data from 4 000 large, listed firms. They explain these differences across firms with a model whereby firms choose different vintages of capital and how much to spend on R&D. Bijnens and Swartenbroekx (2024[7]) use data on 2 479 companies to show that improving the efficiency of just a small number of large unproductive firms could reduce emissions by 38%.

Looking at individual countries, Lyubich, Shapiro, and Walker (2018[4]) document the extent of heterogeneity in energy and carbon productivity across firms in the United States, showing that it is significantly greater than heterogeneity in labour, capital, and multi-factor productivities. Similar evidence is documented for Estonia (Criscuolo et al., 2024[8]) and Germany (Von Graevenitz and Rottner (2020[9]) and Petrick, Rehdanz, and Wagner (2010[10])). Klenow, Pastén, and Ruane (2024[11]) document that there are differences in fossil fuel intensity across firms in Chile, and that this is negatively correlated with their total-factor productivity.

This paper contributes to this literature by providing new detailed evidence on the extent of heterogeneity in energy and carbon productivity across firms within industries in multiple countries. It documents extensive dispersion in both energy and carbon productivity even within very narrowly defined 4-digit industries and within 8-digit product markets in France. It shows the evolution of dispersion over time, and detailed measures of dispersion by industry. It shows the extensive potential gains that could be made by addressing this dispersion. The characteristics of industries with more dispersion are also investigated, providing indicative evidence on the potential frictions related to dispersion.

Second, a broader literature documents the extent of firm heterogeneity in other metrics of performance, mainly productivity. Berlingieri, Blanchenay, and Criscuolo (2017[12]) use representative cross-country data to document large and increasing dispersion in labour and multi-factor productivity. Andrews, Criscuolo, and Gal (2019[2]) show that the increasing divergence is related to weakening catch-up and market selection below the frontier, while Berlingieri et al. (2020[13]) document the characteristics of the less productive laggard firms. Canonical papers on this topic include Syverson (2011[14]) on the drivers of productivity and Hsieh and Klenow (2009[15]) for the implications of extensive misallocation of resources between productive and unproductive firms.

This paper contributes to the literature on heterogeneity between firms by providing further evidence on the huge differences between firms even within narrowly defined industries, namely their differences in energy productivity. It shows that labour productivity is the most important factor in predicting firms' energy productivity. Furthermore, it shows that dispersion in labour productivity is correlated with dispersion in energy productivity, both in levels and changes, and controlling for other industry characteristics. It provides evidence that factors relating to competition and access to cashflow for investments are related to dispersion in energy productivity, as has been discussed for labour productivity. Overall, the evidence suggests that addressing these factors may help to achieve the net zero transition alongside strong economic performance.

The paper proceeds as follows. Section 2 provides detail on the datasets used and the procedures for harmonising micro data across countries. In Section 3, the extent of dispersion in energy productivity is documented. The characteristics of firms with higher energy productivity are explored in Section 4. Section 5 demonstrates the potential implications from addressing dispersion, while Section 6 examines the factors explaining with this dispersion, including some relating to policy. Section 7 concludes and provides a discussion of policy implications.

2 Data

The project relies on firm-level data for multiple countries. As this data is typically confidential and therefore only accessible under strict conditions, a key contribution of the project is to develop a flexible set of codes that can be run on the firm-level micro data to collect both energy use and economic information. These codes clean and harmonise the micro data, conduct analysis at the firm-level, and produce industry-year level datasets that contain within-industry moments of the distribution as well as aggregates on which further analysis can be conducted.

Data on eight countries are used in the analysis: Chile, Croatia, France, Indonesia, Lithuania, Netherlands, Portugal, and Sweden. The data are obtained from a combination of production surveys and surveys on energy use and cover the period 1995-2021, depending on the country. All the sources used are either the population of firms or a representative sample of industry aggregates. The baseline sample for the analysis focuses on the manufacturing sector and on firms with at least 20 employees to harmonise the samples across countries. Measures of dispersion are computed on country-industry-year cells with at least 10 firms.

The two key variables used throughout the analysis are energy and labour productivity. Labour productivity is defined as value added per worker while energy productivity is defined as value added per unit of energy. A third variable – carbon productivity – is also used for certain analyses and is defined as value added per tonne of energy-related carbon dioxide (CO_2) emitted through the use of electricity and fuels in the production process. Carbon productivity does not include the amount of emissions embodied in intermediate inputs, just as the labour and energy embodied in intermediate inputs are also not included in the measures of labour productivity and energy productivity, respectively.[3]

This section is divided into three parts. First, the sources of raw data for each of the eight countries are discussed. Second, details on the data collection process are provided, including the process used to estimate energy and CO_2 from different types of data on energy use, and the procedures to harmonise the data across countries. Third, summary statistics are provided for the firm-level datasets and for the industry-level distributive micro data.

Raw data sources

Key information about each country's raw data is provided in Table 2.1. It first shows the period covered by the data and the type of data on energy use that is available – either quantity of fuel or electricity for each type or just the total spending on energy. Then, it provides information on the samples for both the energy data and production data, which may be overlapping when the energy use data comes from the

[3] A potentially important margin of adjustment, whereby firms outsource emissions to improve their own energy and carbon performance, is therefore omitted. Where applicable, the value of intermediate inputs or intermediate input intensity is controlled for. A full analysis of the role of outsourcing emissions is left for future work.

same source as the economic balance sheet variables. Finally, it provides the units of the raw data. Annex A contains more details on the sources and data for each country.[4]

Table 2.1. Summary of firm-level data sources

Country	Years covered	Type of energy use data	Energy use data sample	Production data sample	Unit level
Chile	1995-2015	Quantity by type of energy	All establishments with at least 10 employees	All establishments with at least 10 employees	Establishment, converted to firm level
Croatia	2002-2019	Total spending on energy	All firms	All firms	Firm
France	2005-2020	Quantity by type of energy	Stratified sample for establishments between 20 and 250 employees with survey weights. All establishments above 250 employees	All firms	Establishment, converted to firm level
Indonesia	2000-2019	Quantity by type of energy	All establishments with at least 20 employees	All establishments with at least 20 employees	Establishment
Lithuania	2004-2020	Total spending on energy	Sample - weights constructed using business register	All firms	Firm
Netherlands	2015-2021	Total spending on energy	Stratified sample with survey weights	Stratified sample with survey weights (same source as energy variable)	Firm
Portugal	2004-2020, exc. 2012-2014	Quantity by type of energy	Stratified sample with survey weights	All firms	Firm
Sweden	2005-2021	Quantity by type of energy	All firms with at least 10 employees	All firms with at least 10 employees	Firm

Note: Sample coverage refers to firms in the manufacturing sector only. The final sample used in the analysis is manufacturing firms with at least 20 employees.
Source: OECD

[4] In Chile, the industry code is provided at the 3 digit level. Throughout the paper, in any analysis at the 4 digit level, Chile was conducted at the 3 digit level.

Data collection procedure

The data collection procedure is flexibly designed to utilise confidential data on firms' environmental and production activities from varying types of sources across countries. This project builds on similar procedures that have been developed at OECD to collect data on productivity (Berlingieri et al., 2017[5]), business dynamics (Calvino, Criscuolo and Verlhac, 2020[16]), AI adoption (Calvino and Fontanelli, 2023[17]), R&D (Appelt et al., 2022[18]), and employment reallocation (Calligaris et al., 2023[19]). Data is provided at the firm-year or establishment-year level, before a set of harmonised cleaning, estimation, and aggregation procedures are conducted.

Multiple data sources may be used within each country to obtain different variables or to help with harmonisation. Three main types of data can be included in the module: an energy use survey; a production survey; and a business register.[5] The data preparation begins by conducting basic checks and cleaning on each data source provided.[6] The data is then processed in numerous steps. First, it computes the energy equivalent of each fuel in joules and estimates the associated emissions (discussed in detail below below). Second, if the environmental data is at the facility or establishment level, it converts it to the firm-level, scaling appropriately, such that all final datasets are at the firm-level.[7] This process also includes a procedure to convert survey weights at the plant level to the firm-level while maintaining aggregate representativeness. Third, all data sources are combined and variables are deflated. Fourth, suspicious outliers – most likely to be data entry errors - are detected using a new methodology that combines comparisons of within-firm variation over time with contemporaneous differences in key variables within the firm. Particular attention is made to avoid over cleaning of genuine frontier and laggard firms. Remaining severe outliers are detected using cross-sectional variation, which is especially useful in countries where panel data is not available, again using ratios of variables within the firm as well levels to minimise dropping genuine outliers.

In addition to being able to conduct analysis directly on the firm-level data for each country, the data is also aggregated to the industry-year level, taking moments of the firm distribution within each cell. These industry-level datasets can be analysed in a cross-country setting.[8] In Annex A, details are provided on the method to convert industry codes, external data sources used to compute emissions, conversion of establishment to firm-level data (used for France only), detection and treatment of extreme outliers (used to try to remove data entry errors), and confidentiality requirements. Figure A A.5 shows the data collection procedure diagrammatically.

[5] Note that for each country it is possible to match every firm with energy data to its production data, either because they are obtained from the same source (Chile, Croatia, Lithuania, Indonesia, Netherlands) or because the production data is available for the census of firms (France, Portugal, and Sweden). Importantly, this means that there is no need to recompute sampling weights to maintain representativeness.

[6] These include, for example, confirming the monetary units and basic properties of the data, checking for negative values in variables that should be non-negative, and checking for missing values. Industry codes are converted to ISIC Revision 4 and any changes in the classification system throughout the period are converted using an internally-constructed crosswalk that exploits the panel data structure of the data.

[7] In Indonesia, no firm identifiers are available, meaning that analysis must be carried out at the establishment level.

[8] Confidentiality checks are applied to the aggregated data before returning the data to the OECD. Throughout the process, diagnostic statistics, notes, and detailed meta data are collected to ensure that the procedure is running correctly and there are no major issues with the process.

Estimating energy use and CO_2 emissions

Energy use and the associated emissions can be computed from different variables, depending on data availability. Accurately and consistently identifying the quantity of energy use and its associated emissions from differing types of data across countries is crucial to this project. The procedures developed in this project help to overcome potential biases that may arise due to, for example, firm characteristics such as size being correlated with the price paid for energy.

The most accurate data from which energy and production emissions can be computed is the quantity of energy used by energy type (electricity, gas, coal, etc). In this case, energy equivalents and emissions factors for each energy type, collected from the International Energy Agency (IEA), are used to obtain energy use in joules and the volume of CO_2 emitted, as well as other greenhouse gases which are converted into CO_2 equivalents and referred to in their totality as carbon throughout for simplicity.[9] Quantity data by energy type is available for Chile, France, Indonesia, Portugal, and Sweden.

In contrast, if only the total spending on all energy is available, publicly available data on energy type usage by industry and energy prices, collected from IEA and Eurostat, must be used alongside the emissions factors to compute estimates of production emissions at the firm-level. These data are available for Croatia, Lithuania, and Netherlands.

The following sub-sections briefly outline the procedure to estimate energy use and production emissions under each type of data availability, and discuss the assumptions required in each case. A comparison of the emissions estimated from total energy spending and from energy use by type is provided using data for France, where both are available. Annex A provides details on the external datasets used to compute emissions as well as the cleaning and imputation procedures used to prepare them. It also provides details on how a continuous relationship for prices as a function of quantity consumed is estimated from available data, including evidence on the accuracy of this procedure.

Estimating energy use and CO_2 emissions with energy use by fuel type data

When data on quantity of energy by type is available, it is straightforward to compute energy use in joules and tonnes of emissions. Quantities are converted to tera joules (TJ) using energy densities, then emissions factors from TJ to tonnes of CO_2 equivalent are used. This is done for each energy type, then total emission is simply the sum of emissions from each energy type.

Estimating energy use and CO_2 emissions when only total energy spending is available

If only total energy spending is available, energy and emissions can be estimated under certain assumptions using external datasets. The procedure takes three main steps: 1) use the industry of the firm to match with energy use by type of fuel, available at the country-industry-year level; 2) match energy types with their prices (which can vary by quantity consumed by the firm) in each country-year; 3) combine firm-level spending with the external data sources to compute estimates of energy and emissions.

An important empirical observation is that energy price is strongly negatively correlated with energy quantity at the firm level. That is, as firms consume more energy, they tend to pay a lower price per joule of energy. It is therefore important to account for this when estimating energy use and emissions from just total energy spending. This is done by developing an algorithm that numerically solves for firms' quantity of each type of energy purchased, using information on their total size and estimated parameters on the price of each energy type as a function of quantity consumed. The procedure is described in detail in Annex

[9] Methane and nitrous oxide emissions are also accounted for and are converted into CO_2 equivalents – throughout the paper, "carbon emissions" will be used to refer to total emissions of carbon dioxide, nitrous oxide, and methane. Process emissions are not available.

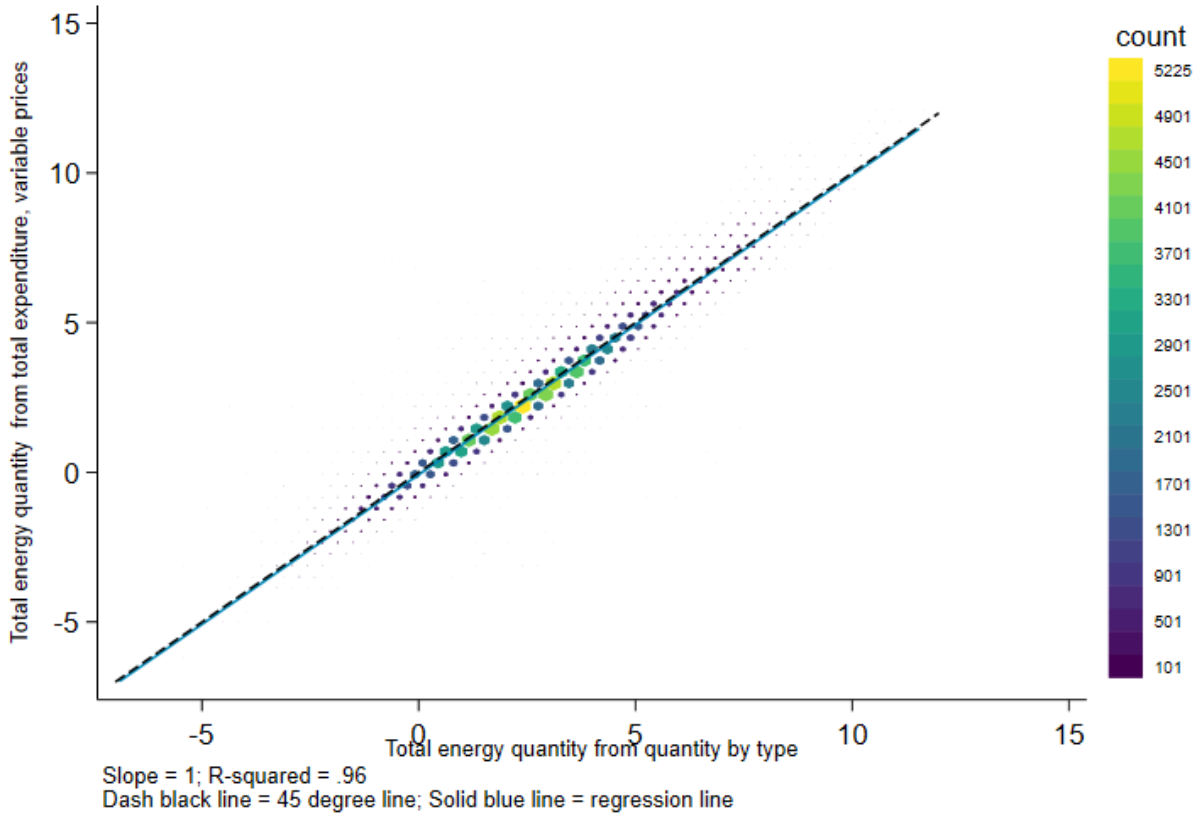
A and is used as the baseline estimate of energy and CO_2 when only total energy spending is available. Tests suggest that this procedure - accounting for different prices paid by firms - significantly improves measurement of energy use and removes the systematic bias that would be introduced by simply assuming fixed prices.

Robustness of estimating energy use and emissions from total energy spending

Using French data, which contains both total energy spending and quantity used of different fuels, it is possible to test the accuracy of the procedures to estimate energy use and emissions from only data on total energy spending. Figure 2.1 shows that there is a close relationship between the log of estimated quantity from total energy spending (y axis) and the log of actual spending (x axis) in the French data, where data on energy quantity and energy spending by type are available. The slope is 1.00, exactly on the 45-degree line, and the data points are close to the estimated line, with an R-squared of 0.96. It is important to account for energy prices that vary with quantity purchased: when instead only using fixed prices, the slope of the relationship is 0.89, reflecting that larger consumers pay lower prices for energy. The relationship for estimated versus actual emissions, shown in Figure A A.2, is close but slightly less accurate than for quantity because there is more variation across energy types in emissions per dollar compared with joules per dollar. Therefore, when estimating emissions from total spending (in Croatia, Lithuania, and Netherlands), any errors in estimating which energy types are used by firms matter more than when estimating energy from quantity.

Figure 2.1. Energy quantity is similar to true quantity when estimated using total energy spending

Log quantity of energy from total energy spending against log quantity of energy from data on energy by type, establishment level data for France



Slope = 1; R-squared = .96
Dash black line = 45 degree line; Solid blue line = regression line

Note: The chart plots the log of total energy quantity estimated using only total expenditure on energy (y axis) against the total energy quantity estimated from detailed information on quantities used of different fuel types and electricity (x axis). The dashed black line is the 45-degree line while the blue line is the fitted OLS regression estimate from y on x. The heat map shows the underlying data points where the size and colour of the areas represent the number of underlying data points. The data is establishment level information from France pooled over 2005 to 2020. Source: authors' calculations.

Summary statistics

Firm summary statistics

It is important to examine summary statistics on the underlying firm-level data for each country used in the project. These data underpin the industry-level micro-distributed datasets and are used in the firm-level analysis such as the regressions in Section 4. In particular, given that the project uses some micro data that cannot be directly accessed by the researchers, the summary statistics provide another important check on the consistency of the data within and across countries.

The tables, included in Annex B, show the firm-level summary statistics for Chile (Table A B.1), Croatia (Table A B.2), France (Table A B.3), Indonesia (Table A B.4), Lithuania (Table A B.5), Netherlands (Table A B.6), Portugal (Table A B.7), Sweden (Table A B.8). Reassuringly, the main key variables exhibit broadly similar distributions across countries, with some differences as expected. For example, the average share of labour in total costs (across all sectors) is similar across the European countries while and higher in Indonesia. Average labour productivity is highest in France, Netherlands, and Sweden – the countries with the highest GDP per capita of the countries included in the sample. The firm-level summary statistics are also indicative of the cross-country differences in the extent of dispersion in energy productivity across firms. The 90th percentile firm is significantly more productive than the 10th percentile firm.

Industry summary statistics

Table A B.9 presents summary statistics on the distributed micro data at the country-industry-year (cell) level. The main sample used throughout the analysis is restricted to firms with at least 20 employees in 3-digit manufacturing industries. The within-cell percentiles and dispersion measures are only computed in cells with at least 10 firms; hence the smaller cell counts in these cells. Comparing the first and fourth rows of the table shows that there are 8 106 cells with at least one firm compared with 5 166 cells with at least 10 firms. Some cells must also be blanked for confidentiality reasons, which can be variable-specific, so cell counts vary slightly across variables. The mean, standard deviation, minimum, maximum, percentiles, and count of cells are provided for each variable. Note that the columns provide different statistics *across* industry-country-years (e.g. the median industry-country-year has value added of 171 million), while some variables are specific moments of the distribution *within*-industry-country-years (e.g. the mean of the 10th percentile firm's log value added per energy within an industry-country-year is 11.61).

Industries vary in terms of their overall size, as measured by the count of firms, total employment, and total gross output. The distributed micro data approach provides cross-country data on within-industry moments of the firm distribution, providing information on dispersion of firms conducting similar production activities. The variables beginning with [p10], [p50] and [p90] present the within-industry moments for both log value added per energy (energy productivity) and log value added per worker (labour productivity).

3 Dispersion in energy productivity

This section explores the patterns of within-industry dispersion in energy productivity. It first documents the great differences between firms in their energy productivity and the robustness of this pattern. It then shows that this dispersion has persisted over time and that it is present across many industries.

Dispersion in energy productivity between firms

There is huge within-industry dispersion in energy productivity in all countries examined, and much more than dispersion in labour productivity. On average across industries and countries, the 90th percentile firm is 21 times more energy productive than the 10th percentile firm in the same industry.[10] These differences are even greater for carbon productivity, where the 90th percentile firm is 26 times more productive than the 10th percentile firm.

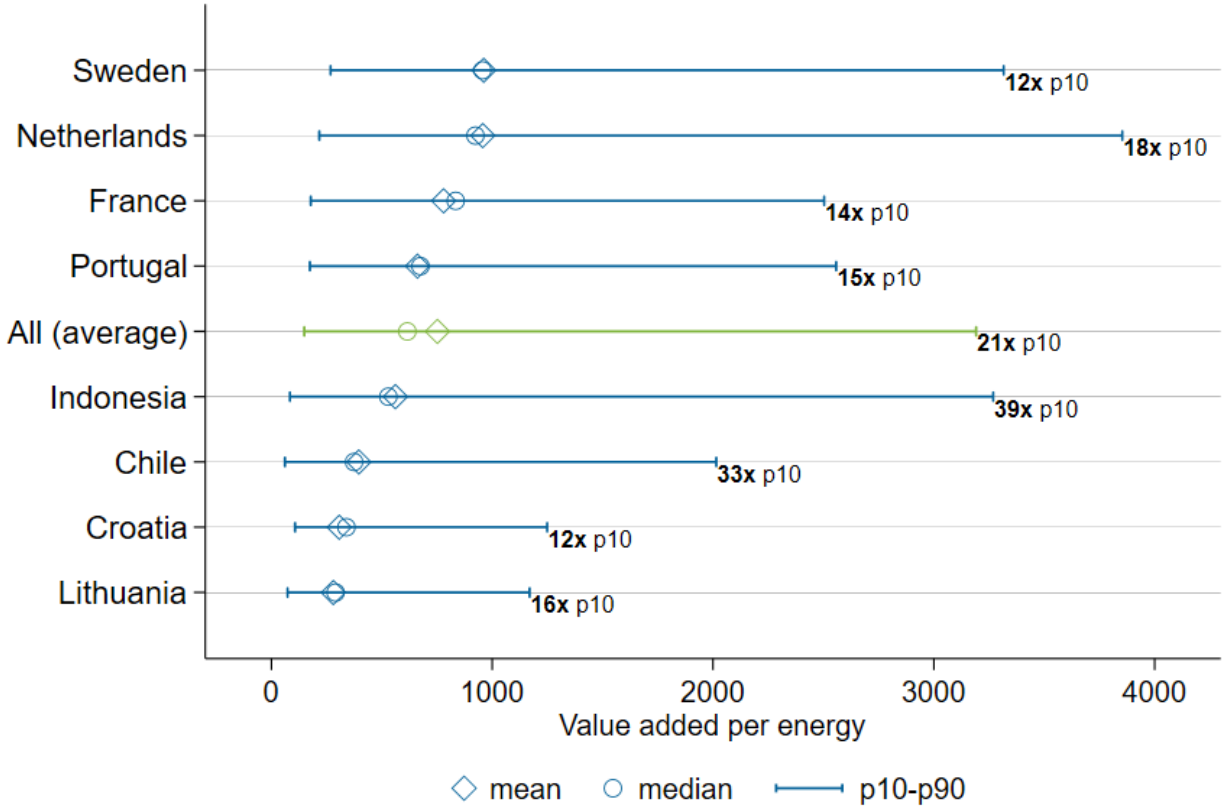
Figure 3.1 shows the extent of dispersion in energy productivity in each country. It plots the average of the 10th and 90th percentiles across industries and the ratio between these values, as well as the mean and the median. Average energy productivity is highest in Sweden and lowest in Lithuania, while there is the largest dispersion in Indonesia. Figure A B.1 in Annex B show the distributions of carbon productivity, where there is even more dispersion than in energy productivity on average, with the additional dispersion generated by differences in the types of fuels used by firms. While Indonesia is the fourth least energy productive country on average, it has the lowest average carbon productivity, because Indonesian firms use dirtier fuels than Chilean, Croatian, Lithuanian firms. Figure A B.2 shows that there is significantly less dispersion in labour productivity compared with energy and carbon productivity across all countries in the sample. For labour productivity the equivalent difference is five times.[11] Therefore, on average, there is four times more dispersion in energy productivity than for labour productivity.

[10] Lyubich, Shapiro and Walker (2018[4]) show that, in the US, the 90th percentile firm is 7 times more energy productive than the 10th percentile firm in the US. This is less dispersed than Sweden (12 times), the least dispersed country in this sample. This lower dispersion in the US relative to other countries may reflect structural economic and environmental differences, or could be that Lyubich et al. are able to use a slightly narrower industry classification in the single-country setting.

[11] These statistics are measured by taking the average of log dispersion across industries and then computing the implied average difference in productivities. This use of logs is preferred to reduce the influence of large outliers in dispersion in the industry level data. The same methodology is used in Lyubich, Shapiro and Walker (2018[4]).

Figure 3.1. Sweden is the most energy productive country while Indonesia is the most dispersed

Percentiles and dispersion in energy productivity across countries



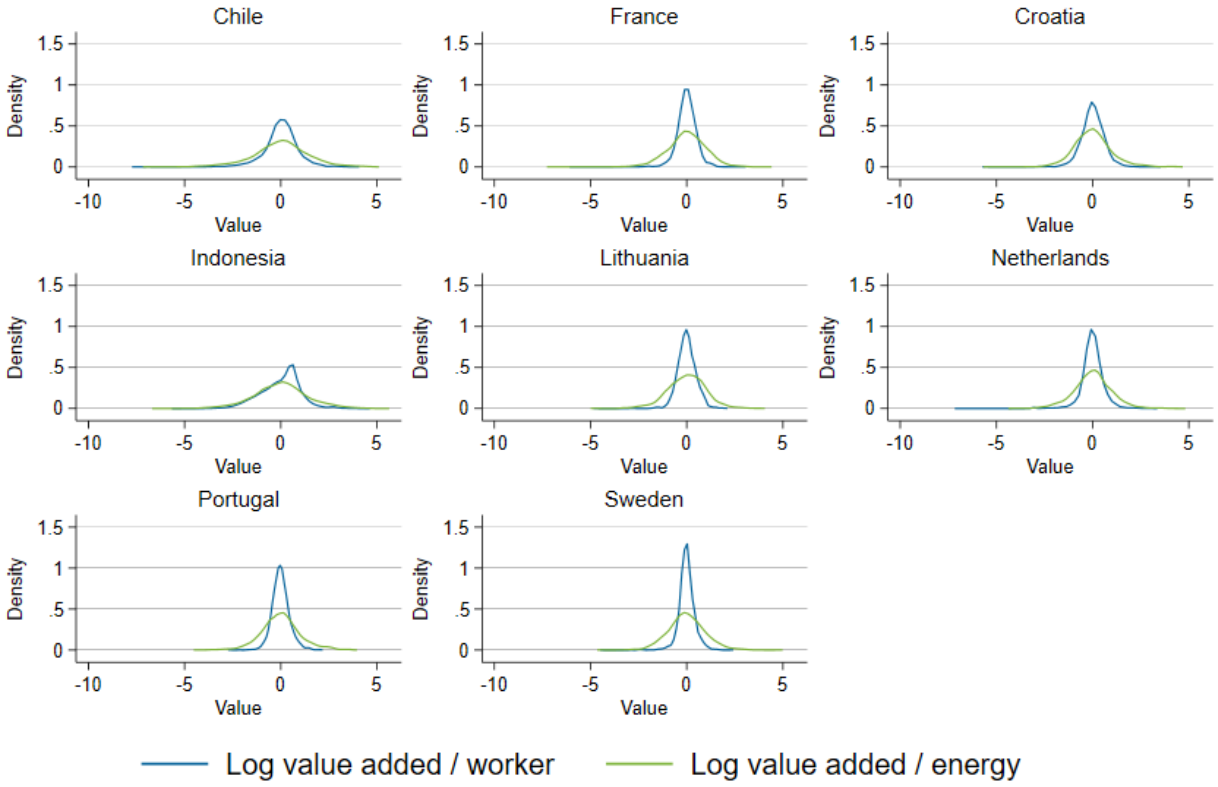
Note: The chart plots the mean and 10th, 50th, and 90th percentiles of energy productivity for each country, averaged across industries, in the latest year of available data for each country up to 2019. The unweighted average of the log values are taken across industries to minimise the influence of outliers, then the exponents are taken. The labels show the ratio between the 90th percentile and 10th percentile firm. The units are thousands of dollars per tera-joule of energy (or dollars per mega joule). The sample is restricted to manufacturing industries with at least 10 firms and firms with at least 20 employees.
Source: authors' calculations.

Figure 3.2 clearly shows that there is more dispersion in energy productivity than in labour productivity in all countries. It plots the within-industry distributions of firms' energy and labour productivity within 4-digit industries for each of the eight countries in the latest year of data for each country (up to 2019, before the effects of Covid-19), with each distribution centred at 0.[12]

[12] Note that it is possible to use 4-digit industries for the density plots as they can be computed at the firm level and produced for each country separately, i.e. they do not require the use of cross-country data.

Figure 3.2. There is more dispersion in energy productivity than labour productivity

Distributions of energy productivity and labour productivity across firms within 4-digit industries



Note: The charts plot the within-industry-year distributions of energy and labour productivity across firms. Data is plotted for the latest year available in each country up to 2019. Industries are defined at the 4-digit level and distributions are centred at 0. The sample is restricted to manufacturing industries with at least 10 firms and firms with at least 20 employees.
Source: authors' calculations.

One possible explanation for this dispersion could be that firms in the same industry are producing different products. To investigate this, detailed data on firms' sales by product for French firms is obtained from the Enquête annuelle de production (EAP), covering 2009 to 2020. This is matched with firm-level measures of energy and labour productivity. For this exercise, firms are assigned to the product in which they have the highest share of sales. As robustness checks, firms with their main product accounting for less than 50% or 80% of their total sales are excluded to ensure that firms producing the same products are being compared. Measures of dispersion – the log difference between the 90th and 10th percentile firm – are then computed within a product or industry at different levels of granularity. Product or industry cells are only included in the sample when they have at least 10 firms, which is required to compute dispersion. This means that, for product concentration measures where narrowly-defined product markets sometimes have few firms, some cells are dropped. For comparability, the main measures are computed at the 4 digit industry and 8 digit product level for the same sample. The 8 digit product markets are very narrowly defined: examples of product markets include "Cartons, boxes and cases, of corrugated paper or paperboard" and "Plastic doors, windows and their frames and thresholds for doors".

The results show that, even comparing firms producing the same 8 digit product, there is huge dispersion in energy productivity. Table 3.1 shows the average dispersion (across years) in energy and labour

productivity within 2, 3, and 4 digit industries, and within 6 and 8 digit products.[13] The table shows the ratio between the 90th and 10th percentile firms by taking the exponent of the average log difference. As expected, when the industry or product is more narrowly defined, the extent of dispersion decreases slightly. When fixing the sample to include the same firms, the 90th percentile firm is 9.4 times more energy productive than the 10th percentile firm in 4 digit industries, compared with 8.9 times in 8 digit products. In other words, even when comparing firms producing the same 8 digit product in the same country (France), there are still huge differences between firms' energy productivity and this difference is far greater than the dispersion in labour productivity, for which the 90th percentile firm is only 2.8 times more productive than the 10th percentile firm. To provide a specific example, Figure A B.3 plots the distribution of carbon, energy, and labour productivity for firms in France in a single year producing only "Plastic doors, windows and their frames and thresholds for doors". It clearly illustrates the extent of dispersion between firms in their productivity even among firms, and that there is more dispersion in energy and carbon productivity than in labour productivity.

Another possible explanation of the extensive dispersion in energy productivity could be differences in firm size. Rows 4 and 5 use the same sample to compare dispersion within 4 digit industries with dispersion in 4 digit industry by employment band (20-49, 50-249, 250+) cells. There is slightly less dispersion when comparing firms of a similar size, with a 90-10 ratio of 12.1 instead of 13.1, but there is still extensive dispersion.

The EAP data also contains production in quantity terms, where the units of quantity are consistent across firms within each 8 digit product, meaning that measures of productivity in quantity terms can be computed.[14] Therefore, the final test shown in the final two rows of Table 3.1 examines whether such dispersion exists when output is measured in quantity terms instead of value added terms. While energy, labour, and value added are measured at the firm level, quantity is measured at the product level. Two approaches are taken to ensure that comparisons are only made within product markets, where units are consistent. First, in row 13, firm-level measures are apportioned to the product level using the share of the firms' sales going to each product market. Second, in row 14, only single product firms are included, accounting for roughly 40% of the sample.

The data reveal that dispersion is more than double when measured as quantity of production per energy compared with value added per energy. This suggests that physical productivity and price are negatively correlated, in line with Foster, Haltiwanger, and Syverson (2008[20]). Note also that dispersion in labour productivity increases significantly when measured in quantity rather than value added terms. This is likely to be due again to differences in markups between small and large producers, as well as differences in the capital and material to labour ratios.

[13] The results are similar when looking at specific years and when taking the weighted average across products or industries, with value added used as the weight.

[14] Note that quantity productivity measures cannot be computed at the industry level as the units of output vary within an industry.

Table 3.1. Firms producing the same product have very different energy productivity

Average energy and labour productivity in France within different definitions of products and industries

Row no.	Product or industry-level	Dispersion in energy productivity	Dispersion in labour productivity	Dispersion in energy productivity (quantity)	Dispersion in labour productivity (quantity)
1	2 digit industry	17.5	3.2		
2	3 digit industry	14	3.2		
3	4 digit industry	12.2	3		
4	4 digit industry (industry by size class sample)	13.1	3.1		
5	4 digit industry by size class	12.1	3		
6	4 digit industry (8 digit product sample)	9.4	2.8		
7	6 digit product (highest firm share)	10.2	2.9		
8	6 digit product (firm share >50%)	9.9	2.9		
9	6 digit product (firm share >50%)	9.6	2.9		
10	8 digit product (highest firm share)	8.9	2.8		
11	8 digit product (firm share >50%)	8.7	2.8		
12	8 digit product (firm share >80%)	8.6	2.9		
13	8 digit product (all firms with proportionality assumption)	10	2.9	28.2	21.6
14	8 digit product (single product firms)	8.5	2.8	18.4	14.3

Note: The table shows dispersion in energy and labour productivity within different definitions of products and industries. Energy productivity is defined as value added per unit of energy and labour productivity is defined as value added per worker, except in the two right columns where quantity is used instead of value added. Dispersion is measured as the log difference between the 90th and 10th percentile firms within the product or industry, and the exponent is taken to present the difference as a ratio. Industries are measured at either the 2, 3, or 4 digit levels. Products are measured at either the 6 or 8 digit level. Firms are assigned to a single product depending on the one that accounts for the largest share of sales, except in row 13 where firm-level values are distributed to their products according to the share of sales accounted for by each respective product. Firms for which the most important product accounts for less than 50% or 80% of sales are excluded in some rows (as indicated in the first column). The final column restricts to firms selling only one 8 digit product. The average dispersion is taken across industries or products and across years.
Source: authors' calculations

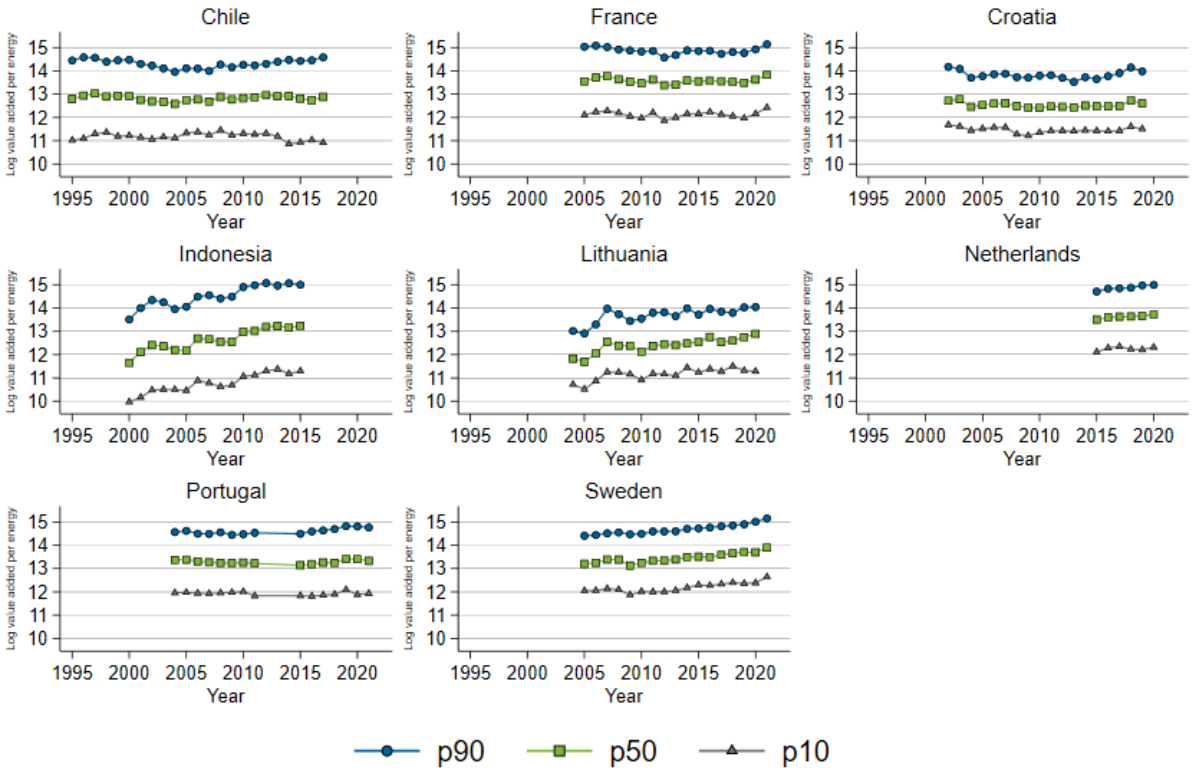
Dispersion over time

Figure 3.3 plots trends in energy productivity at various percentiles of the within-industry distributions over time. Averaging across industries, there have been general improvements in energy productivity over time at various key percentiles of the distribution (10th, 50th, and 90th), notably in Indonesia, Lithuania, and Sweden. In Croatia, Netherlands and Portugal, there appears to have been more improvement at the top than the bottom in recent years.[15] In Chile, there has been improvement at the top over recent years but not at the bottom, leading to increases in dispersion. Finally, in France, there has been an improvement at both the top and bottom in the most recent years, following a period of flatter performance.

[15] Figure A B.4 presents changes in dispersion over time more clearly, by plotting the average ratio between the energy productivity of the 90th and 10th percentile firms within an industry. To allow for an interpretation of differences between firms in ratios, the exponential of average log dispersion is taken and plotted.

Figure 3.3. Energy productivity has improved across all percentiles of the distribution in each country

Energy productivity at the 10th, 50th, and 90th percentiles within 3-digit industries over time



Note: The charts plot the 10th, 50th, and 90th percentile of within-industry energy productivity over time. Industries are defined at the 3-digit level and the unweighted average across industries is taken. The sample is restricted to manufacturing industries with at least 10 firms and firms with at least 20 employees.
Source: authors' calculations

Testing for catch up in energy productivity

The trends in the previous figure suggests that firms below the energy productivity frontier are not catching up to the frontier. This section formally tests whether less productive firms are catching up and whether the rate of catch up has changed over time. Following Acemoglu et al. (2006[21]) and Andrews et al. (2019[2]), the rate of growth in productivity from one year to the next is regressed against the distance of the firm from the frontier in the previous year. The specification is written:

$$\Delta \log y_{ict} = \sum_{s} \beta_c^s GAP_{i,c,k(i),t-1} D_t^s + \boldsymbol{\gamma X_{ict}} + \alpha_{c,k(i),t} + \varepsilon_{ict} \qquad \text{Equation 1}$$

The dependent variable $\Delta \log y_{ict}$ is the annual growth of either energy or labour productivity for firm i in country c in year t. The variable $GAP_{i,c,k(i),t-1}$ determines the distance of firm i from the productivity frontier in its country-industry and is defined as:

$$GAP_{i,c,k(i),t-1} = \log y_{i,c,k(i),t-1}^{frontier} - \log y_{,ic,k(i),t-1}$$

where the frontier is defined as the average productivity of the top 10% of firms in the four-digit industry $k(i)$ of the firm. The sample is restricted to firms that survive in consecutive periods, for which the outcome variable is well-defined. Conditional on survival, the rate of productivity growth is expected to increase with distance from the frontier, so β_c^s is expected to be positive. To test whether the rate of catch up has changed over time, the coefficient β_c^s is allowed to vary for different time periods s, so D_t^s is defined to be a dummy variable that equals one for a specific time period and zero otherwise. The time periods used are three-year intervals up to the latest year of data for the country (e.g. for Croatia, the intervals are 2002-2004, 2005-2007, 2008-2010, 2011-2013, 2014-2016, and 2017-2019). If the rate of catch-up has increased, then the estimated coefficients should increase over time. Finally, the specification controls for basic firm characteristics – log employment and log age (defined to be 1 in the year the firm is born) – written as $\boldsymbol{\gamma X_{ict}}$ and an industry-year fixed effect α_{ckt}.[16] The regression is estimated separately for data on each country, which is necessary as it requires access to the confidential firm-level data.

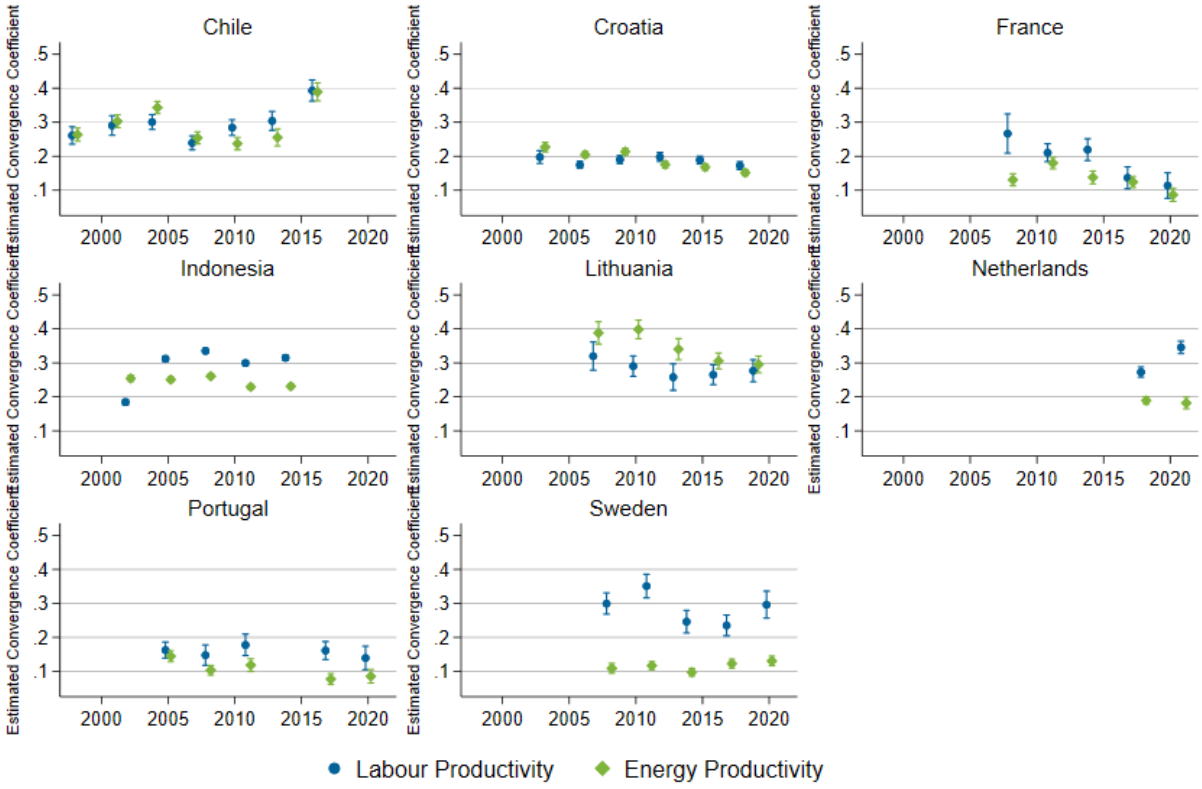
The results are presented in Figure 3.4, with the estimated coefficient for energy productivity in green diamonds and the coefficient for labour productivity, for comparison, in blue circles. As expected, the estimated coefficients are positive in all cases, suggesting that there is catch-up in productivity among surviving firms relative to the frontier. However, catch up has not accelerated in most countries. The rate of catch up decreased in Croatia, France (since 2010), Lithuania, Netherlands, and Portugal. It increased in Chile in recent years, and in Sweden.[17] In most countries, the rate of catch up in energy productivity is lower than the rate of catch up in labour productivity. The results are similar when using 2-digit industries, and when adjusting the weights to account for the probability that firms are re-sampled in countries where data does not cover the population of firms.

[16] The results are similar when including separate industry and year fixed effects and adding a control for the change in growth of frontier firms in the industry.

[17] Indonesia is excluded after 2015 as there is a trend break in the data.

Figure 3.4. There is no evidence of increasing catch up in energy productivity

Estimated difference in growth rates for non-frontier relative to frontier firms



Note: The chart plots estimated coefficients from regressions of the firm-level change in energy or labour productivity growth against the lag of each firm's distance from the productivity frontier in its industry. The rate of catch-up is allowed to vary for each three-year interval. The regressions include controls for log employment and log age, where age is defined to be one in the year of birth of the firm. The regressions are estimated by OLS and 95% confidence intervals based on heteroskedasticity-robust standard errors are plotted. Industries are defined at the four-digit level and the sample is restricted to manufacturing industries and firms with at least 20 employees.
Source: authors' calculations

Dispersion across industries

Looking at aggregates may hide interesting information on the distributions in specific sectors. Figure 3.5 plots percentiles of energy productivity for 2-digit manufacturing sectors in the earliest year available for each country (the dots) and the latest year for each country up to 2019, before Covid-19 (the arrows). The percentiles are computed at the 3-digit level to account for more detailed differences in production characteristics and then averaged to the 2-digit level. The sample is restricted to industries for which at least 10 firms are observed in each of the eight countries in the sample in each year. The unweighted average is then taken across the countries. Industries are sorted by the initial productivity of the 90th percentile.

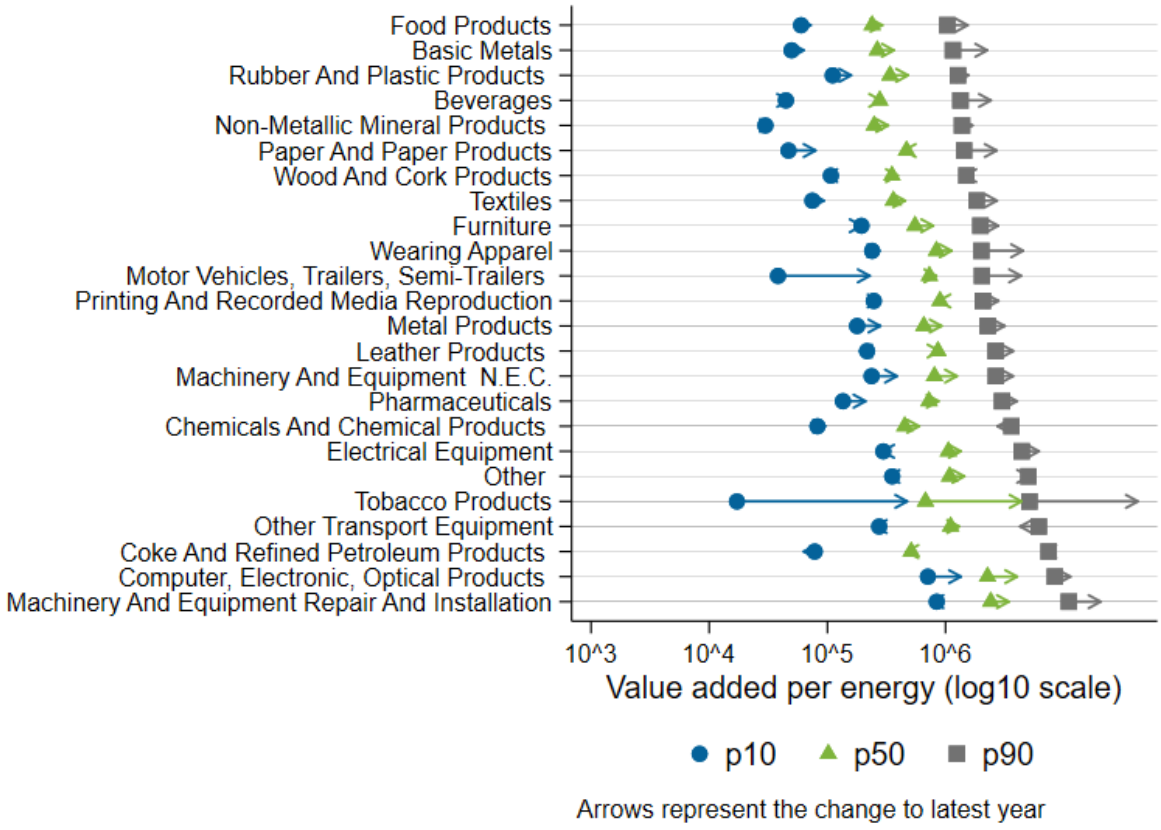
The energy productivity of firms at different percentiles of the distribution within industries differs across industries. The extent of dispersion also differs significantly across industries. Some industries, such as rubber and plastic products, have a smaller difference between the energy productivity of the 10th and 90th

percentile firms than others, such as non-metallic minerals. Despite this, there is a great deal of dispersion in all industries.

There are also substantial differences across industries in the change in energy productivity at different percentiles of the distribution over the sample period. Some industries have experienced an improvement in energy productivity across the distribution, such as metal products, but it is more common that industries have improved differently at different percentiles of the distribution. In particular, dispersion has clearly increased in some industries, whereby the 90th percentile has improved more than the 10th percentile. One notable example of this is basic metals, which includes steel. The production of steel is an industry with a relatively homogenous output but two distinct production technologies with differing levels of emissions intensity (blast furnace or electric arc). This industry therefore illustrates one mechanism through which firms may have such different energy productivities, even when producing a homogenous output. The possible explanations of dispersion and changes in dispersion are explored later in the paper.

Figure 3.5. Evolution of energy productivity across the distribution by industry

Energy productivity at the 10th, 50th, and 90th percentiles within 2-digit industries and their change over time on a log scale



Note: The chart shows the 10th, 50th, and 90th percentiles of energy productivity within industries and their changes from the earliest year in the sample in each country to the latest year up to 2019. The sample is restricted to country-industries with a first year before 2006 and a last year after 2014. The values are plotted on a base 10 log scale. The unweighted average across countries is taken. The sample is restricted to manufacturing industries with at least 10 firms in every country-year included in the plot, and firms with at least 20 employees.
Source: authors' calculations.

Figure A B.5, Figure A B.6, Figure A B.7, Figure A B.8, Figure A B.9, Figure A B.10, Figure A B.11, Figure A B.12 show the equivalent charts for individual countries instead of taking the average across countries. There is huge variation across industries in both the extent of dispersion and changes in dispersion when looking at individual countries. There are also clear differences across countries: Indonesia has experienced a substantial improvement in energy productivity across all percentiles of all industries while Lithuania has improved in most industries; in Chile and Croatia, there has not been much movement at the top but more heterogenous changes at the bottom of the distribution; there has been significant improvement in France, Netherlands, and Sweden in many industries at different points of the distribution; and Portugal has had significant improvement at the top in some industries.

Figure A B.13 shows the equivalent analysis but for labour productivity. It plots the 10^{th}, 50^{th}, and 90^{th} percentiles of labour productivity in the same 2 digit industries, averaged across countries. There is noticeably less dispersion in labour productivity than in energy productivity for all industries in the sample. Furthermore, industries that tend to have more dispersion in energy productivity are also more dispersed in labour productivity.

There is also variation in the performance of firms across countries among firms operating in the same industry. Figure A B.14 plots the average energy productivity of the top (90^{th} percentile) and bottom (10^{th} percentile) firms across countries for selected industries with a high coverage of firms across countries. As before, the percentiles are computed for 3-digit industries and averaged to the 2-digit level.

The chart highlights that the huge productivity dispersion exists both within and between countries. At both the top and the bottom of the distribution, there are often large differences in energy productivity between countries. Sweden has the most productive frontier firms in most industries, followed by France, while the ranking of other countries differs by sector. In some industries, the 10^{th} percentile Dutch, French or Swedish firms are not much less productive than the 90^{th} percentile firms of some other countries.

Table A B.10 shows all 3-digit industries in the sample ranked by their average energy productivity dispersion across countries. The most dispersed industries tend to be heavy industries, with energy costs comprising a higher share of total costs and firms having more energy emissions through energy use on average (shown in the last column).

4 Factors explaining firms' energy productivity

This section identifies the factors associated with firms' energy productivity. In other words, which factors might explain why firms have such great differences in energy productivity? Two complementary analyses are conducted. First, a machine learning (ML) prediction model is used to estimate which factors are most important in explaining variation in energy productivity. Second, regressions are estimated to show the direction and magnitude of the correlations. The results of these two exercises are then discussed together. All analysis is conducted within industry-country-year cells to identify variation across firms producing comparable products.

Factors predicting firm-level energy productivity

The main aim of ML prediction models is to use patterns of observed data to deduce the relationships between variables which help to predict values of a variable out-of-sample. In these models, it is possible to identify which explanatory variables (or "predictors") are most important in affecting the predictions of the dependent (or "response") variable.

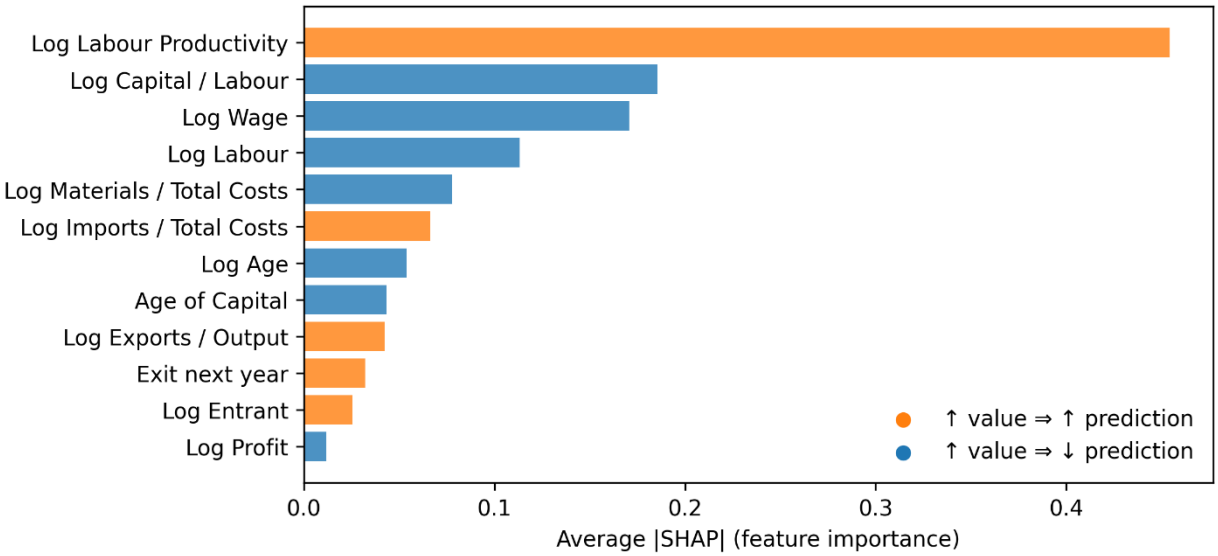
Therefore, to understand the most important factors in explaining firms' energy productivity relative to other firms in their industry, ML prediction models are estimated. The main method used is extreme gradient boosting (XGBoost), while an alternative (light gradient boosting machine - LightGBM) is also tested for robustness. These methods combine many weak prediction methods, such as decision trees, to form a strong predictor. The model is trained on 80% of the data and tested on the remaining 20%, with k-fold cross-validation. Importantly, both these methods can use observations with missing values in the predictor variables, unlike regression (this is particularly important for the analysis in Section 6 where there are more missing values). They also allow for non-linear effects without explicitly specifying the functional form of the non-linearity, as is required with OLS. Shapley additive explanations (SHAP) is then used to identify the factors that are most important in the model predictions, on a random sample of 5 000 observations to reduce computational burden. SHAP is specifically designed to identify the most important predictors, and can measure the direction in which each feature affects the predictions.

Conducting this analysis requires use of the firm-level data with numerous additional packages. Therefore, the analysis is conducted on a subset of countries for which firm-level data can be directly accessed. The results are based on collated firm-level data for five countries: Chile, Croatia, Indonesia, Lithuania, and Portugal. For countries with surveyed data, the models are implemented with sampling weights. All variables are first residualised on country-industry-year fixed effects (separately within the test and train datasets to avoid contamination) to focus on the within-industry differences between firms' energy productivity. Figure 4.1 shows the factors that are most important in predicting firms' (log) energy productivity. It shows the mean of the absolute value of all SHAP values for each predictor. The larger the values, the greater the absolute impact of that factor on the predicted value of energy productivity. The chart also provides an indication of the direction of the marginal effect of the factor on predicted energy productivity: orange bars tend to increase predicted energy productivity while blue bars tend to decrease

it. This is computed by estimating a linear regression of the SHAP value against the actual value for each factor – which are plotted in Figure A B.15. The SHAP values are plotted against the value of the predictor for each observation. A positive SHAP value for an observation means that the predictor made a positive contribution to the predicted value of energy productivity for that observation, and vice versa for negative SHAP values. Therefore, observing a positive correlation between the value of the predictor and the SHAP values suggests that higher values of the predictor tend to increase the expected value of energy productivity (loosely similar to observing a positive coefficient in a linear regression model). The equivalent results for explaining variation in labour productivity (instead of energy productivity) are presented in Figure A B.16 and Figure A B.17 of the Annex.

Figure 4.1. Labour productivity is the most important predictor of energy productivity

Average SHAP values of predictors of energy productivity and direction of average marginal effect



Note: The chart plots the mean of the absolute SHAP value for each predictor. The response variable is energy productivity at the firm level. Both the predictor and response variable are measured subtracting the country-industry-year average value, computed separately within the training and test datasets. The prediction model is estimated using XGBoost and the SHAP analysis is carried out on a random sample of 5 000 firm-year observations. The sample is restricted to firms which have at least 20 employees in the manufacturing industry.
Source: authors' calculations.

Factors correlated with firm-level energy productivity

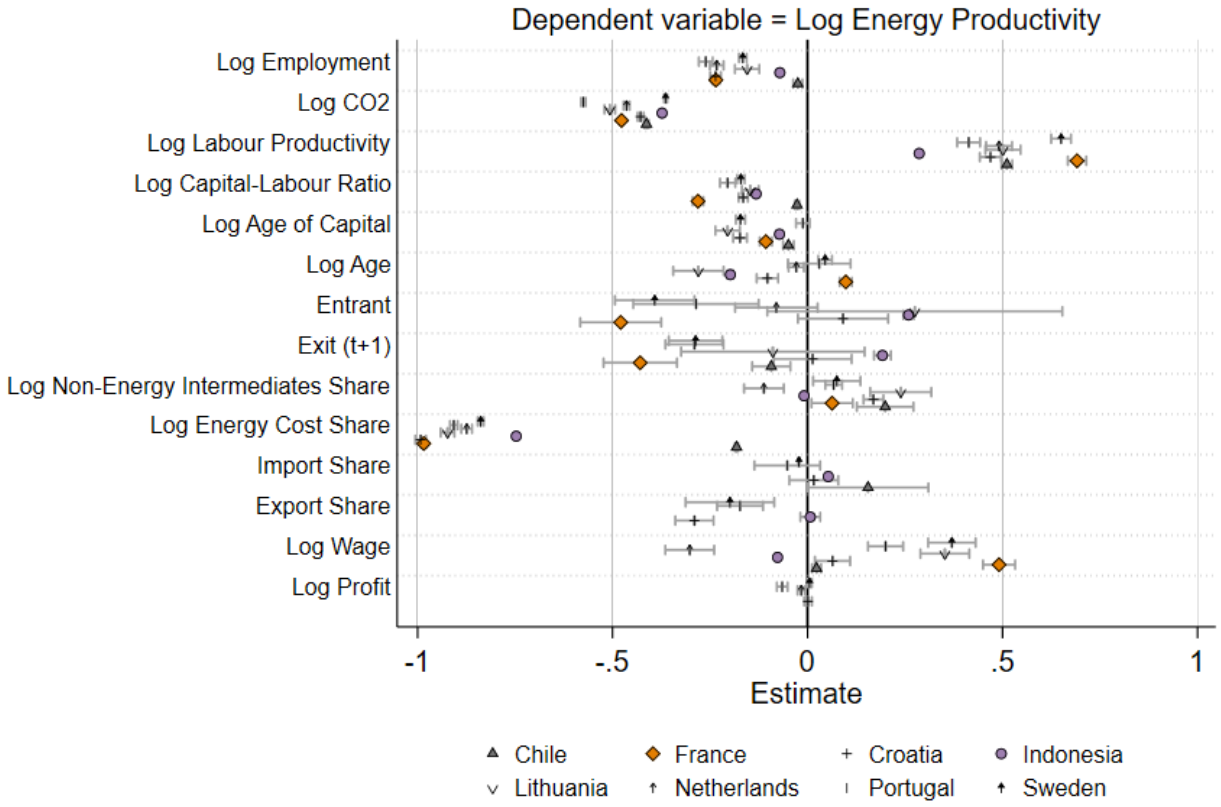
The ML prediction model provides clear insights on which are the most important factors in explaining differences between firms' energy productivity. It does not provide such clear insights on *how* each factor is related to energy productivity. Therefore, this analysis is complemented with linear regressions, which provide estimates of elasticities, i.e. how much each percentage increase in a variable is associated with energy productivity. The regressions directly use firm-level data and so are estimated for each country separately. They estimate the relationship between (log) energy productivity and various firm characteristics. Specifically, the following relationship is estimated as:

$$\log\left(\frac{VA_{ict}}{ENQ_{ict}}\right) = \beta_c\, x_{\text{ict}} + \alpha_{ck(i)t} + \varepsilon_{ict} \qquad \text{Equation 2}$$

for firm i in country c operating in industry $k(i)$ at time t, where $\alpha_{k(i)t}$ is a 4-digit industry-year fixed effect and ε_{ict} is an error term. $\log\left(\frac{VA_{ict}}{ENQ_{ict}}\right)$ is the log of energy productivity, with ENQ_{ict} referring to quantity of energy and VA_{ict} being value added.[18] The relationships are estimated as separate bivariate regressions, meaning that the explanatory variable x_{ict} is adapted across specifications. The regression is necessarily estimated separately for each country because firm-level data cannot be pooled across countries due to access regulations. Therefore, separate estimates of β_c are presented for each country.

Figure 4.2. Relationships between energy productivity and other firm characteristics

Estimates from bivariate OLS regressions of log value added per energy on various explanatory variables



Note: The chart shows estimates from bivariate OLS regressions of log value added per energy on various explanatory variables. They are estimated separately for each country, with country-specific coefficients reported in the plot. The y axis corresponds to the explanatory variable in each regression. The bars indicate 95% confidence intervals with standard errors that are robust to heteroskedasticity. The regressions all include industry-year fixed effects. The sample is restricted to firms with at least 20 employees in manufacturing industries.
Source: authors' calculations

The results are presented in Figure 4.2. A multivariate regression version, whereby the explanatory variables are included together, is also presented in Figure A B.18. Also, for comparison, the equivalent bivariate regressions are also estimated for labour productivity, with the results presented in Figure A B.19. Given the strong correlation between labour productivity and energy productivity, it might be expected that equivalent regressions to the above but with labour productivity as the dependent variable might yield similar results. Figure A B.19 presents results for bivariate regressions with log labour productivity as the

[18] All regressions including high-dimensional fixed effects are estimated using the estimator of Correia (2017[43]).

outcome and the same explanatory variables on the right-hand side. The results are often contrasting to those for energy productivity as the outcome. The results for each set of variables are discussed below.[19]

Discussion of key factors in explaining differences in energy productivity

This sub-section discussed the relationships between energy productivity and each factor, both in terms of its predictive power in the ML model and the direction of relationship from the linear regressions.

Labour productivity

Labour productivity is overwhelmingly the most important factor in predicting firms' energy productivity, even conditional on the other predictors. The relationship is positive, with energy productivity strongly positively correlated with labour productivity in all countries of the sample. This holds in the multivariate regressions and in the ML analysis, which both control for other possible explanatory factors. Interestingly, when looking at ML model predictors of labour productivity, the most important predictor is the firm's wage rate, suggesting that the relationship is not simply mechanical. For energy productivity, higher wages actually predict lower productivity. The results therefore show that firms that are more labour productive tend to be more energy productive. In line with this, Barrows and Ollivier (2018[22]) find that more efficient firms are less emission intensive.

As an additional exercise, Figure A B.18 explores the relationship between changes in labour productivity and changes in energy productivity. The methodology is described in the Annex alongside the Figure. The results show that there is a strong positive relationship between changes in labour productivity and changes in energy productivity, even controlling for firms' initial characteristics and changes in their characteristics.

Together, the evidence suggests that economic and environmental productivity are closely linked and that improvements in labour productivity may also bring improvements in energy productivity.

Capital intensity

The second most important factor is capital intensity. Capital intensity, defined as capital per worker, is negatively correlated with energy productivity in all countries in the sample. This result holds when controlling for other factors in the multivariate regressions and ML analysis. It suggests that energy is likely to be complementary to capital in the production function. In contrast and unsurprisingly, labour productivity is positively correlated with the capital-labour ratio.

Wages

Perhaps surprisingly, the third most important predictor of energy productivity is the wage rate. The estimated correlation between wages and energy productivity is positive in all countries of the sample except Netherlands and Indonesia in the bivariate regressions, suggesting that generally firms that pay higher wages tend to be more energy productive.

However, in the multivariate regressions where other factors like productivity are controlled for, the relationship between energy productivity and wages is negative in all countries except Sweden. The same is true, on average, in the ML prediction model although the relationship is non-linear: at low levels of wages, higher wages tend to increase predicted energy productivity while at high levels it tends to decrease

[19] In each figure, there is one estimated coefficient for export share in Chile that is excluded because it had very large confidence intervals which makes the other coefficients hard to interpret.

it. This could reflect that, controlling for their productivity and capital intensity etc, firms paying higher wages tend to be less efficiently managed, which could also explain their lower energy productivity.

Firm size: energy, emissions, and employment

Employment is also predictive of energy productivity in the ML model. Strikingly, the regressions show that larger firms tend to be less energy efficient and this result holds in all countries in the sample. This is an interesting result and contrasts with the strong empirical and theoretical finding that labour productivity and MFP are positively correlated with firm size (as is also found in the data for this paper, shown in Figure A B.19). The negative relationship between energy productivity and employment holds even when controlling for other factors.[20]

One explanation for this finding is that larger firms tend to pay lower energy prices giving them relatively lower disincentives to use energy more efficiently. Another possible explanation could be that economies of scale may instead be negative for energy use. It is important to note that Table 3.1 showed that there is huge dispersion between firms' energy productivity even when comparing firms with similar employment. Therefore, it appears to be technologically possible to improve energy productivity even without downscaling.

Age of capital, firm age, and entry

Over time, capital is likely to become less energy intensive, meaning that firms with younger vintages of capital are likely to have more energy efficient production. To test this, age of capital is measured following Fiori and Scoccianti (2025[23]): the perpetual inventory method is used to measure capital stock and then the age of capital is the number of years since an investment spike, whereby investment exceeds 20% of capital in that year. An alternative measure is simply the age of the firm itself, whereby younger firms may be expected to have newer vintages of capital.

Interestingly, the age of capital, as well as entry and firm age, are towards the bottom of the ranking of factor importance. Despite this, they do still have some predictive power and there are significant correlations in the regressions.

In all countries in the sample, firms with older capital have significantly lower energy productivity, which also holds controlling for other factors. This suggests that newer vintages of capital may indeed be more energy efficient. The relationship with firm age is more mixed, with some countries having a negative estimated coefficient and others having a positive. Similarly, there are mixed relationships for whether the firm is a recent entrant. However, the results on entry should be interpreted with caution: while entry is identified based on when the firm enters, the sample only includes firms with at least 20 employees, meaning that these estimates are the relationship with entry conditional on entering with at least 20 employees.

Overall, the results suggest that firms making more recent investments in capital tend to be more energy productive.

Input cost shares

The relationship between energy productivity and the share of non-energy intermediate inputs (materials) costs in total costs varies by country. In France and Sweden, for example, the relationship is positive, while

[20] One possible explanation for this result could be that firms employing fewer workers are outsourcing more to other plants within the same firm. However, the result also holds at the establishment level in the French data, where outsourcing is less likely to affect the measurement of energy productivity. Data on value added is not available at the establishment level so energy productivity is proxied with the inverse of energy per worker.

in Netherlands, the relationship is negative and in Indonesia it is close to zero. There are two possible forces at play here. First, firms with a higher share of intermediates have more complex production procedures requiring more parts to be assembled, which may require more energy. A contrasting force would be that firms with high shares of intermediates may be outsourcing a higher share of production or be in a more downstream industry and may therefore be more energy efficient.

Profits and exit

It would be expected that more productive firms make higher profits and are less likely to exit. Indeed, for labour productivity, both these relationships hold for all countries. However, strikingly, the results for energy productivity are far less clear. In all countries, the estimated coefficient on profits is small and is often not significantly different from 0. Furthermore, it is the least important factor in predicting energy productivity in the ML model. Firms in France, the Netherlands, and Sweden – countries that have been subject to carbon pricing – less energy productive firms are more likely to exit. However, the same is not true Croatia and Indonesia. The results therefore suggest that a lack of competitive pressure may allow for firms to survive with poor energy productivity. This is explored in more detail in the final section of the paper.

5 The aggregate implications of dispersion

This section shows that dispersion in energy productivity tends to be associated with worse aggregate outcomes. It then demonstrates the huge gains that could be achieved by improving the productivity of laggards.

Relationship between changes in dispersion and changes in macro performance

Increases in within-industry dispersion in energy productivity may be associated with either increases or decreases in aggregate outcomes of energy, value added, and productivity. Increases in dispersion could be due to top firms becoming more productive, unproductive firms becoming less productive, or through entry and exit of firms at different points of the distribution.

To examine the relationship between dispersion and aggregate outcomes, regressions in both levels and changes are estimated. The regressions in levels mainly exploit cross-sectional variation and measures whether industries that have more dispersion have systematically different aggregate outcomes. The specification can be written:

$$\log(y_{ckt}) = \beta\ DispEP_{ckt} + \varepsilon_{ckt}$$

Equation 3

The outcome variable $\log y_{ckt}$ is the log of the aggregate value of either value added, energy quantity, CO_2 emissions, employment, energy productivity, carbon productivity, or labour productivity, in industry k in country c in year t. The explanatory variable of interest is dispersion in energy productivity, measured as the difference between the 90th and 10th percentile firms within a 3-digit industry and year:

$$DispEP_{ckt} = \log\left(\frac{VA_{ckt}}{ENQ_{ckt}}\right)^{p90} - \log\left(\frac{VA_{ckt}}{ENQ_{ckt}}\right)^{p10}$$

Equation 4

The relationship of interest is the simple correlation between dispersion and aggregate outcomes. Therefore, a simple bivariate regression specification is chosen, in line with the specification of Andrews, Criscuolo, and Gal (2019[2]). The key results are robust to including country-year fixed effects (only the cross sectional results relating to labour qualitatively change, becoming statistically insignificantly different from zero).

The results are presented in Table 5.1, showing that industries with more dispersion in energy productivity have worse aggregate energy, carbon, and labour productivity. They also tend to be significantly larger in terms of labour, value added and, especially, energy use and CO_2 emissions.

To examine how changes in dispersion are correlated with changes in aggregate outcomes, the estimated regressions are instead estimated in 5-year differences:

$$\Delta_5 \log(y_{ckt}) = \beta\, \Delta_5 DispCP_{ckt} + \varepsilon_{ckt}$$

Equation 5

The results are presented in Table 5.2. Increases in dispersion are negatively correlated with energy and carbon productivity, statistically significantly different from 0 at the 5% level. This suggests that increases in dispersion may be driven by less productive firms becoming even less productive, or net entry resulting in more dispersion and lower overall productivity. It suggests that dispersion in energy productivity may be causing a drag on aggregate energy productivity and carbon efficiency. These results appear to be driven by both the numerator and denominator, with increases in energy dispersion being associated with a relative decline in aggregate value added and a relative rise in aggregate energy use. The results are similar with carbon productivity dispersion as the explanatory variable.

Table 5.1. Industries with more dispersion in energy productivity generally have worse aggregate energy productivity

Regressions of aggregate variables on within-industry dispersion in carbon productivity

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	log(VA)	log(Energy)	log(CO_2)	log(L)	Log(VA/Energy)	log(VA/CO_2)	log(VA/L)
log(VA/Energy) p90-p10	0.192***	0.559***	0.856***	0.317***	-0.368***	-0.668***	-0.125***
	(0.0237)	(0.0279)	(0.0292)	(0.0219)	(0.0188)	(0.0209)	(0.0145)
Observations	5085	5013	5002	5088	4998	4980	5085

Note: The table presents results from regressions of aggregate outcomes on within-industry-country-year dispersion in energy productivity, measured as the log difference between the 90th and 10th percentile firm. The outcome variables of interest, in logs, are: value added (VA), energy, CO_2, labour (L), energy productivity (VA/Energy) carbon productivity (VA/CO_2), and labour productivity (VA/L). The data is at the 3-digit industry-country-year level and is restricted to firms with at least 20 employees in manufacturing industries. Standard errors are robust to heteroskedasticity. * p<0.10 ** p<0.05 *** p<0.01.
Source: authors' calculations.

Table 5.2. Increasing dispersion in energy productivity is correlated with declining aggregate energy productivity

Regressions of changes in aggregate variables on changes in within-industry dispersion in energy productivity

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	log(VA)	log(Energy)	log(CO_2)	log(L)	Log(VA/Energy)	log(VA/CO_2)	log(VA/L)
$\Delta_5 DispENQ$	-0.0306*	0.0657***	0.0507***	-0.0116	-0.0966***	-0.0795***	-0.0194
	(0.0181)	(0.0179)	(0.0189)	(0.00993)	(0.0163)	(0.0174)	(0.0140)
Observations	3195	3147	3125	3200	3134	3104	3195

Note: The table presents results from regressions of changes in aggregate outcomes on changes in within-industry-country-year dispersion in energy productivity, measured as the log difference between the 90th and 10th percentile firm. The outcome variables of interest, in logs, : value added (VA), energy, CO_2, labour (L), energy productivity (VA/Energy) carbon productivity (VA/CO_2), and labour productivity (VA/L). The data is at the 3-digit industry-country-year level and is restricted to firms with at least 20 employees in manufacturing industries. Standard errors are robust to heteroskedasticity.* p<0.10 ** p<0.05 *** p<0.01.
Source: authors' calculations.

Counterfactual exercise: improving the energy and carbon productivity of unproductive firms

To what extent could reducing dispersion reduce aggregate energy use or affect the carbon footprint of production? To provide evidence on this, a counterfactual exercise is carried out whereby the energy productivity of the least productive firms is raised to the 25th percentile of the distribution in their 4-digit industry (the most detailed level available in the firm-level data) in their country in 2019 (2015 for Chile and Indonesia). Then, assuming all the firms had at least that energy productivity, the implied amount of energy that would be produced to achieve the actual level of production observed is computed. Put differently, if all firms below the 25th percentile of energy productivity in their sector had an energy productivity equal to the firm at the 25th percentile, how much less energy would be needed to produce the same level of value added? The same exercise is carried out in terms of carbon productivity and aggregate emissions.

Mathematically, define x_i to be the energy (or carbon) productivity of firm i, the counterfactual productivity for each firm is computed as:

$$x_i^* = \begin{cases} x_i & \text{if } x_i \geq x^{p25} \\ x^{p25} & \text{if } x_i < x^{p25} \end{cases}$$

Then, holding value added constant, the quantity of energy (or carbon) implied by this productivity is calculated. The total of this energy across firms is taken, $\sum_i x_i^*$, and compared to the observed total quantity of energy $\sum_i x_i$. Figure 5.1 plots the implied difference for each country in the sample.

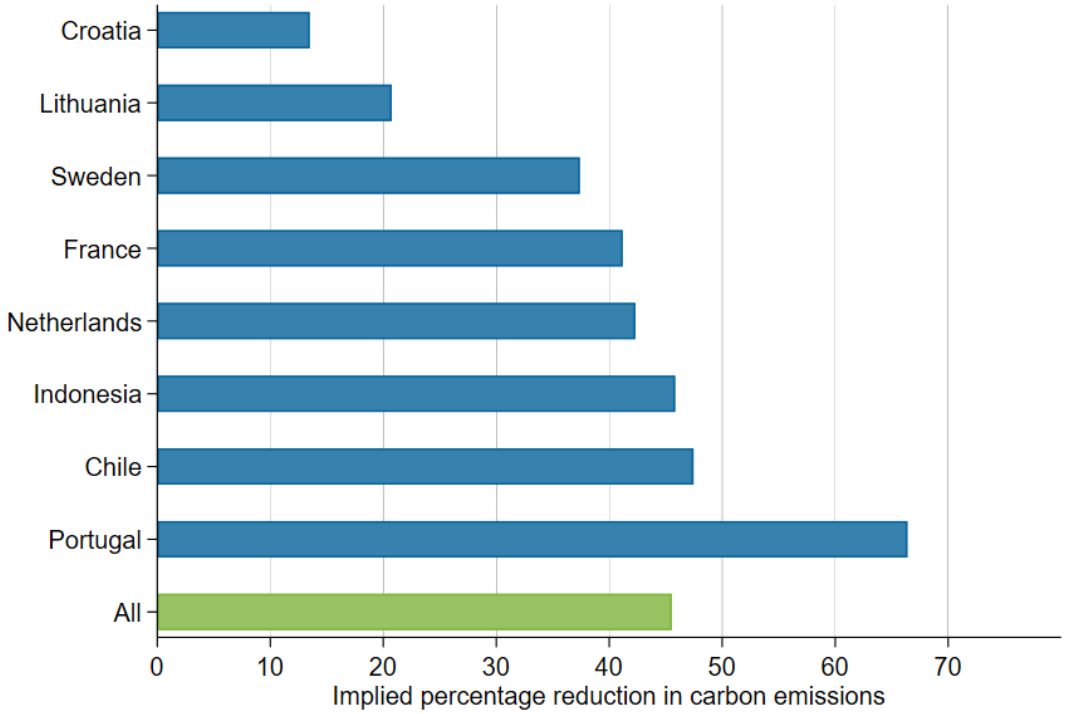
Raising all firms to the 25th percentile of energy productivity within their industry would reduce energy use by 45% to achieve the same level of value added, aggregating across all countries (noting that larger countries have a greater impact on the aggregate). Looking at each country individually, it would reduce energy use by 67% in Portugal, 47% in Chile, 46% in Indonesia, 42% in Netherlands, 41% in France, 37% in Sweden, 21% in Lithuania, and 14% in Croatia. The large potential improvements reflect a higher degree of dispersion at the bottom of the distribution and that larger energy users tend to be less productive. Figure A B.21 shows the equivalent exercise but improving firms to the 50th and 75th percentiles, in addition to the 25th percentile, showing that improving all firms to the median would lead to a reduction in energy use of 64% to achieve the same output.

Performing the same exercise for carbon productivity yields very similar results to that of energy productivity. Overall, emissions could be reduced by 46% if all firms were raised to at least the 25th percentile in their industry. The results are similar to those of Bijnens and Swartenbroekx (2024[7]), who use data on 2 479 companies to show that a limited reallocation from the 20% most emission-intensive firms towards the rest could achieve a decrease in emissions of 38%.

The implied reductions in energy and carbon emissions are large, and this is driven by two factors. First, the huge dispersion in energy and carbon productivity means that the improvement that would be experienced by increasing the productivity of the least productive firms to the 25th percentile is large. Second is the empirical observation that energy use and carbon emissions are strongly negatively correlated with energy and carbon productivity (this is documented in Section 4 within industries but also holds across industries). Hence, the largest energy users are more likely to have low energy productivity, and improving the energy productivity of large energy users will have a greater impact on aggregate energy use and carbon emissions than improving that of smaller energy users.

Figure 5.1. Improving the energy productivity of the least efficient firms can substantially reduce energy used to produce the same output

Counterfactual exercises showing implied reduction in energy to produce the same aggregate value added if all firms were at least as productive as the 25th percentile firms in their industry



Note: The chart shows the implied percentage reduction in energy that would be achieved if all firms had at least an energy productivity equal to the 25th percentile firm in their 4-digit industry. The data is for the latest year available in each country up to 2019 and is restricted to manufacturing sector firms with at least 20 employees.
Source: authors' calculations

Of course, it may be unlikely to expect that a small firm could improve to the energy productivity of a large firm, or vice versa. Therefore, the equivalent exercise is conducted within industry by employment bands. The results, presented in Figure A B.22, show that the implied gains are slightly smaller but the potential reductions in energy use are still huge.

It is also possible that differences in productivity arise because firms produce different products, despite the evidence in Table 3.1 showing that dispersion is similar when measured at the product level to the industry level. To test whether these large potential gains are driven by firms producing different products, Figure A B.23 presents results using French data on production by product market from EAP. It shows that the potential gains are around a third smaller when firms are improved to the 25th percentile within their product market compared with in their industry, but they are still of similar magnitude. The key takeaway for policy that the potential gains from improving the energy productivity of the least productive firms still holds.

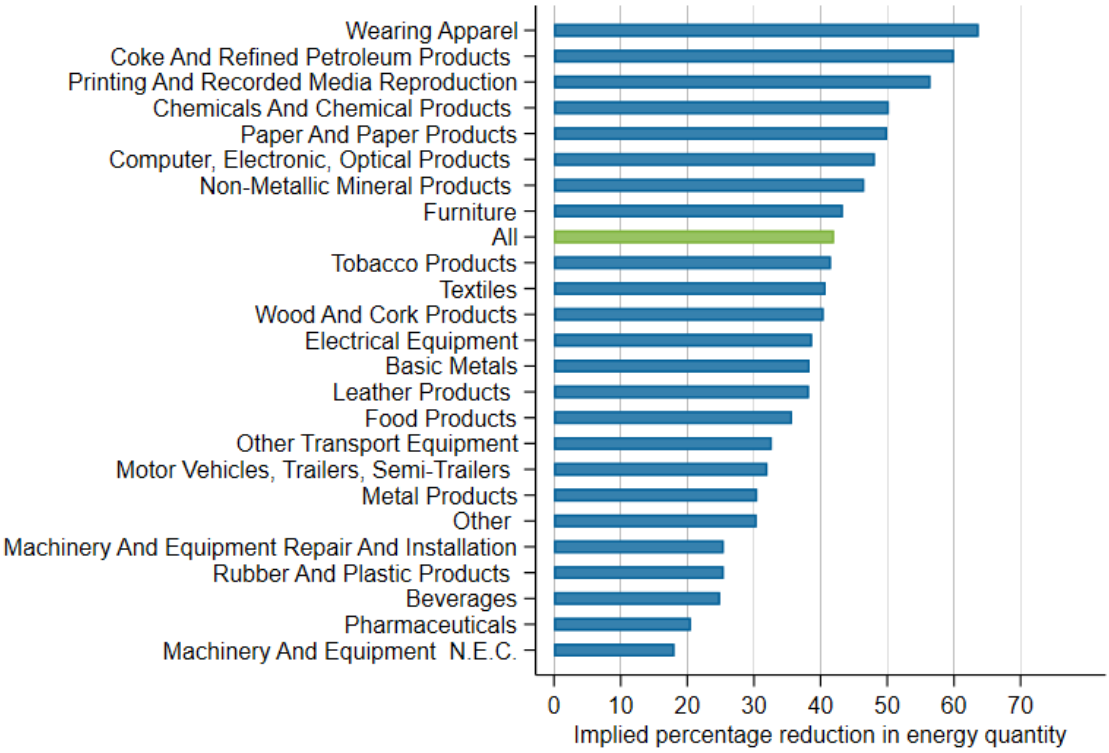
To provide a benchmark comparison, Figure A B.24 presents results from the equivalent exercise for the number of workers that could provide the same level of output if all firms operated at least at the 25th percentile of labour productivity. Reflecting that there is much less dispersion in labour productivity than in energy or carbon productivity, the percentage difference in workers necessary to achieve the same output is much smaller than the percentage reduction in energy or carbon. Bringing all firms to the 25th percentile in their industry and aggregating across countries, for energy productivity the percentage reduction in

energy is 45%, for carbon the equivalent result is 46%, while for labour productivity the percentage reduction in labour is just 8%.

These large potential gains are not driven by just a few sectors. Figure 5.2 plots the potential reduction in energy that could be achieved by improving unproductive firms to the 25th percentile in their 4 digit industry, as in Figure 5.1, presented separately for each 2 digit industry.[21] Even the industry with the smallest potential for improvements could achieve nearly a 20% reduction in energy use, while the industry with the most potential for improvement, wearing apparel, could achieve an improvement of 63%.

Figure 5.2. The potential gains from improving firms with low energy productivity vary significantly across industries

Counterfactual exercises showing implied reduction in energy use to produce the same aggregate value added if all firms were at least as productive as the 25th percentile firms in their industry, presented for each 2 digit industry



Note: The chart shows the implied percentage reduction in energy that would be achieved if all firms had at least a energy productivity equal to the 25th percentile firm in their 4-digit industry, with the results aggregated to the 2 digit industry level. The data is for the latest year available in each country up to 2019 and is restricted to manufacturing sector firms with at least 20 employees.
Source: authors' calculations

[21] Data for Chile and Netherlands are not included in this chart. For Chile, 4 digit industry codes are unavailable while for Netherlands the underlying statistics could not be released for confidentiality reasons.

6 Explaining dispersion in energy productivity

There are many factors which may explain why some industries have more dispersion in energy productivity than others. Like Section 4, this section uses a combination of ML methods and linear regressions to show which factors explain within industry dispersion in energy productivity, and the direction of the relationship. It introduces several policy-relevant variables which may be related to dispersion and discusses the results for each factor at the end of the section.

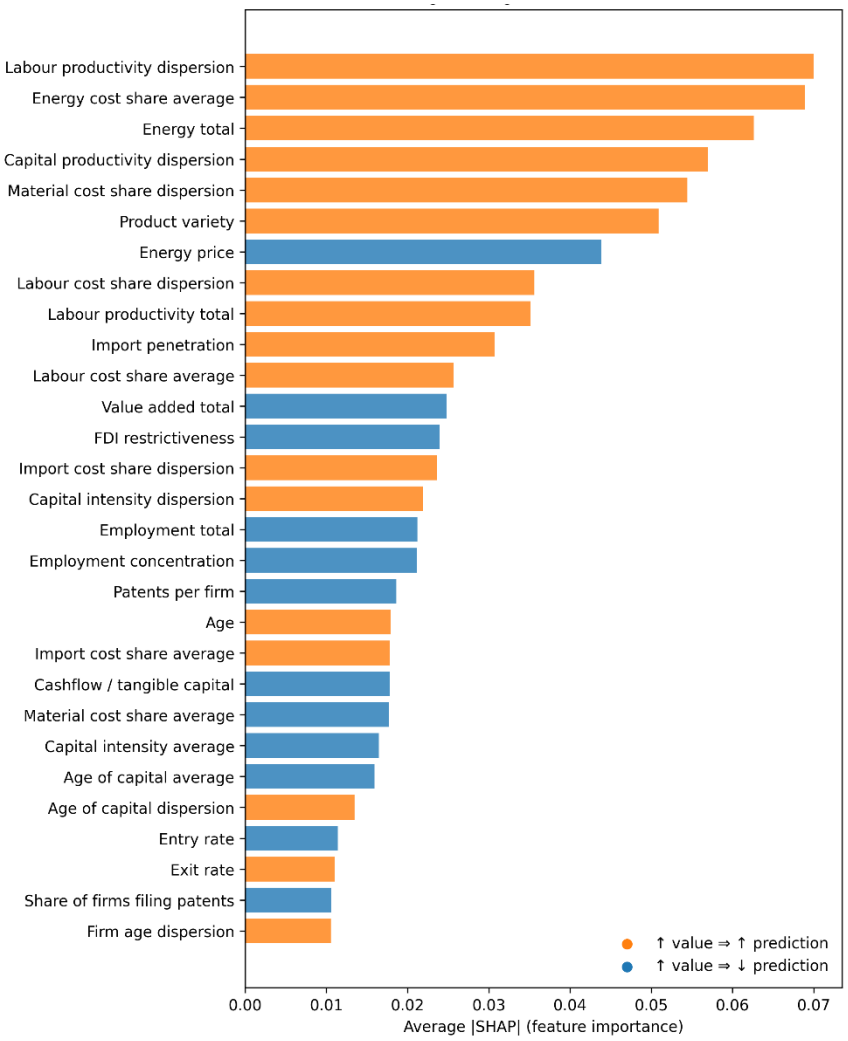
Factors predicting dispersion in energy productivity

The same ML method that was used above is also used to examine the variables with the most predictive power for industries' dispersion in energy productivity. A wide range of predictors are included in the model. Each of these variables is de-meaned by the country-year average, which is computed within the training or test sets separately to avoid contamination. The industry-country-year level data contains numerous missing values, either due to confidentiality requirements or insufficient firm counts. A key advantage of the ML prediction relative to regression is that observations with missing values can be included in the analysis. The performance of the ML model, including a comparison with OLS, is presented and discussed in Table A B.11 of the Annex.

Figure 6.1 shows the most important factors in predicting differences in energy productivity dispersion across industries. The length of the bar reflects the importance of each factor in predicting dispersion. The colours of the bars show the direction in which the factor pushes the predicted value of dispersion, while Figure A B.25 shows this in more detail for the 12 most important factors in predicting dispersion, including potential non-linear relationships. While the cross-sectional differences across industries' energy productivity are interesting, it is also relevant from a policy perspective to look at the factors that predict *changes* in energy productivity dispersion over time. Therefore, all variables (where possible) are expressed in within-industry-country differences and the results are presented in Figure A B.26.

Figure 6.1. Dispersion in labour productivity and the average energy cost share are important predictors of energy productivity dispersion

Average SHAP values of predictors of dispersion in energy productivity in XGBoost model and direction of average marginal effect on model prediction



Note: The chart plots the mean of the absolute SHAP value for each predictor. The response variable is industry dispersion in energy productivity. Both the predictor and response variable are measured subtracting the country-year average value, computed separately within the training and test datasets. The prediction model is estimated using XGBoost. The sample is restricted to manufacturing industries with at least 10 firms which have at least 20 employees.
Source: authors' calculations.

Factors correlated with energy productivity dispersion

Alongside the predictions of the ML model, linear regressions of the same explanatory variables on dispersion in energy productivity are estimated according to the following specification:

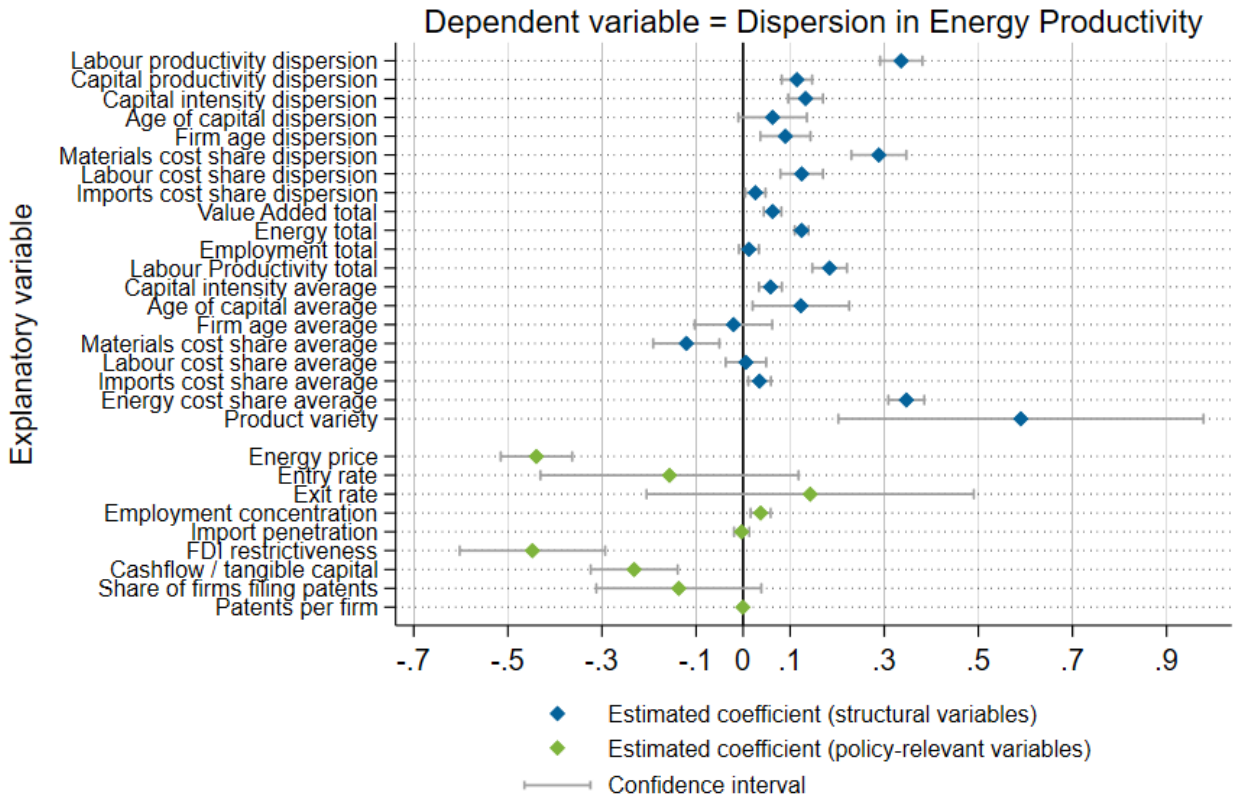
$$DispEP_{ckt} = \beta\, x_{ckt} + \alpha_{ct} + \varepsilon_{ckt}$$

Equation 6

The variable $DispEP_{ckt}$ is within-industry log dispersion in energy productivity, as defined above. The explanatory variable of interest x_{ckt} is changed in each regression specification and is measured in logs, unless otherwise stated. The variables that are most relevant for policy are separated at the bottom of the chart, with the estimated coefficients highlighted in green. The remaining terms are a country-year fixed effect α_{ct} to control for country-specific characteristics and an error term ε_{ckt}. Note that sector-year fixed effects are not included in the regressions because the specification is designed to capture the characteristics of industries with more dispersion. Therefore, it is not desirable to control for averages of industries across countries for this purpose. The regressions are estimated by OLS on the cross-country micro-distributive dataset and standard errors are robust to heteroskedasticity.[22] The observation counts are mostly around 5 000 industry-country-years, with some specifications having counts of around 3 000 because the dependent variables being missing in some countries. The results are presented in Figure 6.2 and the corresponding tables are available on request.

Figure 6.2. Factors correlated with energy productivity dispersion

Estimated coefficients from bivariate OLS regressions of energy productivity dispersion on explanatory variables



Note: The chart shows results from regression estimates of within-industry dispersion in energy productivity on various explanatory variables. The regressions are estimated by OLS on industry-country-year level data, with industries defined at the 3-digit level. Each row is estimated as a separate bivariate regression. Country-year fixed effects are included as controls. The sample is restricted to manufacturing industries and firms with at least 20 employees. Standard errors are robust to heteroskedasticity and 95% confidence intervals are plotted.
Source: authors' calculations.

[22] Note that some countries are excluded from certain regressions as not all variables are available in all countries.

The main results presented above are for bivariate regressions. It is also possible to estimate multivariate regressions, whereby all explanatory variables are included together. As some variables have many missing values, they are excluded from the multivariate regression. The results from the multivariate regressions are shown in Figure A B.27. However, these results should be interpreted with some caution due to limited sample size and a lack of power.[23] The prediction models used in the following section are robust to missing values and have more explanatory power in this specific setting, with the disadvantage of not providing intuitive correlations or elasticity estimates.

Alongside these cross-sectional relationships, the regressions are also estimated in 5 year within-industry differences. This controls for the role of any unobserved time-invariant industry characteristics which may be correlated with both energy productivity dispersion and the explanatory variable can be removed. Defining Δ_5 as 5-year changes in a country-industry unit, the specification in changes is:

$$\Delta_5 DispEP_{ckt} = \beta\ \Delta_5 x_{ckt} + \alpha_{ct} + \varepsilon_{ckt}$$

Equation 7

The results for these are presented in Figure A B.28.

Discussion of factors explaining dispersion in energy productivity

Structural factors

The "structural" factors are discussed first, defined as those which relate to the economic characteristics of an industry and the structural composition of firms in each industry.

Dispersion

Dispersion in energy productivity is strongly positively correlated with other measures of industry dispersion, particularly dispersion in labour and capital productivity. These are among the most important factors in predicting dispersion in energy productivity, even controlling for many other variables. In the ML model in changes, they are the two most important factors in explaining dispersion.

A 10% higher labour productivity dispersion is associated with an 3.3% higher energy productivity dispersion. There is an extensive literature documenting the extent of and possible explanations for dispersion in labour productivity (see for example Berlingieri and Criscuolo (2017[12]), Andrews, Criscuolo, and Gal (2019[2]), Syverson (2011[14])), as well as possible solutions to help improve the productivity of laggard firms (Berlingieri et al., 2020[13]). Several possible explanations of dispersion in labour productivity have been proposed, including a lack of technology transfer between frontier and laggard firms, barriers to entry or scale-up, or a lack of competition, among others. The results presented in this section suggest that such remedies may also help to improve the environmental performance of firms if designed appropriately. Put differently, there may be synergies between improving catch up and allocative efficiency in both economic and environmental outcomes.

For capital productivity, the relationship is less strong but still positive and significant. A 10% increase in capital productivity dispersion is associated with a 1.1% increase in energy productivity dispersion. Capital and energy usage are likely to be complementary to production, so industries with more variation in capital

[23] Even after excluding some less-important variables from the multivariate regression, the observation count drops significantly and the regression may lack power. In particular, all data for Chile and Netherlands is dropped from both multivariate regressions due to missing variables. Furthermore, due to some missing values, there are only 1 837 observations in the multivariate regression sample.

usage across firms are also more dispersed in terms of energy productivity. Industries that are more capital intensive on average are also more dispersed in energy productivity.

Material cost share dispersion is the fifth most important factor in predicting dispersion. The share of materials in total cost is related to differences between firms in their production methods, which are also related to energy use, either as complements or substitutes. Therefore, when firms have more varied shares in these inputs, they also have more varied intensity of energy. Accordingly, dispersion in the material cost share is positively related to energy productivity dispersion. In contrast, dispersion in the cost share of imports is uncorrelated with energy productivity dispersion.[24]

Industry size

Three variables – value added, energy, and employment – capture the relationship between aggregate industry size and energy productivity dispersion. In general, larger industries are more dispersed: for example, 10% higher value added is associated with 0.7% greater dispersion. However, there is a negative correlation between increases in employment and dispersion in energy productivity, suggesting that industries tend to become less dispersed as they grow.

Energy use

Factors relating to the aggregate energy use of the industry – the total energy use and the average energy cost share - strongly influence an industry's predicted level of dispersion. There are multiple ways in which the average share of energy in total costs might be related to dispersion. First, when energy costs make up a higher share of total costs, there is mechanically more scope for dispersion. Second, with energy being a more important input, there may be more competitive pressure on firms to be energy efficient, reducing dispersion. It seems as though the former relationship dominates, as industries with a higher share of energy costs in total costs on average tend to be more dispersed. For the ML model in changes, energy use is only the fifth most important variable in explaining dispersion and it is only marginally statistically significant in the regression in changes.

Age of capital and firm age

The results show that industries with older capital and older firms on average are more dispersed in energy productivity. There is a positive correlation between dispersion in firm age and dispersion in energy productivity, but no statistically significant relationship for dispersion in age of capital. In changes, none of these variables are statistically significant, and they tend to rank relatively low in the prediction models. Therefore, the results suggest some tentative evidence that industries where firms have made older investments in capital may be more dispersed, but that the age of capital is less important than other explanatory factors such as dispersion in productivity and many of the policy-relevant variables discussed below.

Product variety

It is possible that, in some industries, firms simply produce a wider range of different products, which leads to them having different measured energy productivity. To measure the variety of products in each industry, data from Eurostat PRODCOM is used. As this data is only available for European countries, and the product classifications can change in each year, the measure of product variety is defined to be consistent across countries and over time. It uses data from just one year – 2019 – which maps to the ISIC Revision 4 industry classification used throughout the paper. Using the total value of sales for each 8 digit product,

[24] Note that these imports are those reported by the firm and are thus likely to be used as intermediate inputs. The relationships with import competition is studied below.

the Herfindahl-Hirschmann index (HHI) of products per industry can be computed. Product variety is defined as one minus the HHI of products per industry.

The results show that industries with more product variety do indeed have more dispersion in energy productivity, but they do not have systematically different changes in dispersion over time. While product variety is an important, it is not relevant for policymaking: it is simply a result of the nature of industry definitions. Therefore, the important point to take away from the result relating to product variety is that it is not the most important factor in determining dispersion, and that many other factors remain important even conditional on product variety.

Policy-relevant factors

These factors are those which are more closely related to policy in the sense that policymakers have toolkits that may be able to specifically address each factor.

Energy price

Among the policy-relevant variables, the average energy price faced by an industry is the most important predictor of energy productivity dispersion. After dispersion in labour productivity, changes in energy price is the most important predictor of changes in energy productivity dispersion. Higher energy prices faced by an industry are likely to increase the pressure to improve energy efficiency across the distribution of firms. Accordingly, higher energy prices are negatively correlated with dispersion, both in levels and changes. This suggests that policies that increase average energy prices, such as carbon pricing, may reduce dispersion.

Proxies of competition and business dynamism: entry, exit, and concentration

Business dynamism could be an important mechanism for improving aggregate energy productivity. As new businesses enter, they may have more energy efficient production technologies, and business dynamism also promotes Schumpeterian creative destruction, leading to improved allocative efficiency. For these reasons, they may also be related to dispersion between firms. Therefore, entry and exit rates are explored as potential explanators of energy productivity dispersion.

While the entry rate is negatively correlated and exit rate is positively correlated with dispersion, they both have large standard errors, meaning that they are not statistically significantly different from zero. The contribution of entry and exit to aggregate improvements in energy productivity will be examined in future work.

A lack of competition may allow firms to survive without using energy optimally (x-inefficiencies). To test this, the HHI of employment among firms is computed, capturing the extent to which employment is concentrated among leading firms.[25] While concentration is not a perfect measure of competition, it is often used as a proxy. Higher concentration is positively correlated with dispersion in energy productivity and has reasonable predictive power in the ML model, providing suggestive evidence in support of this mechanism.

Import penetration and FDI restrictiveness

Another way in which industries face competition is through imports. To measure import penetration, data on international trade from the “Base pour l'Analyse du Commerce International” (BACI) dataset (Gaulier

[25] Employment is used instead of sales because this data is available for all countries in the sample. Note that computing HHI requires data for the population of firms so, for some countries, this variable is computed from the production dataset or business register.

and Zignago, 2010[25]) is used along with a crosswalk from product to industry classifications from the OECD TiVA project (OECD, 2021[26]). Import penetration is defined as total imports of products relating to that industry (note these are final products, not intermediate inputs made by firms in this industry) divided by total output of the industry. Import competition could be associated with reductions in dispersion if less energy productive firms are forced to exit the market. On the other hand, it could increase dispersion if the most productive firms improve their productivity in response to the competition, as shown in Pavcnik (2002[27]), Trefler (2004[28]), or Bloom, Draca and Van Reenen (2016[29]), among others. There is no significant correlation between import penetration and dispersion in energy productivity, suggesting that these mechanisms may cancel each other out.

Another form of international openness is through FDI. Research shows that inward FDI can raise productivity of incumbents (see for example Javorcik (2004[30])), and FDI is likely to be targeted at more productive firms (Arnold and Javorcik, 2009[31]). The OECD produces measures of FDI restrictiveness by country, industry, and year (OECD, 2024[32]). The results show that more restrictive FDI measures is strongly correlated with less dispersion in energy productivity. The same relationship is seen in changes, although it is not statistically significant with large standard errors. This suggests that inward FDI does indeed promote the energy productivity of firms at the frontier.

A proxy of access to finance: cashflow to capital ratio

Access to finance can determine whether firms are able to invest in newer and potentially more energy-efficient capital. Numerous measures of access to finance have been used in the existing literature, often involving weighted indices (see, for example, Kaplan and Zingales (1997[33]) and Costa et al (2024[34])). For simplicity, this paper uses a simple measure that is often incorporated in the literature and is readily available: the ratio of cashflow to capital stock. For each industry, the median cashflow to tangible capital ratio is taken from the Bureau van Dijk Orbis database.

Higher cashflow is associated with less dispersion, suggesting that when firms are less financially constrained, there is a greater capacity to invest in more energy-efficient production methods. Furthermore, as industries become relatively less cash-constrained over time, dispersion tends to reduce.

A proxy of innovation at the frontier: Patenting intensity

While dispersion can arise from low productivity of laggard firms, it can also arise from leading firms pushing the productivity frontier. Patents are used a proxy of innovation at the frontier. Data on patents are from the OECD STI Microdata Lab infrastructure including industrial property data and links to company level data from Orbis. Patent data rely on the worldwide patent statistical database maintained by the European Patent Office (EPO), PATSTAT, in its Autumn 2024 edition.

Two measures of patenting are used. First is the share of firms filing patents, which captures the (inverse of) concentration of innovative activities. The results show that a higher share of firms with patents – corresponding to lower concentration of innovation – is associated with less dispersion. This is not statistically significant in levels but it is so in changes. The second measure is the total number of patents per firm, which proxies for the overall innovative intensity of the sector. Here there is no statistically significant relationship with dispersion.

7 Conclusions and policy implications

In a world of rising and increasingly volatile energy prices, energy productivity can play a key role for competitiveness. On top of that, it is a vital component of achieving the transition to net zero: while the energy sector can produce more carbon-efficient energy, industrial firms – which accounts for 23% of global emissions – must also use that energy more efficiently.

This paper highlights the extent to which adoption of more efficient production practices – that are already in place among leading firms – can make substantial gains towards reducing energy use. It develops unique cross-country data on the environmental and economic performance of firms to show that there is huge dispersion across firms in their energy productivity within narrowly defined industries: the 90th percentile firm is 21 times more energy productive than the 10th percentile firm on average. The extent of dispersion has not reduced over time and has even increased in some countries. Helping the least energy productive firms to improve their efficiency could substantially help improve energy independence and reduce emissions.

Furthermore, the results highlight the relationship between economic and environmental performance. Firms with higher labour productivity also tend to have higher energy productivity and improvements in labour and energy productivity are strongly correlated, even controlling for other factors. The paper shows that the most important factor predicting industry dispersion in energy productivity is dispersion in labour productivity, suggesting that the same structural factors may determine differences in both economic and environmental efficiency between firms.

In light of the huge dispersion between firms in energy productivity, policies should help improve the environmental performance of unproductive firms while increasing long term allocative efficiency. The evidence of this paper suggests that improving the energy productivity of the least productive firms to that of the 25th percentile firm in their industry would reduce energy use and carbon emissions by nearly a half to achieve the same economic output. The analysis also shows that policy-relevant factors including energy prices, competition, and access to finance appear to be important in determining differences in energy productivity between firms. Therefore, such policies could be effective tools to help catch up of unproductive firms and improvements in allocative efficiency of resources between firms. A holistic policy framework should aim to push the technological frontier while also boosting catch-up of unproductive firms and providing incentives to switch away from inefficient production processes.

First, as the evidence shows that increases in energy prices are correlated with a relative fall in dispersion, carbon pricing policies across the firm distribution may be effective in incentivising improved energy efficiency of less productive firms. The data also reveals that larger firms tend to be less energy productive, in stark contrast to the strong positive relationship between labour productivity and firm size. Part of the reason for this is that larger firms pay considerably lower prices for energy than small firms. Carbon and energy pricing policies must ensure that they do not accentuate this pattern.

Second, the evidence suggests that policies beyond carbon energy pricing can be effective in improving catch up of unproductive firms and improving allocative efficiency. Measures of competition are associated with less dispersion, suggesting that policies to promote healthy competition can encourage firms to become more energy productive. Industries in which firms are more cash-constrained on average tend to be more dispersed and there is evidence that firms with older capital stocks are less energy productive,

highlighting the importance of access to finance in energy efficiency upgrading. Finally, increases in the share of firms with patents is associated with falling dispersion, suggesting that policies relating to technology diffusion and expanding innovation across the firm distribution could help improve energy productivity across the distribution.

Third, these policies - relating to technology diffusion, competition, and access to finance – are regularly proposed to help promote labour or total-factor productivity. Furthermore, there is a strong relationship between environmental and economic performance of firms, as well as their dispersion at the industry-level. Together, the evidence suggests that there are likely to be synergies rather than trade-offs between environmental and economic policies, especially when aiming to reduce dispersion between firms by helping unproductive firms improve. Policies to improve labour productivity may have co-benefits for energy productivity and aggregate emissions. This also provides encouragement that the green transition can be achieved without being detrimental to aggregate economic performance.

References

Acemoglu, D., P. Aghion and F. Zilibotti (2006), "Distance to frontier, selection, and economic growth", *Journal of the European Economic Association*, Vol. 4/1, pp. 37-74, https://doi.org/10.1162/jeea.2006.4.1.37. [21]

Andrews, D., C. Criscuolo and P. Gal (2019), "The Best versus the Rest: Divergence across Firms during the Global Productivity Slowdown", *CEP Discussion Paper* 1645. [2]

Appelt, S. et al. (2022), *Micro-data based insights on trends in business R&D performance and funding: Findings from the OECD microBeRD+ project*, OECD Science, Technology and Industry Working Papers. [18]

Arnold, J. and B. Javorcik (2009), "Gifted kids or pushy parents? Foreign direct investment and plant productivity in Indonesia", *Journal of International Economics*, Vol. 79/1, pp. 42-53, https://doi.org/10.1016/j.jinteco.2009.05.004. [31]

Aulie, F. et al. (2023), "Did Covid-19 accelerate the green transition?", *OECD SCIENCE, TECHNOLOGY AND INDUSTRY POLICY PAPERS*. [1]

Barrows, G. and H. Ollivier (2018), "Cleaner firms or cleaner products? How product mix shapes emission intensity from manufacturing", *Journal of Environmental Economics and Management*, Vol. 88, pp. 134-158, https://doi.org/10.1016/j.jeem.2017.10.008. [22]

Berlingieri, G., P. Blanchenay and C. Criscuolo (2017), "The Great Divergence(s)", *OECD Science, Technology, and Innovation Policy Papers*. [12]

Berlingieri, G. et al. (2017), *The MultiProd Project: A Comprehensive Overview*. [5]

Berlingieri, G. et al. (2020), "Laggard firms, technology diffusion and its structural and policy determinants", *OECD Science, Technology and Industry* 86. [13]

Bijnens, G. and C. Swartenbroekx (2024), "Hunting "Brown Zombies" to Reduce Industry's Carbon Emissions", *Economie et Statistique*, Vol. 2024/544, pp. 27-44, https://doi.org/10.24187/ecostat.2024.544.2122. [7]

Bloom, N., M. Draca and J. Van Reenen (2016), "Trade induced technical change? The impact of chinese imports on innovation, IT and productivity", *Review of Economic Studies*, Vol. 83/1, pp. 87-117, https://doi.org/10.1093/restud/rdv039. [29]

Calel, R. and A. Dechezleprêtre (2016), "Environmental policy and directed technological change: Evidence from the european carbon market", *Review of Economics and Statistics*, Vol. 98/1, pp. 173-191, https://doi.org/10.1162/REST_a_00470. [35]

Calligaris, S. et al. (2024), "Measuring competition through different proxies: some OECD evidence", *Protecting competition in a changing world, EC report*, https://competition-policy.ec.europa.eu/system/files/2024-06/Exploring_the_evolution_and_the_state_of_competition_in_the_EU_launch.pdf (accessed on 5 August 2024). [24]

Calligaris, S. et al. (2023), *Employment dynamics across firms during COVID-19*. [19]

Calvino, F., C. Criscuolo and R. Verlhac (2020), “Declining business dynamism: Structural and policy determinants”, *OECD Publishing*, Vol. November/94, pp. 1-74. [16]

Calvino, F. and L. Fontanelli (2023), *A portrait of AI adopters across countries*. [17]

Capelle, D. et al. (2023), *Mitigating Climate Change at the Firm Level: Mind the Laggards*. [6]

Colmer, J. et al. (2020), *Does Pricing Carbon Mitigate Climate Change? Firm-Level Evidence From the European Union Emissions Trading Scheme Does Pricing Carbon Mitigate Climate Change? Firm-Level Evidence from the European Union Emissions Trading Scheme* *, https://ssrn.com/abstract=3917883. [40]

Correia, S. (2017), “Linear Models with High-Dimensional Fixed Effects: An Efficient and Feasible Estimator”. [43]

Costa, H. et al. (2024), “Making the grass greener: The role of firm’s financial and managerial capacity in paving the way for the green transition”*, OECD Economics Department Working Papers*, No. 1791, OECD Publishing, Paris, https://doi.org/10.1787/cdffe7eb-en. [34]

Criscuolo, C. et al. (2024), “Digital and green spillovers across firm-level production networks”, Forthcoming OECD Working Paper. [8]

Dechezleprêtre, A., D. Nachtigall and F. Venmans (2023), “The joint impact of the European Union emissions trading system on carbon emissions and economic performance”, *Journal of Environmental Economics and Management*, Vol. 118, https://doi.org/10.1016/j.jeem.2022.102758. [39]

Fiori, G. and F. Scoccianti (2025), *Aggregate Dynamics and Microeconomic Heterogeneity: The Role of Investment-Specific Technology*, http://www.giuseppefiori.net. [23]

Foster, L., J. Haltiwanger and C. Syverson (2008), “Reallocation, firm turnover, and efficiency: Selection on productivity or profitability?”, *American Economic Review*, Vol. 98/1, pp. 394-425, https://doi.org/10.1257/aer.98.1.394. [20]

Gaulier, G. and S. Zignago (2010), *BACI: International Trade Database at the Product-level The 1994-2007 Version*, http://www.cepii.fr/anglaisgraph/bdd/baci.htm. [25]

Hsieh, C. and P. Klenow (2009), *Misallocation and Manufacturing TFP in China and India*, https://academic.oup.com/qje/article/124/4/1403/1917179. [15]

Javorcik, B. (2004), “Does Foreign Direct Investment Increase the Productivity of Domestic Firms? In Search of Spillovers through Backward Linkages”, *The American Economic Review*, Vol. 94/3, pp. 605-627, http://www.jstor.org/stable/3592945. [30]

Kaplan, S. and L. Zingales (1997), “Do Investment-Cash Flow Sensitivities Provide Useful Measures of Financing Constraints?”, *Quarterly Journal of Economics*, Vol. 112/1, pp. 169-215, https://academic.oup.com/qje/article/112/1/169/1870889. [33]

Klenow, P., E. Pastén and C. Ruane (2024), *Carbon Taxes and Misallocation in Chile*. [11]

Levinson, A. (2015), “A direct estimate of the technique effect: Changes in the pollution intensity of US manufacturing, 1990–2008”, *Journal of the Association of Environmental and Resource Economists*, Vol. 2/1, pp. 43-56, https://doi.org/10.1086/680039. [38]

Lyubich, E., J. Shapiro and R. Walker (2018), “Regulating Mismeasured Pollution: Implications of [4]

Firm Heterogeneity for Environmental Policy", *AEA Papers and Proceedings*, Vol. 108, pp. 136-142, https://doi.org/10.1257/pandp.20181089.

OECD (2024), *OECD FDI Regulatory Restrictiveness Index: Methodological Framework*. [32]

OECD (2021), "OECD Inter-Country Input-Output Database". [26]

Pavcnik, N. (2002), *Trade Liberalization, Exit, and Productivity Improvements: Evidence from Chilean Plants*, https://academic.oup.com/restud/article/69/1/245/1584480. [27]

Petrick, S., K. Rehdanz and U. Wagner (2010), *Energy Use Patterns in German Industry: Evidence from Plant-level Data*, http://ssrn.com/abstract=1689044https://ssrn.com/abstract=1689044Electroniccopyavailableat:http://ssrn.com/abstract=1689044. [10]

Petrick, S. and U. Wagner (2014), *The impact of carbon trading on industry: Evidence from German manufacturing firms*, http://hdl.handle.net/10419/94357. [41]

Shapiro, J. and R. Walker (2018), *Why is pollution from us manufacturing declining? the roles of environmental regulation, productivity, and trade*, American Economic Association, https://doi.org/10.1257/aer.20151272. [37]

Syverson, C. (2011), *What determines productivity*, https://doi.org/10.1257/jel.49.2.326. [14]

Trefler, D. (2004), "The long and short of the Canada-U.S. free trade agreement", *American Economic Review*, Vol. 94/4, pp. 870-895, https://doi.org/10.1257/0002828042002633. [28]

Von Graevenitz, K. and E. Rottner (2020), *Energy use patterns in German manufacturing since 2003; Energy use patterns in German manufacturing since 2003*, http://hdl.handle.net/10419/214229. [9]

Wagner, U. et al. (2020), *Carbon Footprints of European Manufacturing Jobs: Stylized Facts and Implications for Climate Policy Carbon Footprints of European Manufacturing Jobs: Stylized Facts and Implications for Climate Policy* *, https://ssrn.com/abstract=3577550. [3]

Werner, S. (2016), "European District Heating Price Series", Energiforsk. [36]

Yamano, N. and J. Guilhoto (2020), "CO2 emissions embodied in international trade and domestic final demand: Methodology and results using the OECD Inter-Country Input-Output Database"*, OECD Science, Technology and Industry Working Papers*, No. 2020/11, OECD Publishing, Paris, https://doi.org/10.1787/8f2963b8-en. [42]

Annex A. Data collection procedure

Raw data sources

Chile

Data for Chile is obtained from the Encuesta Nacional Industrial Anual (ENIA) – the annual industrial survey. A publicly available version is used, covering 1995 to 2015, with establishment level information on balance sheet information and quantities of energy use by fuel type. Establishments have the same anonymised identifiers over time. Firm identifiers are provided, meaning that establishments can be aggregated to the firm level. The ENIA target population includes manufacturing establishments with ten or more workers but some establishments with fewer than 10 workers may be included if they are part of companies that perform more than one economic activity or have more than one plant. Beyond this threshold, the population of manufacturing establishments are covered. For confidentiality reasons some information is perturbed.

Croatia

Data for Croatia are from the Register of Annual financial reports of enterprises obtained from the Financial Agency (Financijska agencija, Fina). This is a mandatory administrative register: all legal and natural persons subject to corporate income tax must submit their balance sheets and profit and loss accounts, as well as some additional data such as on employment (GFI-POD forms).

The Fina data cover the whole non-financial business sector, and the statistical unit is the enterprise. An addendum to the administrative register requests voluntary information on total expenditure on energy used. Most firms provide this information, and those that do not respond are mainly very small (mostly one employee). The data cover 2002 to 2019.

France

Energy use data for France come from the Annual Survey of Energy Consumption in the Industry (EACEI) carried out by INSEE, the French statistical office.[26] The data cover the manufacturing sector and are based on a stratified sample for establishments between 20 and 250 employees; survey weights are provided. All establishments with more than 250 employees are included. From 2013 onwards, only establishments with more than 20 employees are covered (threshold of 10 employees up to 2012). The data used in the analysis cover the period 2005 to 2020.

It collects detailed information on the quantity of different energy types used, as well as the total expenditure and average price paid by type. A wide range of energy types is covered, of which electricity, natural gas, refinery gas, biofuels, heat, coal and coke. In the data construction, the energy variables are restricted to only include energy that is used by the firm for manufacturing or electricity generation purposes (which covers the majority of energy use).

[26] https://www.casd.eu/en/source/survey-on-energy-consumption-in-industries/

The energy data are collected at the establishment level. They are therefore supplemented with the business register of establishments, the Répertoire des entreprises et des établissements (REE), which contains the employment of all registered establishments and the corresponding identifier of the firm.[27]

Detailed balance sheet data are obtained for the universe of firms from Fichier complet unifé de SUSE (FARE) and Fichier approché des résultats Ésane (FICUS).

Indonesia

Statistics Indonesia conducts a survey of all manufacturing establishments with more than 20 employees, although the response rate varies year-on-year. The data contain detailed information on the quantities of different energy types used, as well as financial information of the establishment. Unfortunately, firm identifiers are not available, so the data are used at the establishment level (for all variables). Additional cleaning on the raw data is carried out for Indonesia to drop a small number of establishments with extremely volatile employment. The data cover 1999 to 2019, excluding 2016, but there is a trend break in 2016 so data for 1999-2015 is used for time series analysis .

Lithuania

Data for Lithuania come from the sampled part of the Structural Business Statistics. The data source is the questionnaire F-01 (annual enterprise activity report) administered by Statistics Lithuania. The survey population is comprised of all active public and private companies, state and municipal enterprises, foreign affiliates, all types of partnerships (with non-agricultural activity), individual enterprises, natural persons carrying out economic (commercial) activity (included from 2005) and public bodies covering the major share of their expenses with income from selling goods and providing services. The economic activities of the enterprises surveyed cover sections B-J, L-N and P-R and divisions A02, A03, S95 and S96 of EVRK Rev. 2, and data are collected at the firm-level. The variable used is total energy expenditure, and data between 2004 and 2020 are used.

As survey weights are not available in the raw data for all years, estimated survey weights are constructed based on the stratification system of the data collection method. These weights are refined by trying to match population aggregates on variables that are observed for all firms (such as employment).

Portugal

Energy data for Portugal is obtained from the Inquérito Anual à Produção Industrial survey. The survey provides information on products produced and sold, materials purchased and consumed, and energy used. The data covers around 6 000 companies per year in the manufacturing sector, with weights provided to ensure that the results are representative of the population. Energy use data is provided in quantity terms by energy type. The data is available from 2004 to 2021 but the period 2012-2014 is currently unavailable to researchers.

Alongside this, Portugal collects balance sheet data for the population of firms through the Sistema de Contas Integradas das Empresas (SCIE), which can be linked at the firm level to the energy use data. These are collected every year for the population of firms in the private sector, except for the financial sector, from 2006 to 2021, covering around 400 000 firms per year.

[27] https://www.insee.fr/en/metadonnees/source/serie/s1346

Sweden

Data for on energy use by type for Sweden is obtained from Industrins energianvändning (ISEN), collected by the Swedish Energy Agency. The data cover all manufacturing firms with at least 10 employees, over the period 2005-2021. The data is collected at the plant level and converted to the firm-level by the data owners. The data can be matched with Structural Business Statistics to obtain balance sheet information.

External datasets used to estimate emissions

Quantity of energy by industry

Data on quantities of energy by fuel type and industry are collected from two sources. First, Eurostat provides data on the quantity of energy used by type for each industry. The data is only available for manufacturing industries and has been collected since 2018, although many countries do not have data until later years. All quantities are converted into TJ using conversion factors provided by Eurostat. This data has the advantage of having very detailed energy types available but is not available for services industries, non-European countries, or for many years.[28]

Therefore, this is supplemented with data from the IEA, which includes the services sector (as a whole) and countries outside Europe, as well as being available for many years. The data collects information on six main energy types. These data are used as the main source of information on energy, and are complemented with the Eurostat data to split out the more aggregated energy types to more detailed energy types (e.g. gas to natural gas and refinery gas) using the shares of each sub-energy type from Eurostat to distribute the IEA values.[29] Specifically, in the Eurostat data, the sum of detailed energy types is taken within each match to a more aggregated energy type in the IEA data. Then the shares are computed of each detailed energy type in the Eurostat data. These shares are used in the IEA data to split out the more aggregated energy types.

Data on energy use by industry is not available from IEA or Eurostat for Indonesia but can instead be constructed from the microdata (and is anyway not necessary for the construction of the baseline energy and carbon measures because firm-level information is available).

Energy prices

Energy price data is available by country, energy type, and year. Four sources of data are used. First, IEA provides energy prices of many types of energy.[30] Second, Eurostat provides data on prices of electricity and gas faced by industry, with a breakdown of price according to the amount of energy consumed.[31] Third, Eurostat HICP provides a heat price index, but no indication of the level of the price for heat.[32] Fourth, data from Werner (2016[36]) is used to provide the level of heat price in a specific year, which can

[28] https://ec.europa.eu/eurostat/cache/metadata/en/nrg_quant_esms.htm

[29] https://www.iea.org/data-and-statistics/data-product/energy-efficiency-indicators

[30] https://www.iea.org/data-and-statistics/data-product/energy-prices

[31] https://ec.europa.eu/eurostat/cache/metadata/en/nrg_pc_202_sims.htm and https://ec.europa.eu/eurostat/cache/metadata/en/nrg_pc_204_sims.htm

[32] https://ec.europa.eu/eurostat/cache/metadata/en/cult_exp_esms.htm

be combined with the Eurostat HICP heat price index to produce the level of heat prices over the full sample period.[33]

Prices include taxes and levies, as is desirable for the analysis, and are converted into units of national currency per TJ using conversion factors from Eurostat and OECD exchange rate data. Eurostat data is available on a half-yearly basis: the average is taken over each year. The datasets are collated to provide panel data on energy prices by type and country.

Imputation method

There are missing values in the collated dataset. To impute values to fill these gaps, a weighted average across other countries' energy price is used. The inference is conducted for each energy type separately. The main rationale for using the prices of other countries to infer missing prices is that many energy prices are determined on a world market, meaning that the values of other countries is likely to give a better estimate of the true price than by, for example, simply making a linear interpolation.

The procedure uses the average value from a group of similar countries to impute missing values. To define this group of countries, optimal weightings are determined by comparing the prices of the target country with prices of other countries for the specific energy type of interest in the years where the target country has an available price. An algorithm (the Stata package "synth") is used to assign weights to all other countries based on their similarity to the target country. These weights are then applied to the years with missing values to impute as value as the weighted average across similar countries. If no data is available for the energy type-country in any year, the average weights constructed using other energy types are used.

For example, Croatia has data on electricity prices from 2005 onwards but not before 2004. For 2005 onwards, the values of Croatia's electricity price is compared with all other available countries' prices. The countries with the biggest weights for Croatia's electricity price are Germany (28.5%), Slovakia (19%), and Türkiye (16.7%).

The procedure is tested by setting some values to missing, applying the procedure to impute these values, and then comparing the imputed values to the true values. This testing suggests that the imputation procedure works well and provides a relatively accurate prediction of missing values.

Estimating price as a function of quantity consumed using available data

The price paid for each fuel type may depend on the quantity of that fuel type purchased. In particular, firms that consume larger quantities may get preferential rates. Therefore, using average prices that do not vary by quantity purchased may underestimate the true dispersion in emissions across firms.

Eurostat provides prices that vary by quantity consumed for electricity and natural gas – the two main energy types for most countries. For each year and country, the price is provided for different consumption bands. For example, in France in Q1 2019, the price per TJ of electricity was 51111 EUR if consuming less than 0.072 TJ of electricity, but was 15 916 EUR if consuming more than 540 TJ.

In practice, it is likely that, on average, the price that is paid for electricity or gas is a continuous function of quantity (instead of being fixed within a quantity band). Furthermore, to compute emissions, it is preferable to work with price as a continuous function of quantity. Therefore, prices are estimated as a continuous function of quantity, taking the mid-point of each quantity band. The price data is available on a half-yearly basis, so the average over each year is taken to be in line with other data used in the project.

[33] https://energiforskmedia.blob.core.windows.net/media/21926/european-district-heating-price-series-energiforskrapport-2016-316.pdf

Testing suggests that a log-log specification provides the best fit. Specifically, linear regression is used to estimate:

$$logP_{nct} = \alpha + \beta logQ_{nct} + \varepsilon_{nct}$$

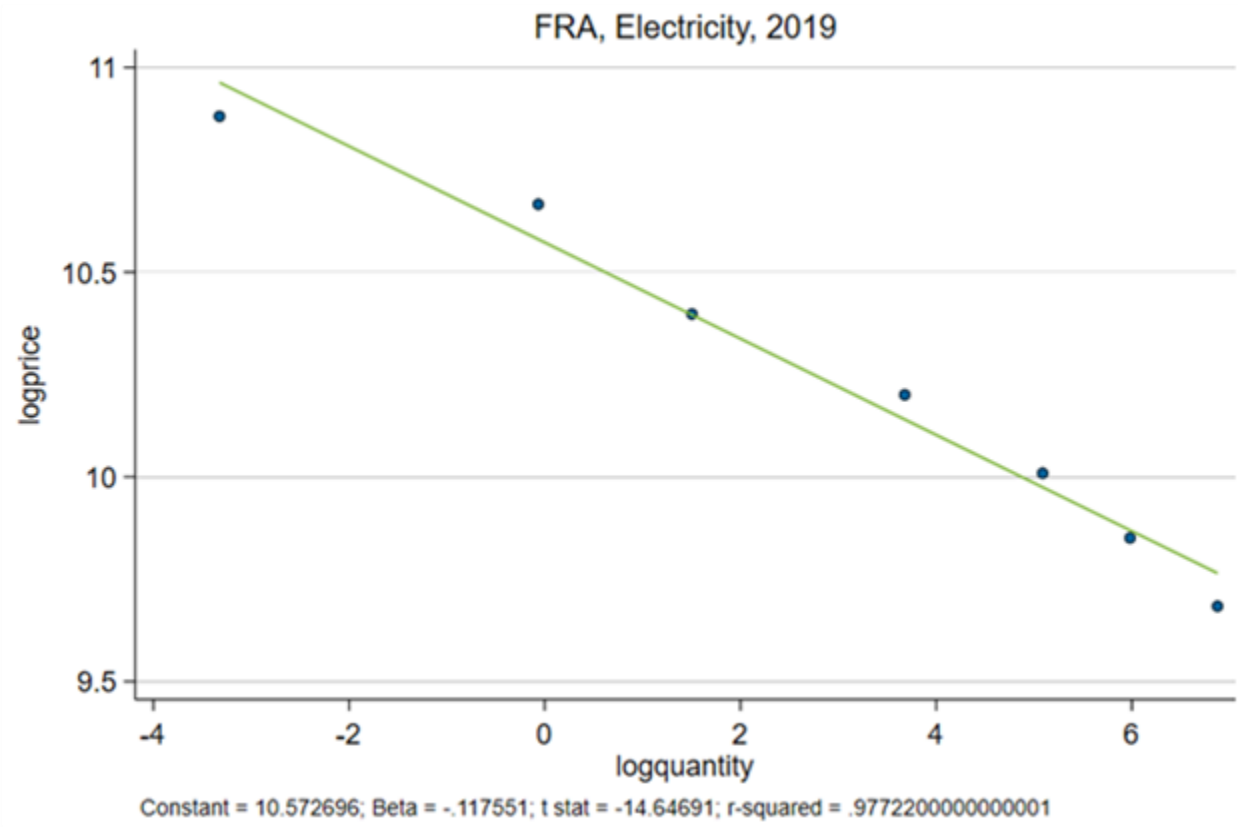
Equation 8

For each country, energy type, and year. In each regression, there are only between 5 and 9 data points because prices are only provided for this many quantity bands. Despite this, the regressions fit well. Finally, to avoid extrapolating extreme prices far outside of the observed price range, the price is restricted to be within the range of 0.1 times the minimum and 2 times the maximum actually observed in the data (i.e. even if a firm consumes a huge quantity of energy, it is not assumed to pay a price lower than the minimum price). These bounds are chosen for the minimum and maximum based on observations of prices in the French data, which were often below the minimum price observed in the Eurostat data, but less often observed to be greater than the maximum. Note that this characteristic of energy prices is further motivation for using the estimated monotonic relationship between price and quantity instead of simply using the bands, as it allows for lower prices than the minimum, which is commonly observed in the data.

The chart below shows this procedure for electricity in France in 2019. The R-squared is 0.977, suggesting that the model in logs provides a very tight fit.

Figure A A.1. Estimating price as a continuous function of quantity fits well

Log price of electricity against the log of the quantity band midpoints and associated regression line for electricity in 2019 for France



Note: The chart the relationship between log price and the log of the quantity band mid points. It also shows the associated OLS regression line and coefficients. The chart is for electricity in 2019 in France.
Source: Author's calculations using data from Eurostat

For countries where variable price data is missing, values are imputed using the same procedure described above. Furthermore, in cases where prices are variable in year t-h and t+l (for h,l>0), but not variable in period t, the coefficients are linearly interpolated (this applies to relatively few cases).

Finally, prices that vary with quantity are only available for electricity and natural gas, but the prices of all energy types are likely to vary with quantity. Therefore, the price gradients for electricity and gas are applied to other energy types within a country-year. First, the minimum price is compared with the average price for electricity and/or gas. If gradients are available for both then the average of this ratio is taken, otherwise it is taken for the only energy type with an available gradient. This ratio is then used to compute a minimum price for the new energy type using the available average price. The same is then done for the maximum. The gradient can then be applied to the new energy type, again taking the average slope across both electricity and gas if both are available. Doing this improves the estimation of true energy quantity and emissions in the French data. The same procedure can be applied to other countries in the sample.

Emissions factors

Emissions factors provide estimates of the amount of carbon dioxide (CO2), methane (CH4), and nitrous oxide (N2O) emitted from the consumption of a unit of each energy type. The emissions factors vary by country and over time. The data is provided by IEA.[34] The units are converted to tonnes of CO2 equivalent per TJ of energy, for each energy type. For electricity, the emissions factor represents the emissions from electricity produced from all sources, according to the production methods of that country. For fossil fuels, the emissions factor is from the direct combustion of that energy type.

As with the energy price data, there are some missing values. An equivalent method of imputation is used to fill these values, taking the weighted average across similar countries.

Estimating emissions

Estimating energy and emissions when only total energy spending is available

To outline the procedure, the following variables are defined:

S_{it} = total spending on energy by firm i, at time t. Each firm operates in an industry k and in a country c.

Q_{nkct} = quantity of fuel type n used by industry k in country c at time t. All quantities are in tera joules (TJ).

P_{nct} = price of fuel type n in country c at time t. This can also vary according to the quantity of the fuel type consumed, in which case it is denoted $P_{nct}(Q_n)$. All prices are for a tera joule of energy.

F_{nct} = emissions factor for fuel type fuel type n in country c at time t. Emissions can be computed as F×Q.

In all cases, the energy types included in the analysis are restricted to be those for which data is available on all variables. In practice, price data is available for the fewest energy types, so only the energy types with price data are included (these are coking coal, electricity, fuel oil, gas oil and diesel oil, heat, kerosene-type jet fuel, natural gas, other bituminous coal, other kerosene, and petroleum coke). Where possible, energy types that are included in the energy use data are matched to the most similar energy type in the price data (e.g. lignite to coking coal), making the assumption that they follow similar price trends. The shares of quantity consumed always sum to 1, so proportionality assumptions ensure that emissions are not underestimated due to these restricted energy types.

[34] https://www.iea.org/data-and-statistics/data-product/emissions-factors-2

First, the simple procedure used to estimate emissions using constant prices is outlined before the preferred procedure which implements an algorithm to estimate emissions using prices that vary by quantity consumed is described.

Using consumption-invariant prices

Basic data on energy prices are available at the energy type-country-year level, meaning that they do not depend on the volume of energy consumed. When these are used, it is still relatively straightforward to compute emissions. Total quantity consumed by the firm is estimated first. Note that the following identity for spending holds:

$$S = \sum_n Q_n \times P_n = Q \sum_n \frac{Q_n}{Q} \times P_n$$

Equation 9

Where the second equality holds because quantity and price are independent in this setting (i.e., this requires prices that do not depend on the quantity of energy consumed). Applying this to the available data, the total quantity of energy used by the firm is estimated as:

$$\widehat{Q_{it}} = \frac{S_{it}}{\sum_n \frac{Q_{nkc}}{Q_{nc}} \times P_{nct}}$$

Equation 10

Note that all estimated values are denoted with a hat. The denominator is the price index for a TJ of energy. This requires two assumptions:

Assumption (1): the firm's quantity shares of each energy type are assumed to be equal to the industry-country quantity shares.

Assumption (2): prices are determined at the national level (there is no variation in prices across firms).

Now that firm-level quantity has been estimated, total emissions $\widehat{M_{it}}$ can be computed as:

$$\widehat{M_{it}} = \widehat{Q_{it}} \sum_n \frac{Q_{nkc}}{Q_{nc}} \times F_{nct}$$

Equation 11

Using prices that vary with quantity consumed

The preferred estimate of emissions uses information on prices as a function of quantity for each energy type. Intuitively and in the data, firms that consume more energy pay a lower average price per unit. Therefore, ignoring the variation in prices would lead to an underestimate of energy consumed and of emissions for large firms and an overestimate for small firms.

The firm's quantity of each type using observed firm total spending, industry quantity shares by type, and prices must be solved. Recall that the equation for spending and quantity by type can be written as:

$$S = \sum_n Q_n \times P_n$$

Equation 12

The problem now is that the price of each energy type is a function of the quantity of that energy type, while the quantity consumed of that energy type is also a function of the price. Furthermore, the functions are non-linear and cannot be easily solved analytically.

To solve for the total quantity of energy used by each firm, the following maximisation problem is solved:

$$\max_{q} Q$$
$$s.t.$$
$$log P_n = \alpha_n + \beta_n log Q_n$$
$$S = Q \sum_n P_n \frac{Q_n}{Q}$$

Where the share of each quantity is fixed and is observed in the external data (but is updated ex-ante based on firm size) and the price parameters are determined by this estimation procedure Annex A with limits on the minimum and maximum price also imposed. The parameters in blue are fixed and observed in the data, while the variables in red are to be determined.

This maximisation problem finds the largest quantity for which the implied prices of each energy type are consistent with the total spending of the firm. Note that by maximising Q instead of minimising Q, it is assumed that firms behave optimally in the sense that they consume the maximum quantity of energy possible, given their spending. Put differently, there may be smaller values of Q that satisfy the constraints, but it is assumed that firms would not choose these combinations as they could consume more energy for the same total spending.

The parameters vary with each industry, country, and year. Therefore, solutions to the optimisation must be found for all values of total spending in each industry-country-year. The algorithm to estimate the quantity variable is as follows:

1. Order observations in decreasing order of total spending within an industry-country-year (biggest spending first).
2. Beginning with the firm with the largest expenditure.
    a. Propose that the firm pays the smallest possible prices (with bounds imposed on the minimum and maximum possible price based on the energy price data).
    b. Compute the largest possible quantity that could be consumed, given the firm's expenditure.
    c. If this quantity implies an expenditure larger than the firm's actual expenditure, reduce the value of the observed quantity.
    d. Continue to reduce the proposed quantity until implied expenditures equal observed expenditures.
3. Use the final quantity of the previous firm as the first proposed quantity of the firm with the next biggest expenditure.
    a. Reduce proposed quantity until implied spending equals observed spending.
4. Repeat this procedure for all firms.

To improve efficiency of the algorithm by minimising the number of computations, quantities are updated downwards in large jumps initially until the difference between implied and proposed expenditure becomes negative, then is updated upwards again until it is positive, and so on until convergence to a difference of 0.

In this setting, assumption (2) from before can be relaxed to:

Assumption (2*): Prices do not vary across firms, conditional on their quantity consumed.

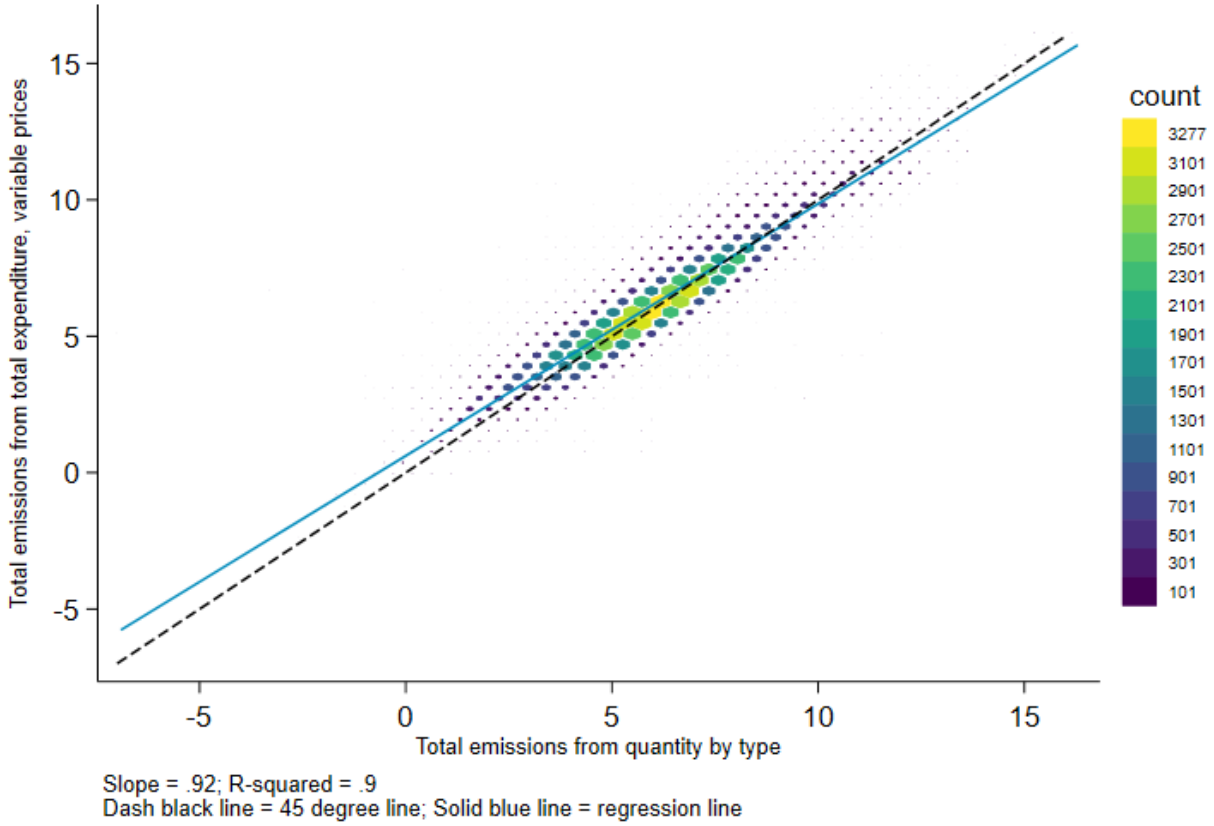
Two additional adjustments are also made to the quantity shares of each energy type $\frac{Q_n}{Q}$ using detailed French data. First, in some cases, quantity shares are only available at the 2 or 3 digit level. The French

data is used to compare the average share of electricity at the 2 (or 3, accordingly) digit level with the average share at the 4 digit level. This ratio is then used to adjust the 2 (or 3) digit share data for electricity in line with the ratio, and the inverse for fossil fuels. Second, even within 4 digit industries, there is some systematic variation in the types of energy that firms use, in particular with firm size. Therefore, a firm-level regression is estimated in France for each 2 digit industry, with the log of the electricity share as the dependent variable and the log of employment as the explanatory variable. The estimated coefficient is then applied to other countries to adjust up or down the electricity share, and the inverse for fossil fuels.[35]

Once total quantity by type has been computed, it is straightforward to estimate emissions using the same procedure as described above. The goodness-of-fit for the predicted energy quantity is presented in the main paper, along with a discussion of the sizeable improvement in the estimation of energy quantity compared with simply using fixed energy prices. For emissions, the goodness-of-fit is presented below.

Figure A A.2. Energy emissions is similar to true emissions when estimated using total energy spending

Log emissions from total energy spending against log emissions from data on energy by type, establishment level data for France



Note: The chart plots the log of total emissions estimated using only total expenditure on energy (y axis) against the total emissions estimated from detailed information on quantities used of different fuel types and electricity (x axis). The dashed black line is the 45-degree line while the blue line is the fitted OLS regression estimate from y on x. The heat map shows the underlying data points where the size and colour of the areas represent the number of underlying data points. The data is establishment level information from France pooled over 2002 to 2020.

[35] Alternative regression specifications were explored with more explanatory variables but this simple version seemed to be most accurate. Only parameters on the electricity share are used to reduce issues with adding up constraints after the adjustments.

Source: authors' calculations.

Industry code conversion

Most OECD Member countries changed from using the ISIC revision 3.1 to ISIC revision 4 industry classification system around 2008. The correspondence between ISIC rev. 3.1 and ISIC rev 4 is many-to-many (m:m), so it is not straightforward to obtain a consistent industry classification going back before 2008. Therefore, a conversion procedure in the data is used. The industry conversion procedure uses firm-level panel data to identify the relevant correspondences between the two systems.

First, changeover proportions are produced by taking firms observed in the data in both the year before the changeover and the year of the changeover. This produces a m:m crosswalk. If there are any industries for which a crosswalk cannot be created (because there are no panel firms in those industries), the crosswalk is supplemented with an external crosswalk taken from the MultiProd project (Berlingieri et al., 2017[5]).

Then an industry code is assigned according to the following procedure:

1. If the firm already has the new industry code, this is the baseline industry code.
2. If the proposed match from the crosswalk is unique, then the uniquely proposed industry code is used.

If panel data is available for a firm and it is observed to transition to a stable industry code in the ISIC rev 4 classification, assign this as the industry code for the firm in earlier years so long as the industry code in the earlier years has a possible correspondence to the later industry code.

   a. For example, suppose the ISIC rev 3.1 industry A maps to an ISIC rev 4 industry Y with 90% probability and Z with 10% probability. A firm is observed with industry A in 2007 and industry Z from 2008 onwards. Then this firm is assigned to have industry Z in 2007. However, if in 2005 the same firm is observed with an ISIC rev 3.1 industry code B, which only maps to industry X, the firm would be assigned industry X.

   b. This procedure aims to use information about the firm. For example, publishing from ISIC rev 3.1 splits into a manufacturing sector (Printing and reproduction of recorded media) and a services sector industry (publishing activities). If a firm is observed to be in the services sector after 2008 in ISIC rev 4, it makes sense to assume that they were in the services sector before, rather than going straight for probabilistic matching.

3. Finally, probabilistic matching is applied using the crosswalk constructed. The firm is assigned a new industry code probabilistically in the first year in which it is observed without the new industry code. Then the remaining years of the firm are filled, keeping the industry code constant. A seed is set to ensure that the same probabilistic mapping is applied in each run of the code.

The most common industry observed for each firm is assigned to the firm throughout the entire timespan. This ensures that industry trends are not driven by firms switching in and out of the industry classification. Of course, the disadvantage is that there may be true industry switches that are held constant.

Industry variables may be included in each data source. Therefore, a preferred industry code must be chosen. The first choice is the business register, followed by the production survey, and finally the environmental survey.

Converting establishment to firm-level data

The analysis of this project is conducted at the firm-level while some countries (including France) collect environmental data at the establishment level. It is therefore necessary to aggregate establishment level data to the firm-level.

If all the establishments within a firm are included in the sample, then it is straightforward to aggregate variables to the firm-level by simply taking the sum (because all variables considered by the module are additive). However, not all establishments within a firm may be included in the environmental survey. In this case, simply aggregating the observed values from the establishment level data would lead to an underestimate of firm-level values.

To overcome the issue of incomplete coverage of establishments within a firm, the employment covered by surveyed establishments is compared to the employment of the firm. There are numerous possible methods to obtain a value for the share of employment covered in the environmental survey, depending on the data sources available. If an establishment level business register is available, then the environmental survey is matched to the register at the establishment level and the share of employment covered it computed. Otherwise, the total employment of establishments in the environmental survey is computed and compared to the best available data source for firm-level employment (production survey or firm-level business register).

The total values (e.g. in energy quantity or emissions) of the establishments in the environmental survey are then scaled up by the ratio of the firm's employment to the sum of establishment employment. For example, if the total employment of establishments is 50 while the firm's employment is 100, the sum of energy and emissions from the environmental survey would be scaled up by a factor of 2 for this firm. The scaling factor is bounded to be at least 1 (so that the sum of establishment values within a firm is never scaled down) and at most 100 (so that the values are not scaled by a huge value). Note that if establishment level values are scaled by an implausibly large factor, the outlier detection methodology described below will smooth or remove implausible outlier values by comparing with other variables available at the firm-level.

The key assumption of this method is that all environmental variables can be scaled proportionally to employment shares. Testing by randomly dropping establishments within a firm for which all establishments are observed (checked in the business register data) and conducting the methodology, then comparing results against the firm's true value, suggests that this performs well.

Method for converting sample weights to the firm-level

While most variables, such as energy, can simply be summed across establishments, it is not accurate to simply sum weights across establishments. Furthermore, it would not be accurate to take the average of weights. To see this, consider the example below of a firm with two establishments, where the sample weight is an inverse probability weight, and suppose this is the only firm in the sample which is representative of the full economy. To obtain the total weighted energy of the aggregate economy from the sampled establishment data, the computation would be 1 * 100 + 10 * 10 = 200, where the first number in each multiplication is the weight and the second is the observed value of energy. Once the data is aggregated to the firm-level, if the average or sum of the sample weights is taken, the estimate of the total economy energy would be wrong (it would be estimated as 5.5 * 110 = 605 or 11 * 110 = 1210, respectively).

The proposed solution to aggregate weights from the establishment to the firm-level is the following. Suppose a sample weight w_s is observed for establishment s. To obtain an accurate sample weight at the firm-level i for a target variable x_s (in the above example the target variable is energy), the following formula is used:

$$w_{x,i} = \sum_{i \in f} \frac{x_s}{\sum_{s \in i} x_s} w_s$$

Equation 13

This formula corresponds to the within-firm target variable weighted average of the establishment weights. Returning to the example from above, the firm's weight for energy would be (100/110)*1 + (10/110)*10 = 20/11. Then, it is possible to obtain the total energy of the economy using this weight: 110*20/11 = 200.

Table A A.1. Illustrative example for aggregating establishment weights to the firm-level

Establishment ID	Firm ID	Sample weight	Energy (establishment)	Energy (firm)
1	1	1	100	110
2	1	10	10	110

This method can also be used consistently with the scaling described above. Under the assumption required for the scaling to be accurate – that the target variables increase proportionally with employment – the weight scaling method still obtains macro aggregates. To see this, assume that only establishment 2 is observed in the environmental survey data so that establishment 1 has a sample weight of 0 (because it is not in the sample) while the sample weight of establishment 2 is increased to 20. This is shown in the example below. The true total energy in the economy is still 200. Assume that the perfect scaling factor of 11 is observed so that energy can be accurately scaled to the firm-level. Then, the sample weight of the firm is 20 * 10 / 110 = 20/11, as before.

Table A A.2. Illustrative example for aggregating establishment weights to the firm-level with scaling

Establishment ID	Firm ID	Sample weight	Energy (establishment)	Energy (firm)
1	1	0	100	110
2	1	20	10	110

Detection of extreme outliers to reduce data entry errors

For some countries, the raw data cannot be directly accessed. Therefore, it is important to include procedures to detect and clean extreme outliers on the firm-level data as these may be data entry errors, and these may be particularly common in energy use surveys where the units of measurement are variable, and conversions can be complex.

Significant errors in data entry, such as entering a value in single units when it is supposed to be entered in thousands, can lead to major incorrect outliers that significantly affect results. To try to detect such errors, two methods are used. First, and most reliably, the within-firm time series is used to detect major jumps in specific variables. Second, extreme cross-sectional outliers are detected by comparing observations with other firms within an industry in each year. The procedures are described in detail in this sub-section.

The aim of these procedures is to detect data entry errors while not affecting genuine outliers. Indeed, outliers may be some of the most important observations economically and should not be dropped. Yet, given the nature of the survey data and the decentralised manner in which the codes are run, it is important to impose some provisions for errors.

Within-firm detection

With sufficient firm-level panel data, it is possible to identify implausible jumps in variables within a firm over time. Numerous measures are taken to try to avoid incorrectly identifying actual changes in a firm's activities as an error. The within-firm detection is carried out on a firm-by-firm basis, meaning that detection for one firm is independent from that of another.

Since this method relies on identifying sharp changes within a firm, it is necessary to have a sufficiently long and consistent panel of observations for the firm's potential errors to be identified. This requires two conditions. First, there must be a minimum number of periods surrounding an observation. On this, it is required that there is at least one period before and three periods after an observation (adjustable), or vice versa, to identify whether that observation is a significant error (therefore, at least four comparison periods total). The reasoning for requiring periods both before and after the observation in question is that firms are likely to have extreme outliers in their first and last year of operation, so this should not be incorrectly identified as an error. There are two exceptions to this. First, it is not necessary to observe a period before the observation is if the firm is observed to be older than one year, so it is not a new firm, in which case the first observation is still checked, so long as there are at least three observations after. The second exception is that, in the last period of the entire sample, outliers are still checked. The second requirement is that the observed periods must not be too many years apart. Specifically, if consecutive observations are more than three years apart (adjustable), they are not used to identify errors, as it is plausible that the firm may have changed significantly in this time.

Figure A A.3 provides four illustrative examples. In the top left panel, firm 1 is only observed in two periods, so neither period is checked using within-firm detection. In the top right panel, the firm's first two observations are many years apart, so cannot be checked. The observation in 2015 can be checked as the firm is at least two years old, and there are three periods observed after. Beyond that, it is not possible to check as there are not at least three periods either side of the observations. In the bottom left panel, the firm is known to have been born in the first period, so it is not checked. After that, it is checked in every period except the last, when it may have ended existence. In the bottom right panel, the firm is known to be 25 years old in the first period so it is checked, along with all others until the final period (where the sample of data covers periods later than 2008).

Figure A A.3. Illustrative examples of sample requirements for within-firm outlier detection

Firm ID	Year	Firm Age	Identify?
1	2007	1	No
1	2008	2	No

Firm ID	Year	Firm Age	Identify?
2	2000	1	No
2	2005	6	No
2	2015	16	Yes
2	2016	17	No
2	2017	18	No
2	2018	19	No

Firm ID	Year	Firm Age	Identify?
3	2002	1	No
3	2003	2	Yes
3	2005	4	Yes
3	2006	5	Yes
3	2007	6	Yes
3	2008	7	No

Firm ID	Year	Firm Age	Identify?
4	2002	25	Yes
4	2003	26	Yes
4	2005	28	Yes
4	2006	29	Yes
4	2007	30	Yes
4	2008	31	No

For the observations that are identified as candidates for checking within-firm errors, the procedure used relies on comparing the value of a variable with the moving average of that variable in the consecutive periods (excluding the contemporary observation). Furthermore, to identify a variable as an error, it must not only be an extreme outlier in its level but must also generate extreme outliers in an associated ratio (for example, if there is a sharp decrease in energy used that happens at the same time as a sharp decrease in the number of employees or revenue, the sharp drop in energy used would not be considered as an error).

Three more conditions are required for an observation to be considered an outlier. First, there must not be a clear within-firm trend break downwards. This aims to avoid incorrectly identifying firm observations as outliers when they have simply downsized. This was identified in the data as relatively common, while corresponding sudden jumps upwards were not generating issues for the outlier detection. Second, outliers are not detected if there is a persistent negative growth rate of the variable. Third, if more than half of the observations are zero, then it is not inspected for outliers.

Figure A A.4 provides an illustrative example. For firm 5 in the top panel, value added (VA) is relatively constant over time. Focusing on 2005, if VA is compared with the moving average (MA) of value added in neighbouring periods, the values are similar. However, its energy (ENQ) in 2005 is less than 1% of the size of the associated moving average. This imbalance is identified by looking at the ratio of VA/ENQ, which jumps sharply upwards in 2005. Therefore, it is likely that the value for ENQ was entered with error.

In contrast, in the second panel, the entry for value added also dropped sharply in 2005. Therefore, even though there is the same sharp drop in energy as before, the ratio of value added to energy is relatively similar to the moving average. It is feasible that, in this year, the firm simply downsized its operations, so this entry would not be considered as an error.

Figure A A.4. Illustrative examples of within-firm outlier detection

Firm ID	Year	VA	MA of VA	ENQ	MA of EN(	VA/ENQ	MA of VA/ENQ
5	2002	10000		2000		5.00	
5	2003	11000		2500		4.40	
5	2004	10500		2200		4.77	
5	2005	12000	11000	22	2400	545.45	4.615545
5	2006	11000		2300		4.78	
5	2007	11500		2500		4.60	
5	2008	12000		2900		4.14	

Firm ID	Year	VA	MA of VA	ENQ	MA of EN(	VA/ENQ	MA of VA/ENQ
6	2002	10000		2000		5.00	
6	2003	11000		2500		4.40	
6	2004	10500		2200		4.77	
6	2005	150	11000	22	2400	6.82	4.615545
6	2006	11000		2300		4.78	
6	2007	11500		2500		4.60	
6	2008	12000		2900		4.14	

It is necessary to specify a parameter beyond which the variable is considered to be sufficiently different from the within-firm moving average to consider it a potential error (for both the level and the ratio). After testing and eyeballing the data, a parameter of 15 is chosen. If a variable x is at least a factor of 15 different

from the moving average, and a key ratio involving X is also 15 times different from its respective moving average, then the variable x is considered to be an error. Then, x is adjusted to be equal to its moving average and all corresponding ratios and log values are also updated.

Specifically, the following two conditions are used to identify the distance of variable x from its moving average, ma, for the distance parameter α. The observation x_{it} is considered an error if either:

$$abs\left\{\frac{(x_{it} - ma(x_{it}))}{ma(x_{it})}\right\} > \alpha$$

Equation 14

$$abs\left\{\frac{ma(x_{it})}{x_{it}}\right\} > \alpha \text{ if } sign(x_{it}) = sign\big(ma(x_{it})\big) \text{ and } abs(x_{it}) < abs\big(ma(x_{it})\big)$$

Equation 15

The first condition detects large absolute differences of x from its moving average and is robust to the two variables having different signs. The second condition identifies cases where the value of x is very small relative to the moving average, for example if it had been entered in millions instead of single units.

The procedure is implemented on just one observation per firm at a time. That is, only the largest outlier for each firm is identified and smoothed, if necessary. Then the procedure is repeated, identifying just one outlier at a time using the updated dataset, until no more outliers remain.

There are, of course, limitations to this approach. It is only possible to carry out the procedure on firms with panel data. In surveys with stratified samples, it is more likely to observe a sufficient panel for larger firms which, in turn, may be less likely to submit data with errors.

Cross-sectional outlier detection

Implausible values can also be identified by comparing values across firms within an industry. This is particularly useful if firm-level panel data is not available, as is often the case with environmental surveys. To detect data entry errors, the value of a variable is compared with the difference between the 25th and 75th percentile value of that variable within the firm's industry in each year (the Tukey method).

Like with the within-firm detection, it is important to avoid incorrectly identifying genuine outliers as data entry errors. In particular, when trying to identify outliers, it is important not to incorrectly identify particularly large firms as outliers. Therefore, to identify a variable as a data entry error and not just an outlier, it must be that the value of the level of the variable and the ratio of the variable with respect to the firm's employment are sufficiently different from the respective within-industry 75-25 percentile ranges.

Firms may have plausible values of two variables relative to other firms but the comparison of these two variables within the firm may be implausible. For example, value added of hundreds of millions is plausible, as is employment of just five, but a firm with both of these together may be implausible. Therefore, ratios of key variables relative to both labour and energy are checked. Firm-year observations are dropped if these ratios exceed the specified plausibility threshold for the industry. In practice, this mainly drops observations with implausibly small values of energy which, from examination of the data, most likely to be data entry errors.

If a variable is detected as being a data entry error, the observation is dropped. The variables on which this procedure is implemented, if available, are gross output; value added; energy expenditure; carbon dioxide emissions; total environmental investments; total environmental spending; and total pollution. The parameter which determines whether a variable is sufficiently different from the within-industry 75-25 percentile range to be considered an error is 50 times larger or smaller.

Confidentiality

The program can automatically apply three forms of confidentiality. The user can specify that any or all of these should be applied.

First, the program can apply a threshold rule based on the number of units per cell. If the minimum elements per cell option is specified, the code blanks cells for statistics based on fewer than X units, where X is specified by the user, for example, automatically blanking cells with fewer than five firms. If the raw data is at the establishment level, then the count of establishments is used for the cell.

Secondly, the program can implement a dominance rule. If the user chooses to apply dominance, the program blanks cells in which the largest firm or the largest two firms account for more than Y% of the cell's total, in terms of employment (L), output (GO), or on a variable-by-variable basis. The code program can also implement blanking based on the P-percent rule. To allow for the possibility of negative value added, the p-percent rule is applied after converting to absolute values.

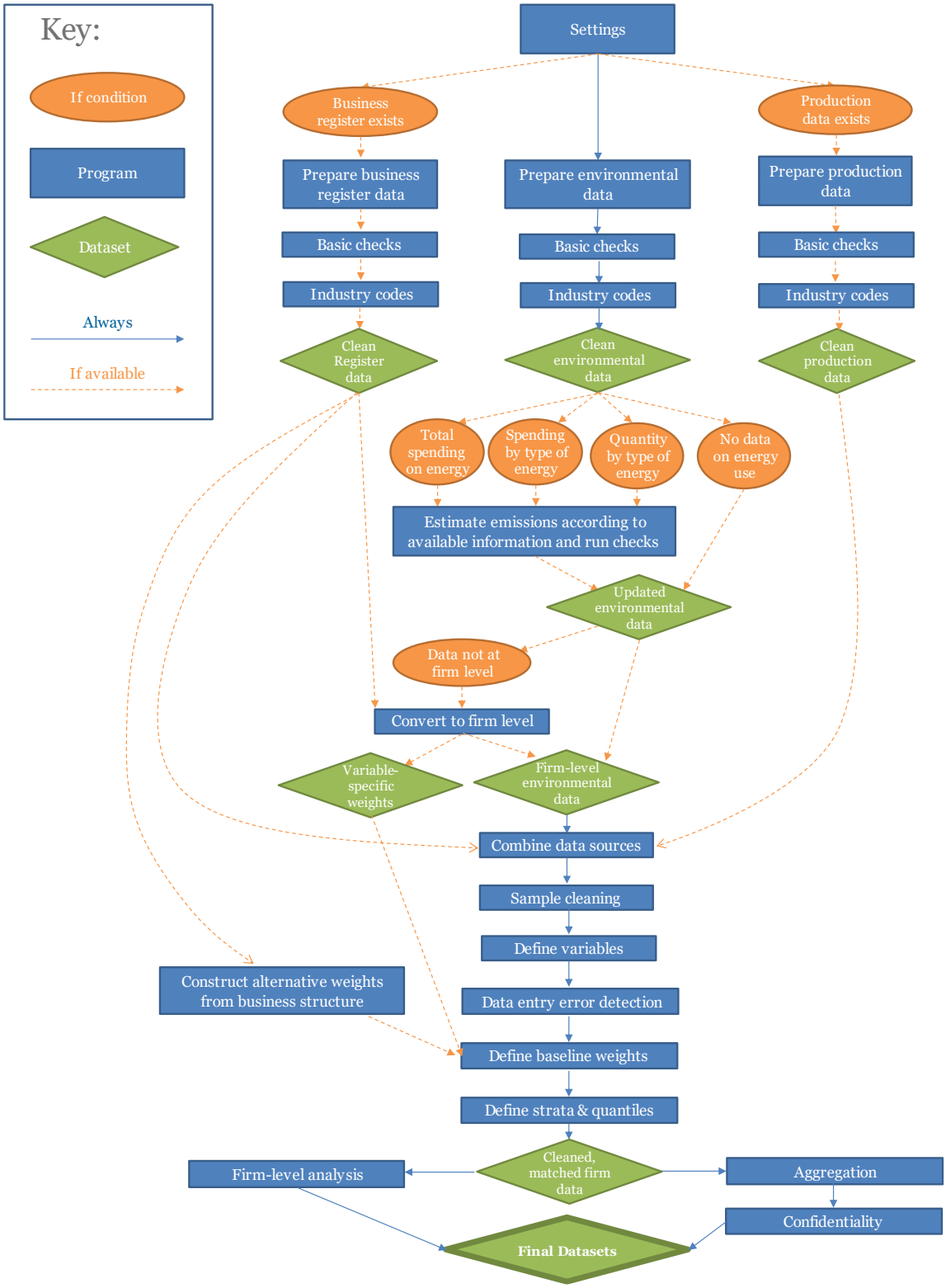
Third, some countries, though not any of those currently included in the project, do not allow specific percentiles of the data to be released. In these cases, percentiles are replaced with the average over the nearest X observations, where X is a number to be determined by the country.

Table A A.3. Confidentiality requirements by country

Country	Min units per cell	Dominance
Chile	-	-
Croatia	-	-
France	3 (establishments)	1 unit, 85%.
Indonesia	-	-
Lithuania	-	-
Netherlands	10	15% for p-percent rule based on top two firms
Portugal	-	-
Sweden	3	1 unit, 50%. 2 units, 69%
United Kingdom	10	1 unit, 43.5%, plus 12.5% p-percent rule

Diagrammatic representation of the procedure

Figure A A.5. Diagrammatic representation of the data collection process



Annex B. Additional analysis

Table A B.1. Firm-year-level summary statistics for Chile

variable	mean	sd	p10	p50	p90	count
Employment	119.46	259.95	23	48	254	61798
Value added (millions)	4.45	27.76	0.13	0.64	6.71	61798
CO2 Emissions (tonnes)	2620.69	26123.02	23.45	180.6	2640.39	61798
Log Value added / energy quantity	12.72	1.44	10.89	12.75	14.5	59895
Log Value added / CO2	8.3	1.38	6.57	8.34	9.99	59895
Log Value added / labour	9.54	0.94	8.5	9.5	10.67	59895
Energy spending / intermediate inputs	0.03	0.05	0	0.01	0.07	61768
Energy spending / total costs	0.02	0.05	0	0.01	0.07	61776
Intermediate input costs / total costs	0.89	0.19	0.59	1	1	61782
Age of Capital	2.86	2.8	1	2	7	60058
Electricity quantity share (share of joules)	0.11	0.25	0	0	0.51	61798

Note: The table reports summary statistics of firm-level data for manufacturing firms with at least 20 employees in Chile. The statistics are unweighted and cover close to the population of firms. Note that intermediate inputs includes energy spending.
Source: authors' calculations

Table A B.2. Firm-year-level summary statistics for Croatia

	mean	sd	p10	p50	p90	count
Employment	126.1	334.9	22	48	258	25907
Value added (millions)	5.32	38.04	0.32	1.24	8.07	25907
CO2 Emissions (tonnes)	2215.22	19693.27	37.43	296.15	3498.78	25332
Log Value added / energy quantity	12.58	1.29	11.1	12.52	14.13	25144
Log Value added / CO2	8.41	1.3	6.92	8.36	9.98	25144
Log Value added / labour	10.06	0.79	9.16	10.04	11.02	25719
Firm age	14.17	8.58	3	13	25	20279
Energy spending / intermediate inputs	0.4	0.25	0.1	0.35	0.79	20260
Energy spending / total costs	0.31	0.21	0.08	0.26	0.62	24725
Intermediate input costs / total costs	0.65	0.23	0.27	0.71	0.89	25900
Labour costs / total costs	0.35	0.23	0.11	0.29	0.74	25907
Age of Capital	3.73	3.28	1	3	9	25458

Note: The table reports summary statistics of firm-level data for manufacturing firms with at least 20 employees in Croatia. The statistics are unweighted and cover close to the population of firms. Note that intermediate inputs includes energy spending.
Source: authors' calculations

Table A B.3. Firm-year-level summary statistics for France

variable	mean	sd	p10	p50	p90	count
Employment	239.27	1989.92	26	79	432.02	105741
Value added (millions)	21.49	115	1.32	5.13	36.61	99049
CO2 Emissions (tonnes)	6455.63	103852.37	23.84	275.47	4756.76	105741
Log Value added / energy quantity	13.3	1.36	11.59	13.36	14.93	98130
Log Value added / CO2	9.89	1.52	7.98	9.95	11.72	98125
Log Value added / labour	11.1	0.59	10.47	11.09	11.77	98151
Firm age	30.01	21.11	8	26	54	99042
Energy spending / intermediate inputs	0.04	0.06	0	0.02	0.08	98652
Energy spending / total costs	0.02	0.04	0	0.01	0.05	98761
Intermediate input costs / total costs	0.7	0.14	0.52	0.72	0.87	98855
Labour costs / total costs	0.3	0.14	0.13	0.28	0.48	98855
Age of Capital	5.2	4.05	1	4	11	105741
Electricity quantity share (share of joules)	0.63	0.29	0.23	0.63	1	100986

Note: The table reports summary statistics of firm-level data for manufacturing firms with at least 20 employees in France. The statistics are weighted by the sample weight. Note that intermediate inputs includes energy spending.
Source: authors' calculations

Table A B.4. Establishment-year level summary statistics for Indonesia

variable	mean	sd	p10	p50	p90	count
Employment	165.13	650.14	21	42	318	386340
Value added (millions)	3.84	26.48	0.03	0.25	6.22	386340
CO2 Emissions (tonnes)	2603.61	23142.29	6.41	127.09	3289.81	386340
Log Value added / energy quantity	12.77	1.65	10.74	12.75	14.81	385961
Log Value added / CO2	7.8	1.64	5.76	7.79	9.85	385961
Log Value added / labour	8.59	1.44	6.81	8.53	10.47	385961
Firm age	5.36	4.79	0	4	13	386340
Energy spending / intermediate inputs	0.53	0.3	0.12	0.53	0.94	383183
Energy spending / total costs	0.21	0.19	0.03	0.14	0.48	382775
Intermediate input costs / total costs	0.4	0.25	0.09	0.37	0.78	383208
Labour costs / total costs	0.64	0.71	0.23	0.64	0.92	386340
Age of Capital	3.37	2.94	1	2	7	239238
Electricity quantity share (share of joules)	0.73	0.44	0	1	1	369045

Note: The table reports summary statistics of establishment level data for manufacturing establishments with at least 20 employees in Indonesia. Note that intermediate inputs includes energy spending.
Source: authors' calculations

Table A B.5. Firm-year level summary statistics for Lithuania

variable	mean	sd	p10	p50	p90	count
Employment	126.08	203.86	25	64	281	13512
Value added (millions)	2.85	10.5	0.18	0.98	5.61	13512
CO2 Emissions (tonnes)	1673.31	17671.55	22.1	147.54	2091.81	12919
Log Value added / energy quantity	12.42	1.36	10.69	12.43	14.13	12792
Log Value added / CO2	8.65	1.43	6.81	8.69	10.45	12792
Log Value added / labour	9.55	0.85	8.47	9.55	10.61	13383
Firm age	13.21	6.44	5	13	22	13512
Energy spending / intermediate inputs	0.07	0.09	0.01	0.04	0.17	13503
Energy spending / total costs	0.05	0.05	0.01	0.03	0.1	13512
Intermediate input costs / total costs	0.72	0.18	0.48	0.77	0.91	13508
Labour costs / total costs	0.28	0.18	0.09	0.23	0.52	13512
Age of Capital	4.74	4.09	1	3	11	13468

Note: The table reports summary statistics of firm-level data for manufacturing firms with at least 20 employees in Lithuania. The statistics are weighted by the sample weight. Note that intermediate inputs includes energy spending.
Source: authors' calculations

Table A B.6. Firm-year level summary statistics for Netherlands

variable	mean	sd	p10	p50	p90	count
Employment	119.46	326.21	22.7	50	225	33600
Value added (millions)	15.54	119	1.16	3.87	23.65	33615
CO2 Emissions (tonnes)	4886.14	59392.67	0	196.59	3436.86	26839
Log Value added / energy quantity	13.74	1.27	12.1	13.82	15.26	22396
Log Value added / CO2	9.44	1.25	7.84	9.49	10.93	22340
Log Value added / labour	11.26	0.74	10.58	11.25	12.06	32913
Firm age	31.43	18.91	7	31	53	33616
Energy spending / intermediate inputs	0.03	0.05	0	0.01	0.06	32083
Energy spending / total costs	0.02	0.03	0	0.01	0.04	32083
Intermediate input costs / total costs	0.72	0.16	0.52	0.74	0.89	33614
Labour costs / total costs	0.28	0.16	0.11	0.26	0.48	33615

Note: The table reports summary statistics of firm-level data for manufacturing firms with at least 20 employees in Portugal. Note that intermediate inputs includes energy spending.
Source: authors' calculations

Table A B.7. Firm-year level summary statistics for Portugal

variable	mean	sd	p10	p50	p90	count
Employment	83.71	145.91	23	45	165	52647
Value added (millions)	3.05	10.37	0.4	1.2	5.54	52370
CO2 Emissions (tonnes)	1773.62	47770.92	23.43	141.22	1987	52647
Log Value added / energy quantity	13.4	1.41	11.64	13.52	15	52003
Log Value added / CO2	8.96	1.38	7.26	9.04	10.58	52003
Log Value added / labour	10.14	0.61	9.43	10.13	10.88	52003
Firm age	7.18	5.49	1	6	15	52647
Energy spending / intermediate inputs	0.08	0.13	0.01	0.03	0.17	50956
Energy spending / total costs	0.04	0.08	0	0.02	0.08	52325
Intermediate input costs / total costs	0.62	0.22	0.32	0.66	0.88	52367
Labour costs / total costs	0.38	0.22	0.12	0.34	0.68	52370
Age of Capital	4.79	4	1	4	12	48909
Electricity quantity share (share of joules)	0.84	0.28	0.31	1	1	49568

Note: The table reports summary statistics of firm-level data for manufacturing firms with at least 20 employees in Portugal. Note that intermediate inputs includes energy spending.
Source: authors' calculations

Table A B.8. Firm-year level summary statistics for Sweden

variable	mean	sd	p10	p50	p90	count
Employment	135.9	606.45	23	45	227	50713
Value added (millions)	16.58	123	1.35	3.51	22.4	50713
CO2 Emissions (tonnes)	2044.62	33261.38	5.56	67.68	1222.16	50713
Log Value added / energy quantity	13.42	1.26	11.89	13.43	14.95	50318
Log Value added / CO2	10.99	1.69	8.89	10.97	13.12	50318
Log Value added / labour	11.25	0.49	10.76	11.22	11.81	50318
Firm age	22.41	8.88	9	23	33	50713
Energy spending / intermediate inputs	0.03	0.05	0	0.02	0.07	50651
Energy spending / total costs	0.02	0.03	0	0.01	0.05	50692
Intermediate input costs / total costs	0.76	0.12	0.61	0.78	0.9	50701
Labour costs / total costs	0.24	0.12	0.1	0.22	0.39	50701
Age of Capital	4.62	3.77	1	3	10	50494
Electricity quantity share (share of joules)	0.71	0.25	0.35	0.75	1	48104

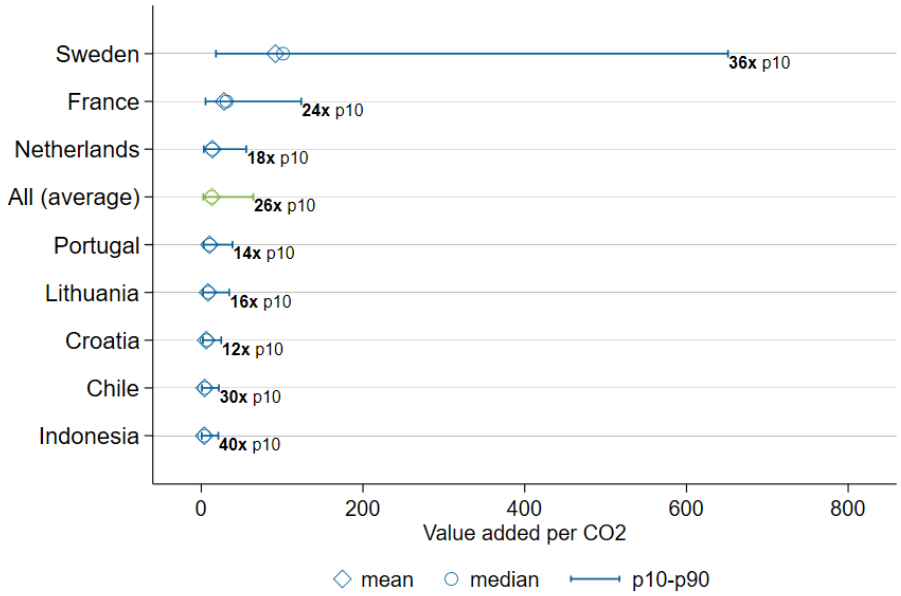
Note: The table reports summary statistics of firm-level data for manufacturing firms with at least 20 employees in Sweden. Note that intermediate inputs includes energy spending.
Source: authors' calculations

Table A B.9. Country-industry-year level summary statistics

	mean	sd	min	p10	p25	p50	p75	p90	max	count
Aggregate statistics computed for all cells										
Firm Count	109	254	1	2	6	27	96	267	3278	8106
[sum] Employment	16184	37902	20	237	1027	4250	13610	39174	514142	7593
[sum] Value Added (millions)	733	1500	-735	4	32	171	673	2085	18003	7571
Firm Count (10 plus firms)	158	295	10	14	25	63	149	393	3278	5547
Dispersion statistics computed on cells with at least 10 firms										
[p10] Log Value added / energy	11.61	1.13	6.80	10.16	10.77	11.60	12.41	13.08	15.47	5166
[p50] Log Value added / energy	13.07	0.98	10.24	11.80	12.36	13.05	13.70	14.33	16.38	5166
[p90] Log Value added / energy	14.49	1.00	11.24	13.24	13.87	14.47	15.08	15.73	18.04	5166
[p10] Log Value added / CO_2	7.49	1.54	2.67	5.56	6.35	7.36	8.55	9.52	12.87	5165
[p50] Log Value added / CO_2	9.04	1.46	5.72	7.32	7.92	8.84	9.97	11.05	14.84	5165
[p90] Log Value added / CO_2	10.58	1.55	6.51	8.85	9.46	10.26	11.51	12.78	16.31	5165
[p10] Log Value added / labour	9.37	1.33	4.85	7.41	8.48	9.47	10.55	10.91	12.61	5307
[p50] Log Value added / labour	10.23	1.02	6.22	8.90	9.55	10.26	11.10	11.40	13.59	5307
[p90] Log Value added / labour	11.08	0.88	8.18	9.96	10.49	11.13	11.69	12.13	14.98	5307
Log(VA/energy) p90-p10	2.88	0.88	0.28	1.83	2.25	2.78	3.44	4.03	7.17	5166
log(VA/CO_2) p90-p10	3.09	0.91	0.27	1.96	2.43	3.05	3.66	4.30	7.13	5165
log(VA/L) p90-p10	1.71	0.88	0.29	0.84	1.06	1.43	2.18	3.07	5.21	5307

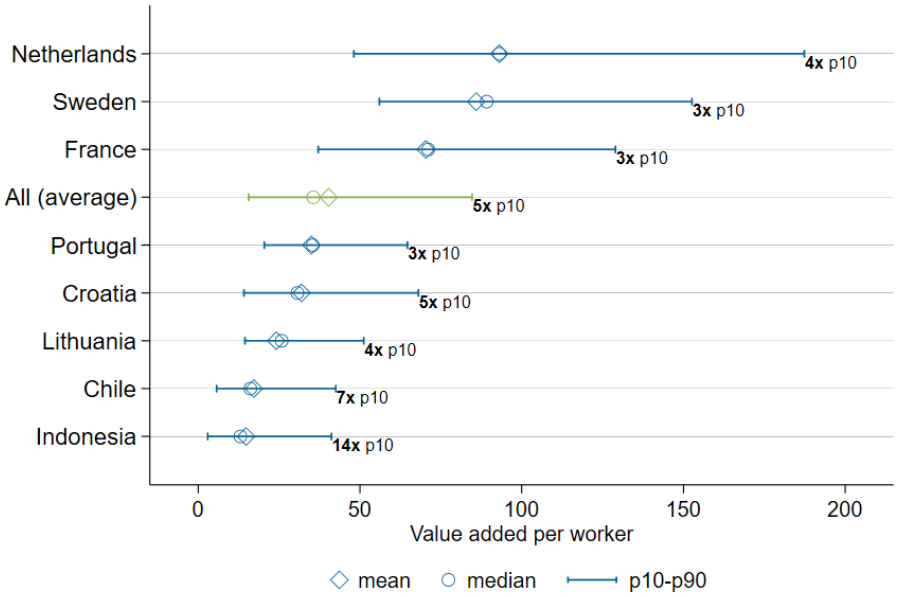
Note: The table presents summary statistics on the industry-year level data for 3-digit ISIC industries in the manufacturing sector. Only firms with at least 20 employees are included in the sample. The statistics are unweighted across industries (but weighted by sample weights to the industry-level in France and Lithuania), and cover Croatia, France, Indonesia, Lithuania, and Sweden. The variables beginning [p10], [p50], and [p90] represent within-industry percentiles of those variables while the variables p90-p10 measure within-industry dispersion. These within-industry percentiles and dispersion measures are only computed for cells with at least 10 firms.
Source: authors' calculations.

Figure A B.1. Percentiles and dispersion in carbon productivity across countries



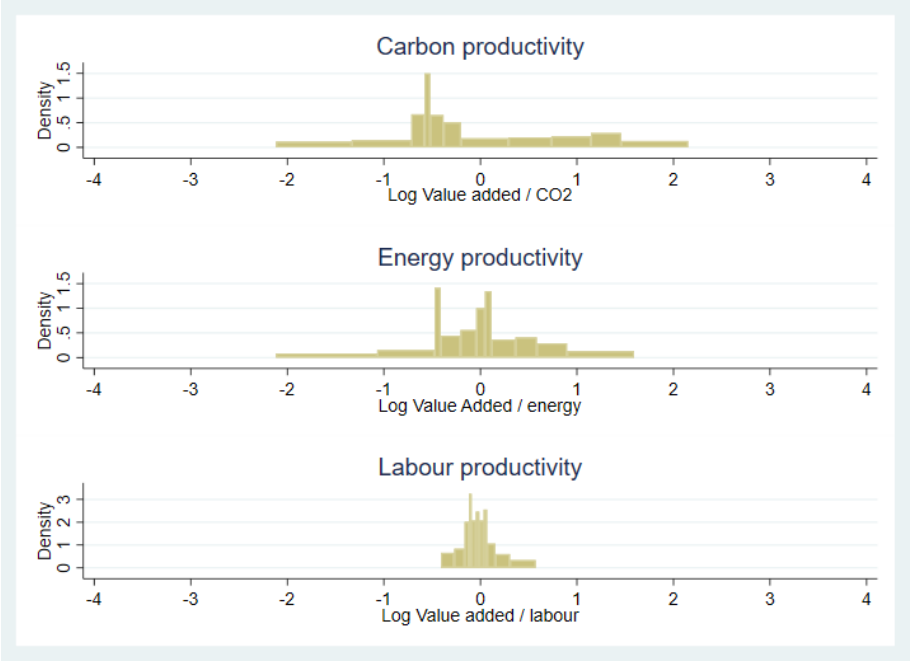
Note: The chart plots the 10th, 50th, and 90th percentiles of carbon productivity for each country, averaged across industries, in the latest year of available data for each country up to 2019. The unweighted average of the log values are taken across industries to minimise the influence of outliers, then the exponents are taken. The labels show the ratio between the 90th percentile and 10th percentile firm. The units are thousands of dollars per tonne of CO_2. The sample is restricted to manufacturing industries with at least 10 firms and firms with at least 20 employees.
Source: authors' calculations.

Figure A B.2. Dispersion and percentiles in labour productivity across countries



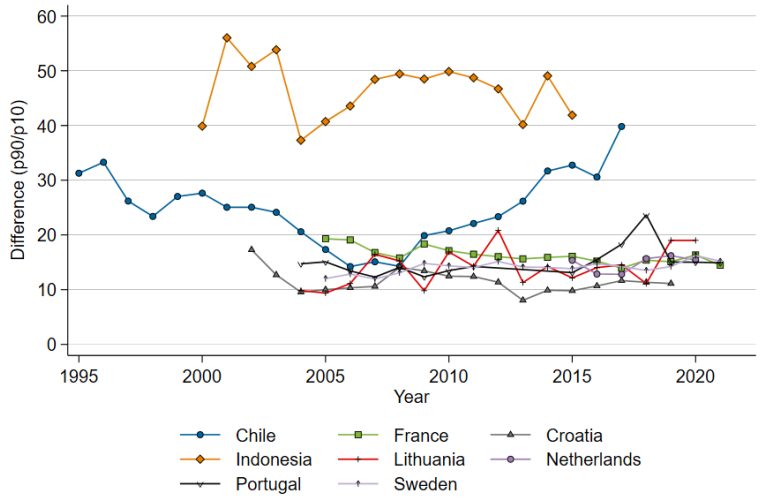
Note: The chart plots the 10th, 50th, and 90th percentiles of labour productivity for each country, averaged across industries, in the latest year of available data for each country up to 2019. The unweighted average of the log values are taken across industries to minimise the influence of outliers, then the exponents are taken. The labels show the ratio between the 90th percentile and 10th percentile firm. The units are thousands of dollars per worker. The sample is restricted to manufacturing industries with at least 10 firms and firms with at least 20 employees.
Source: authors' calculations.

Figure A B.3. Distributions of productivity in "plastic doors, windows and their frames and thresholds for doors"



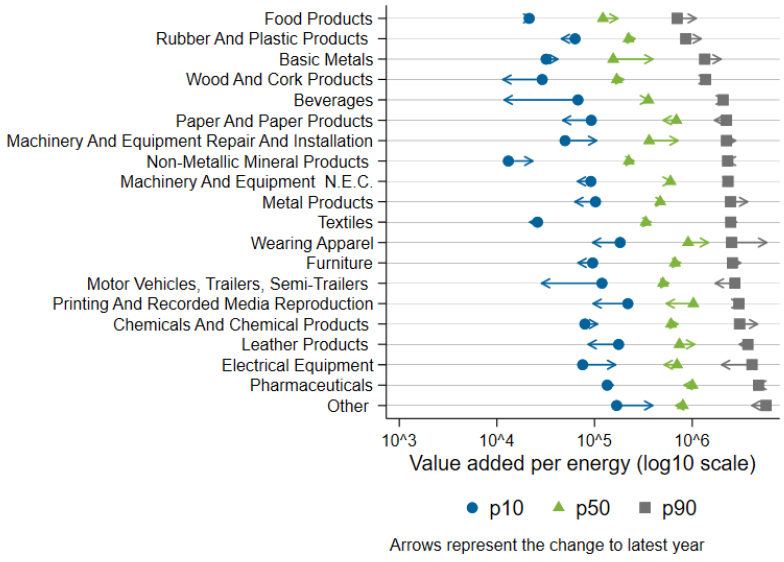
Note: The chart plots the distributions of carbon, energy, and labour productivity among firms producing only plastic doors, windows and their frames, and thresholds for doors. Each bar represents at least three firms and the distributions are centred at 0. Data is for France.
Source: Authors' calculations

Figure A B.4. Ratio between energy productivity at the 90th and 10th percentile within industries



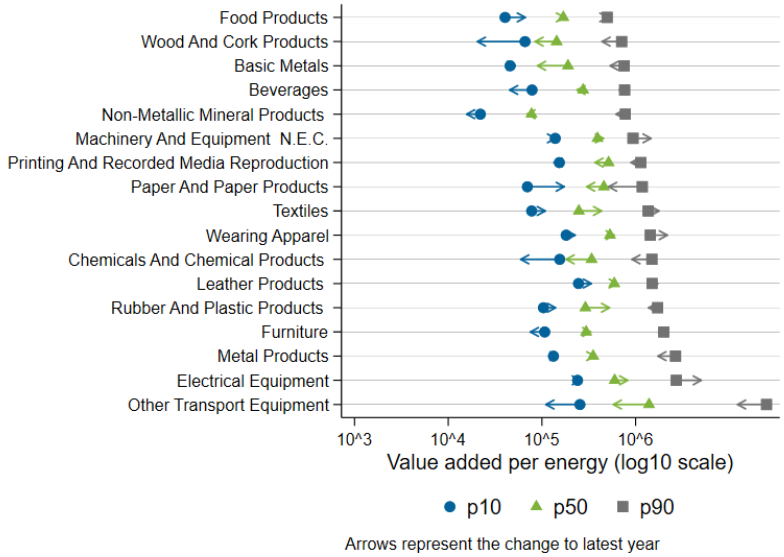
Note: The charts plot the ratio between energy productivity at the 90th and 10th percentile within industries. The average of log dispersion is taken, then the exponent of this is computed so that dispersion can be interpreted as the ratio difference between the 90th and 10th percentile firm. Industries are defined at the 3-digit level and the unweighted average across industries is taken. The sample is restricted to manufacturing industries with at least 10 firms and firms with at least 20 employees.
Source: authors' calculations.

Figure A B.5. Energy productivity at the 10th, 50th, and 90th percentiles within 2 digit industries and their change over time on a log scale in Chile



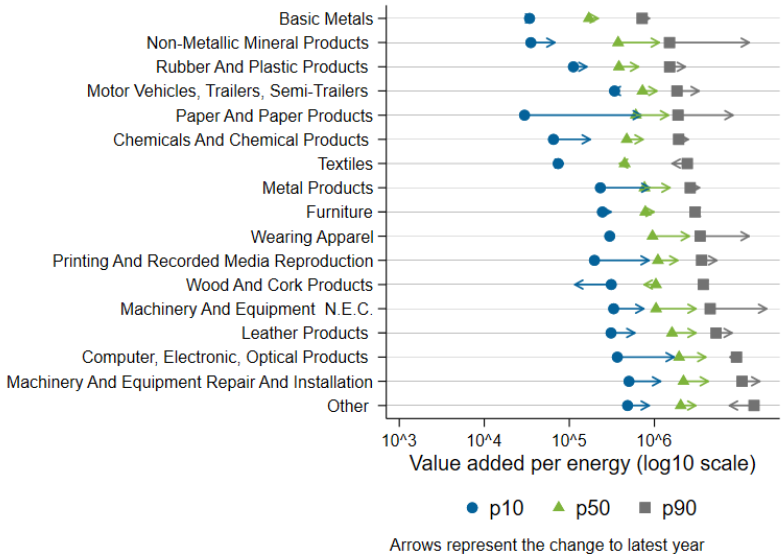
Note: The chart shows the 10th, 50th, and 90th percentiles of energy productivity within industries and their changes from 1995 to 2017 in Chile. The values are plotted on a base 10 log scale. The sample is restricted to manufacturing industries with at least 10 firms in every country-year included in the plot, and firms with at least 20 employees.
Source: authors' calculations

Figure A B.6. Energy productivity at the 10th, 50th, and 90th percentiles within 2 digit industries and their change over time on a log scale in Croatia



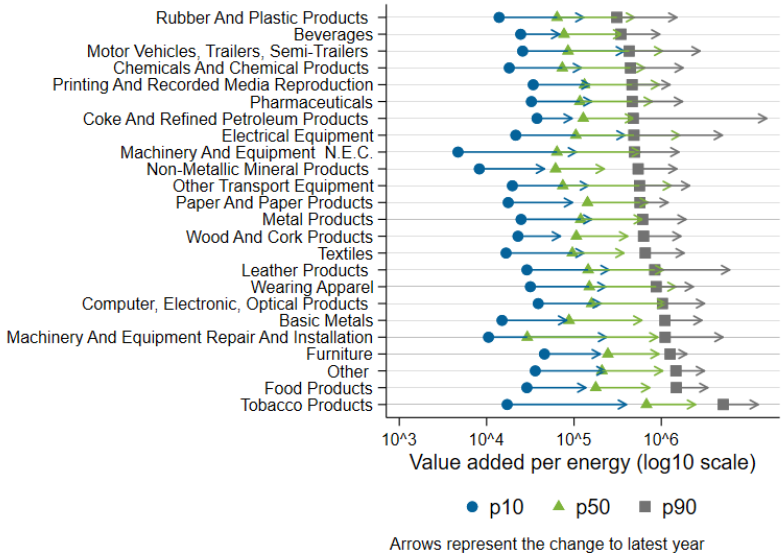
Note: The chart shows the 10th, 50th, and 90th percentiles of energy productivity within industries and their changes from 2002 to 2019 in Croatia. The values are plotted on a base 10 log scale. The sample is restricted to manufacturing industries with at least 10 firms in every country-year included in the plot, and firms with at least 20 employees.
Source: authors' calculations

Figure A B.7. Energy productivity at the 10th, 50th, and 90th percentiles within 2 digit industries and their change over time on a log scale in France



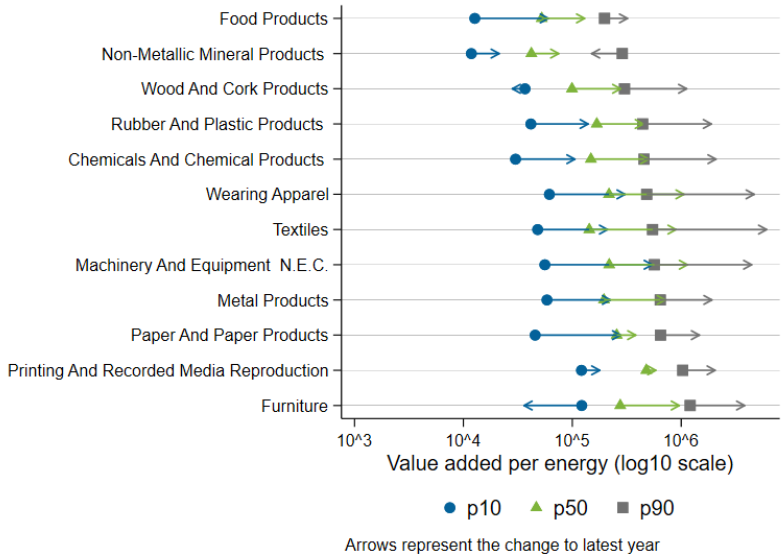
Note: The chart shows the 10th, 50th, and 90th percentiles of energy productivity within industries and their changes from 2005 to 2019 in France. The values are plotted on a base 10 log scale. The sample is restricted to manufacturing industries with at least 10 firms in every country-year included in the plot, and firms with at least 20 employees.
Source: authors' calculations

Figure A B.8. Energy productivity at the 10th, 50th, and 90th percentiles within 2 digit industries and their change over time on a log scale for Indonesia



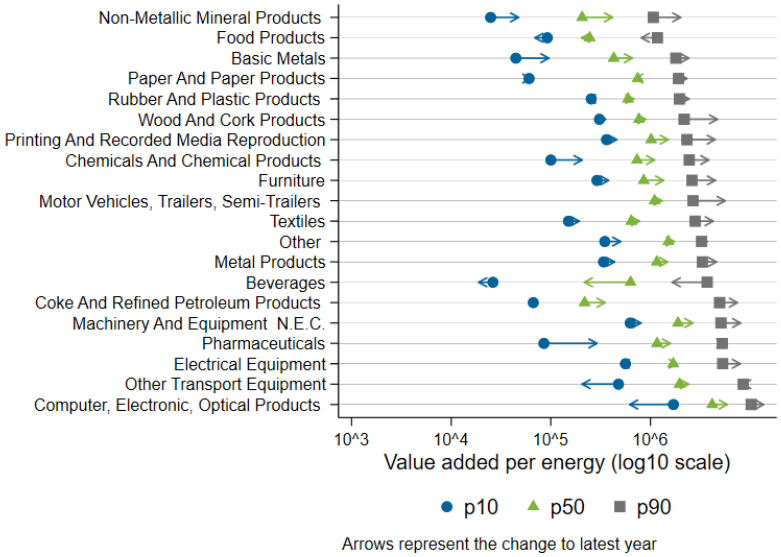
Note: The chart shows the 10th, 50th, and 90th percentiles of energy productivity within industries and their changes from 2000 to 2015 in Indonesia. The values are plotted on a base 10 log scale. The sample is restricted to manufacturing industries with at least 10 firms in every country-year included in the plot, and firms with at least 20 employees.
Source: authors' calculations

Figure A B.9. Energy productivity at the 10th, 50th, and 90th percentiles within 2 digit industries and their change over time on a log scale for Lithuania



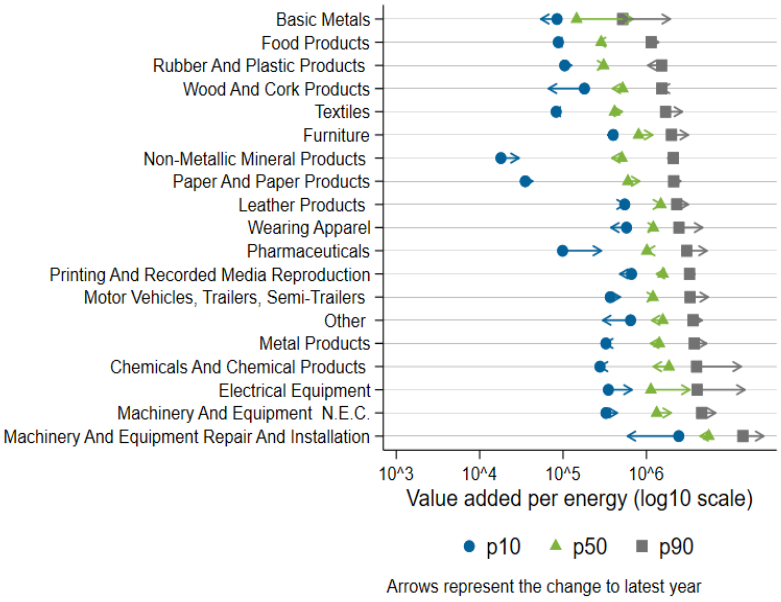
Note: The chart shows the 10th, 50th, and 90th percentiles of energy productivity within industries and their changes from 2004 to 2019 in Lithuania. The values are plotted on a base 10 log scale. The sample is restricted to manufacturing industries with at least 10 firms in every country-year included in the plot, and firms with at least 20 employees.
Source: authors' calculations

Figure A B.10. Energy productivity at the 10th, 50th, and 90th percentiles within 2 digit industries and their change over time on a log scale for Netherlands



Note: The chart shows the 10th, 50th, and 90th percentiles of energy productivity within industries and their changes from 2015 to 2019 in Netherlands. The values are plotted on a base 10 log scale. The sample is restricted to manufacturing industries with at least 10 firms in every country-year included in the plot, and firms with at least 20 employees.
Source: authors' calculations

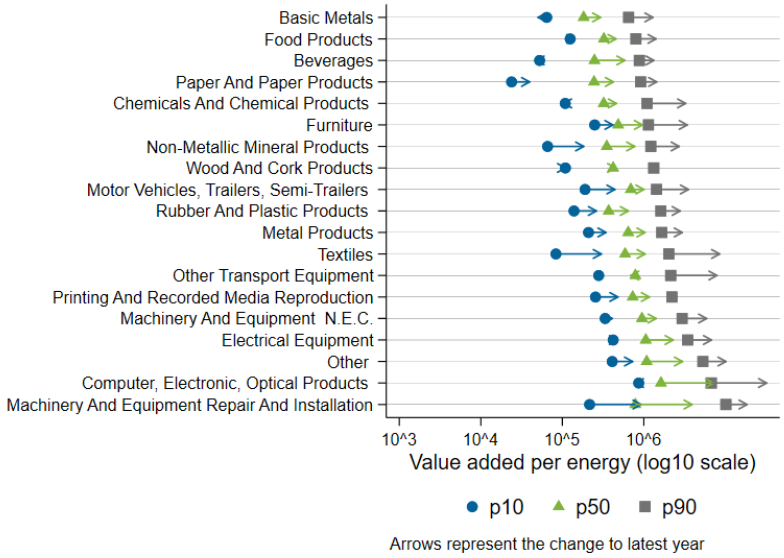
Figure A B.11. Energy productivity at the 10th, 50th, and 90th percentiles within 2 digit industries and their change over time on a log scale for Portugal



Note: The chart shows the 10th, 50th, and 90th percentiles of energy productivity within industries and their changes from 2004 to 2019 in Portugal. The values are plotted on a base 10 log scale. The sample is restricted to manufacturing industries with at least 10 firms in every country-year included in the plot, and firms with at least 20 employees.

Source: authors' calculations

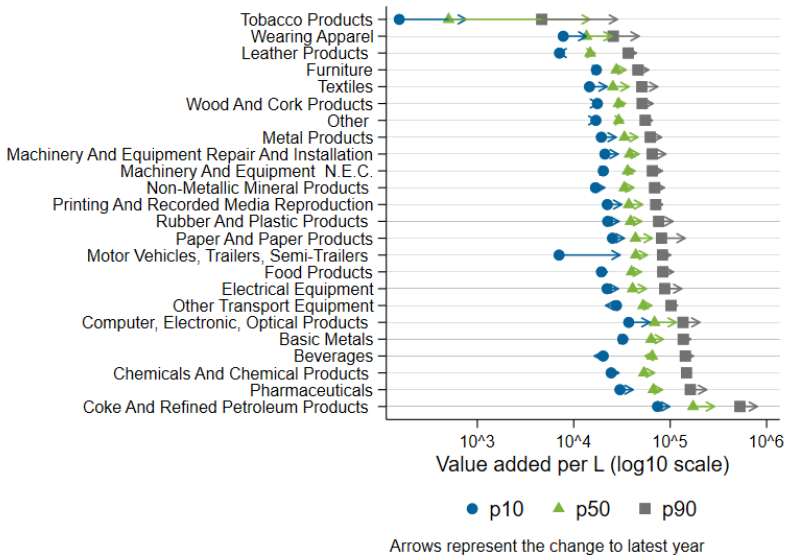
Figure A B.12. Energy productivity at the 10th, 50th, and 90th percentiles within 2 digit industries and their change over time on a log scale for Sweden



Note: The chart shows the 10th, 50th, and 90th percentiles of energy productivity within industries and their changes from 2004 to 2020 in Sweden. The values are plotted on a base 10 log scale. The sample is restricted to manufacturing industries with at least 10 firms in every country-year included in the plot, and firms with at least 20 employees.
Source: authors' calculations

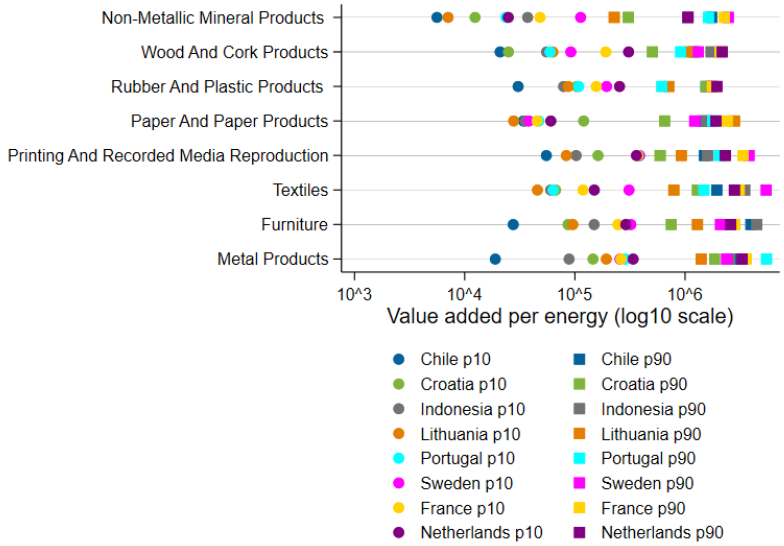
Figure A B.13. Variation across industries in labour productivity dispersion

Labour productivity at the 10th, 50th, and 90th percentiles within 2 digit industries and their change over time on a log scale



Note: The chart shows the 10th, 50th, and 90th percentiles of labour productivity within industries and their changes from the earliest year to the latest year, up to 2019. The values are plotted on a base 10 log scale. The unweighted average across countries is taken. The sample is restricted to manufacturing industries with at least 10 firms in every country-year, and firms with at least 20 employees.
Source: authors' calculations

Figure A B.14. Variation across countries and industries in frontier and lagging firms' energy productivity



Note: The chart shows the 10th and 90th percentiles of energy productivity within industries for each country in 2015. The values are plotted on a base 10 log scale. Industries are restricted to those that include all countries with data in that year. The sample is restricted to manufacturing industries with at least 10 firms in every country-year included in the plot, and firms with at least 20 employees.
Source: authors' calculations.

Table A B.10. Ranking of industries by energy productivity dispersion

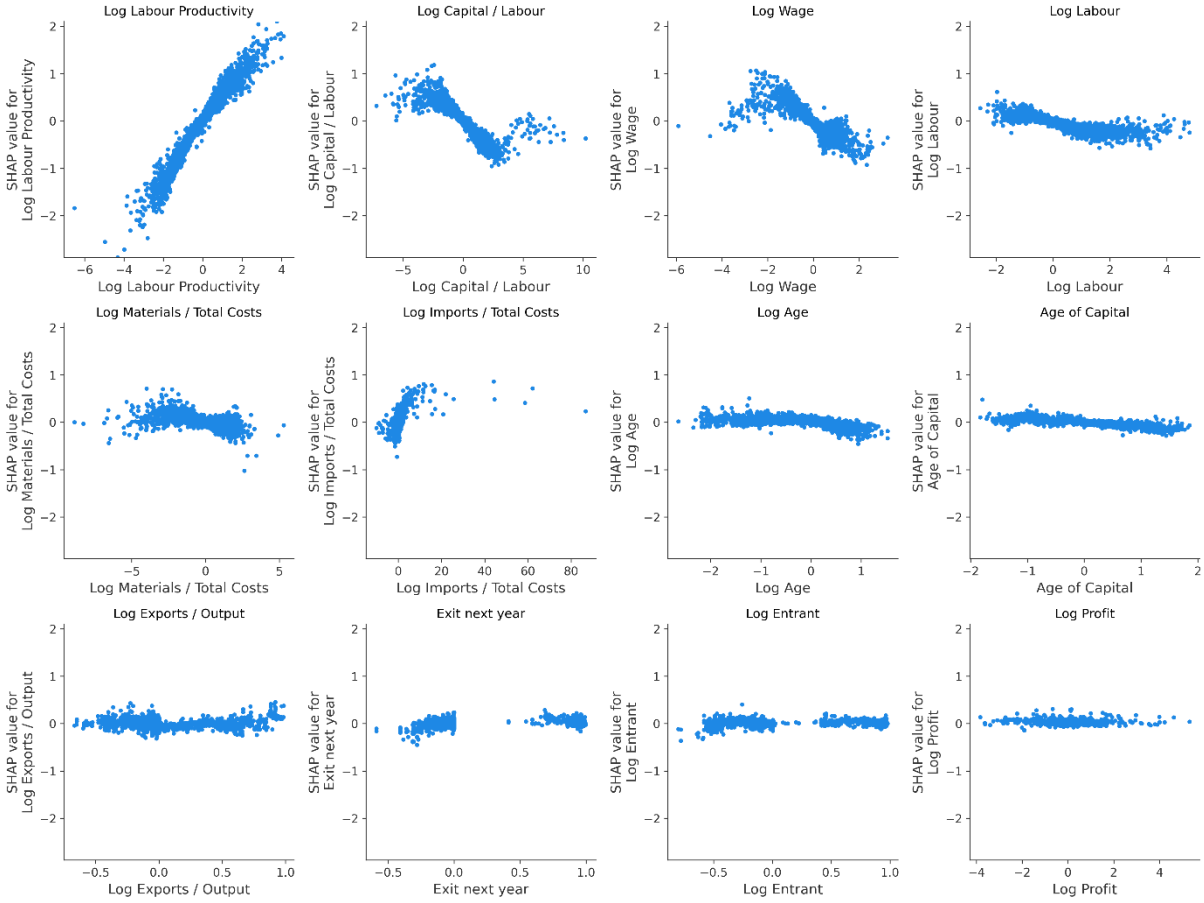
Industry (3 digit)	Log dispersion in VA/energy	Log dispersion in VA/worker	Difference in log dispersion of VA/energy and VA/Worker	Average of energy cost share	Average of CO2 emissions
Manufacture of consumer electronics	4.83	3.36	1.47	0.05	2181
Manufacture of tobacco products	4.84	3.98	0.86	0.11	4951
Manufacture of glass and glass products	4.02	1.77	2.25	0.10	11601
Manufacture of basic iron and steel	3.80	2.00	1.80	0.11	29605
Manufacture of basic chemicals and plastics	3.75	2.35	1.40	0.12	29138
Batteries and accumulators manufacturing	3.91	2.83	1.07	0.21	6673
Manufacture of games and toys	3.76	2.09	1.67	0.15	321
Manufacture of man-made fibres	3.70	3.27	0.42	0.16	14656
Non-metallic mineral products manufacturing n.e.c.	3.61	1.90	1.71	0.15	5341
Manufacture of sports goods	3.65	1.95	1.70	0.06	158
Vegetable and animal oils and fats manufacturing	3.49	2.92	0.57	0.13	4686
Industrial machinery and equipment installation	2.86	0.92	1.94	0.04	171

Repair of metal products and machinery	3.18	1.43	1.75	0.08	610
Manufacture of other textiles	3.29	1.48	1.81	0.08	745
Leather tanning and luggage manufacturing	3.40	1.76	1.63	0.07	519
Manufacture of paper and paper products	3.29	1.55	1.74	0.09	19971
Jewellery and related articles manufacturing	3.33	1.97	1.36	0.07	206
Manufacture of musical instruments	3.71	2.93	0.78	0.08	2889
Transport equipment manufacturing n.e.c.	2.83	1.78	1.05	0.05	1046
Textiles spinning, weaving, finishing	3.39	1.74	1.65	0.16	3186
Building of ships and boats	2.94	1.87	1.07	0.07	1251
Fruit and vegetables processing and preserving	3.03	1.68	1.35	0.10	3426
Manufacture of pharmaceuticals	3.04	1.94	1.11	0.07	3020
Basic precious and non-ferrous metals manufacturing	3.00	2.19	0.81	0.14	18893
Manufacture of wiring and wiring devices	2.96	1.62	1.34	0.06	1732
Electronic components and boards manufacturing	2.80	1.63	1.16	0.09	2152
Manufacture of lighting equipment	2.85	1.68	1.17	0.08	1790
Manufacture of communication equipment	2.80	1.41	1.39	0.07	1322
Manufacture of beverages	2.93	2.40	0.53	0.11	3146
Fish and seafood processing and preserving	2.96	1.76	1.20	0.09	2180
Manufacture of domestic appliances	2.97	1.57	1.39	0.06	1645
Other manufacturing n.e.c.	2.75	1.63	1.12	0.07	377
Manufacture of wood products	2.89	1.57	1.31	0.10	1241
Manufacture of measuring and control equipment	2.64	1.41	1.23	0.06	364
Manufacture of other chemical products	2.83	1.80	1.04	0.07	1326
Electrical equipment manufacturing	2.59	1.39	1.20	0.05	725
Manufacture of dairy products	2.67	2.01	0.66	0.11	7907
Knitted and crocheted apparel manufacturing	2.84	1.57	1.26	0.09	1313
Manufacture of electric motors and transformers	2.74	1.55	1.20	0.06	1274
Computers and peripheral equipment manufacturing	2.57	1.42	1.15	0.06	708
Sawmilling and planing of wood	2.68	1.56	1.12	0.14	1078
Manufacture of metal products	2.73	1.44	1.29	0.09	773
Manufacture of grain mill products	2.59	1.99	0.60	0.12	6838
Manufacture of structural metal products	2.63	1.45	1.18	0.08	609
Railway locomotives and rolling stock manufacturing	2.46	1.08	1.38	0.08	2377
Motor vehicle parts and accessories manufacturing	2.63	1.49	1.15	0.08	2379
Medical and dental instruments and supplies manufacturing	2.51	1.38	1.13	0.06	763
Manufacture of prepared animal feeds	2.82	2.23	0.59	0.11	4941
Manufacture of special-purpose machinery	2.61	1.52	1.09	0.08	715
Manufacture of general-purpose machinery	2.63	1.48	1.15	0.07	801
Manufacture of optical instruments	2.37	1.36	1.02	0.03	248
Casting of metals	2.40	1.44	0.96	0.18	3980
Manufacture of other food products	2.52	1.54	0.98	0.13	1659
Air and spacecraft manufacturing	2.42	1.27	1.14	0.13	1610
Manufacture of furniture	2.52	1.30	1.22	0.07	440
Manufacture of rubber products	2.47	1.69	0.78	0.11	1822
Manufacture of plastics products	2.54	1.51	1.03	0.09	1062
Manufacture of motor vehicle bodies	2.31	1.39	0.93	0.05	324

Manufacture of footwear	2.55	1.46	1.08	0.08	642
Wearing apparel manufacturing, except fur	2.52	1.47	1.05	0.07	294
Manufacture of motor vehicles	2.08	1.29	0.79	0.05	9293
Processing and preserving of meat	2.18	1.57	0.61	0.09	2288
Printing and related services	2.22	1.31	0.92	0.07	463
Manufacture of weapons and ammunition	1.94	1.38	0.56	0.08	2574
Manufacture of medical equipment	1.54	0.96	0.58	0.09	414

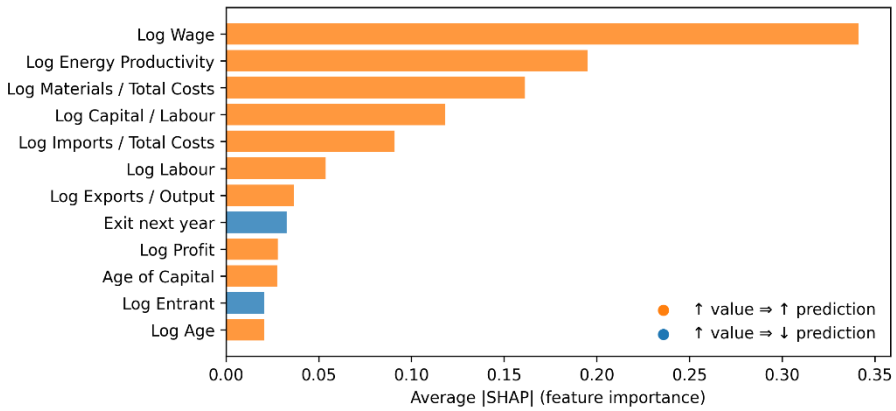
Note: The table shows 3 digit industries, ranked by their energy productivity dispersion – largest to smallest. Dispersion in energy and labour productivity are defined as the log difference between the 90th percentile and 10th percentile firms in their 3 digit industry. The fourth column shows the difference between energy and labour productivity dispersions. The remaining two columns show the average share of energy in total costs and average CO2 emissions per firm. The measures are averaged across countries and years.
Source: authors' calculations

Figure A B.15. Relationship between SHAP and predictor values for firms' energy productivity



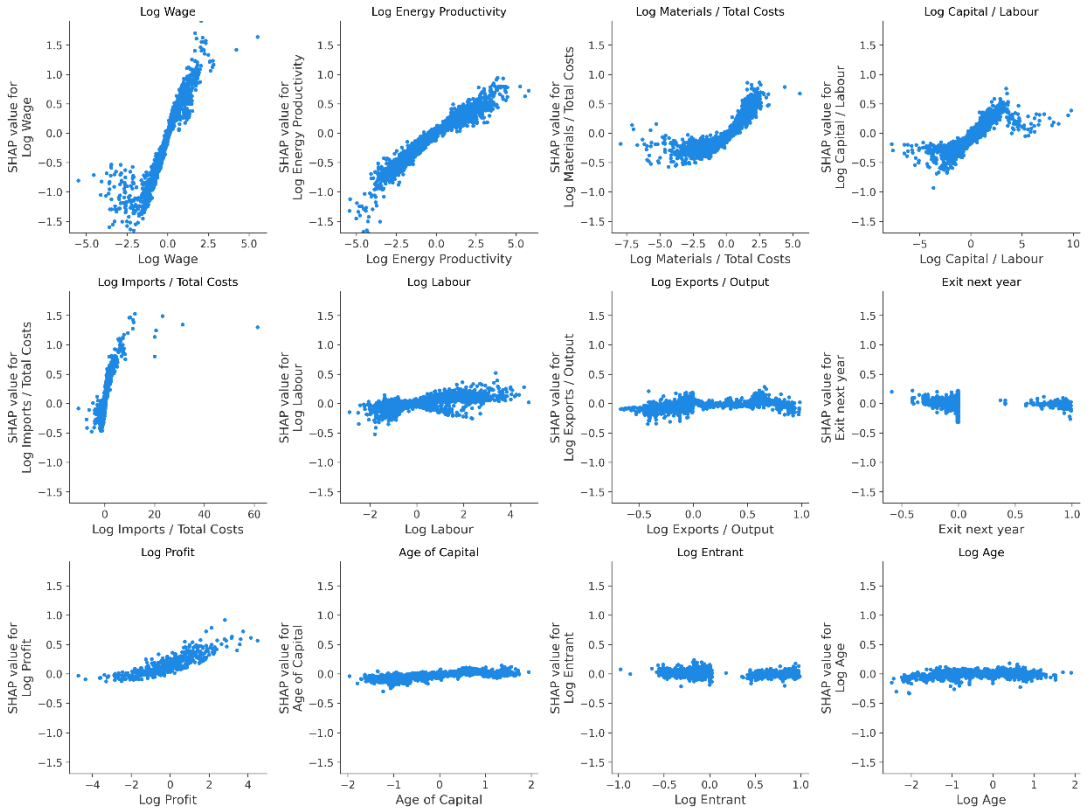
Note: The charts plot SHAP values against the value of the predictor for each observation. Each chart represents a different predictor variable. The response variable is energy productivity at the firm level. Both the predictor and response variable are measured subtracting the country-industry-year average value, computed separately within the training and test datasets. The prediction model is estimated using XGBoost and the SHAP analysis is carried out on a random sample of 5 000 firm-year observations. The sample is restricted to firms which have at least 20 employees in the manufacturing industry.
Source: authors' calculations.

Figure A B.16. Average SHAP values of predictors of labour productivity in XGBoost model and direction of marginal effect on prediction



Note: The chart plots the mean of the absolute SHAP value for each predictor. The response variable is labour productivity at the firm level. Both the predictor and response variable are measured subtracting the country-industry-year average value, computed separately within the training and test datasets. The prediction model is estimated using XGBoost and the SHAP analysis is carried out on a random sample of 5 000 firm-year observations. The sample is restricted to firms which have at least 20 employees in the manufacturing industry.
Source: authors' calculations.

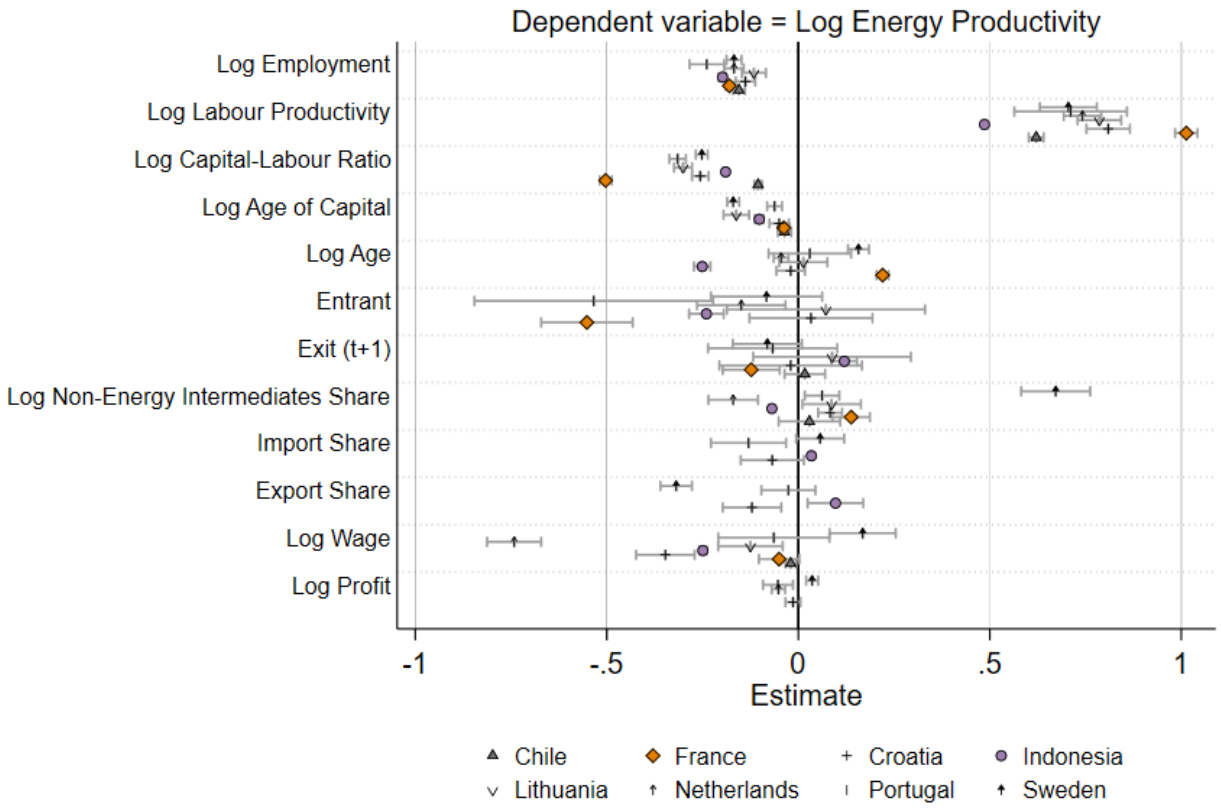
Figure A B.17. Relationship between SHAP and predictor values for firms' labour productivity



Note: The charts plot SHAP values against the value of the predictor for each observation. Each chart represents a different predictor variable. The response variable is labour productivity at the firm level. Both the predictor and response variable are measured subtracting the country-industry-year average value, computed separately within the training and test datasets. The prediction model is estimated using XGBoost and the SHAP analysis is carried out on a random sample of 5 000 firm-year observations. The sample is restricted to firms which have at least 20 employees in the manufacturing industry.
Source: authors' calculations.

Figure A B.18. Factors correlated with firms' energy productivity – multivariate regressions

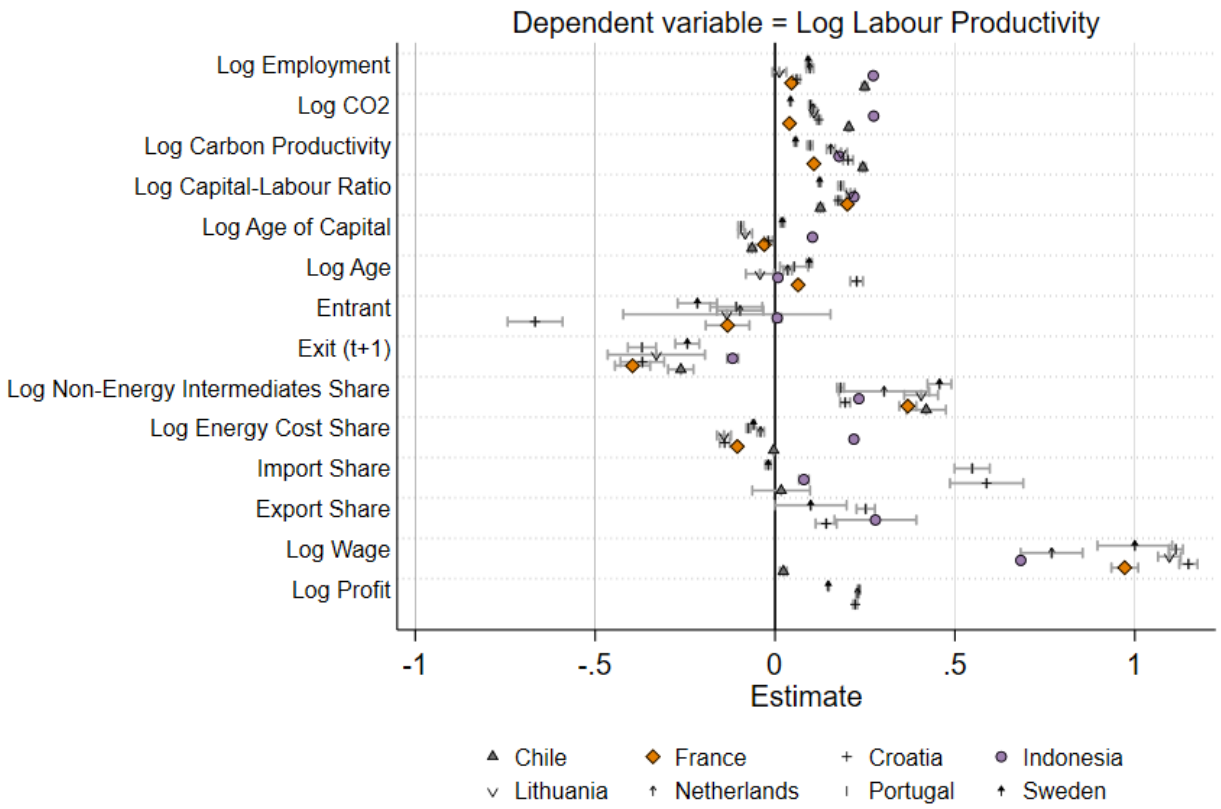
Estimates from multivariate OLS regressions of log value added per energy on various explanatory variables



Note: The chart shows estimates from multivariate OLS regressions of log value added per energy on all the explanatory variables listed simultaneously. Note that log CO2 is excluded from the multivariate regression to avoid saturating the model alongside labour productivity and employment (with energy productivity, the dependent variable, strongly related to carbon productivity). They are estimated separately for each country, with country-specific coefficients reported in the plot. The y axis corresponds to the explanatory variable in each regression. The bars indicate 95% confidence intervals with standard errors that are robust to heteroskedasticity. The regressions all include industry-year fixed effects. The sample is restricted to firms with at least 20 employees in manufacturing industries.
Source: authors' calculations

Figure A B.19. Factors correlated with firms' labour productivity

Estimates from bivariate OLS regressions of log value added per worker on various explanatory variables



Note: The chart shows estimates from bivariate OLS regressions of log value added per worker on various explanatory variables. They are estimated separately for each country, with country-specific coefficients reported in the plot. The y axis corresponds to the explanatory variable in each regression. The bars indicate 95% confidence intervals with standard errors that are robust to heteroskedasticity. The regressions all include industry-year fixed effects. The sample is restricted to firms with at least 20 employees in manufacturing industries.
Source: authors' calculations

Relationship between changes in energy and labour productivity

The relationship between changes in labour productivity and changes in energy productivity within firms over time is also explored. To do so, the following regression specification is estimated:

$$\Delta_l \log\left(\frac{VA_{ict}}{ENQ_{ict}}\right) = \beta \Delta_l \log\left(\frac{VA_{ict}}{L_{ict}}\right) + \alpha_{k(i)t} + \Delta_l X_{ict} + X_{ic,t=0} + \gamma_1 FirstYear_{ict} + \gamma_2 FinalYear_{ict} + \gamma_3 YearSpan_{ict} + \varepsilon_{ict}$$

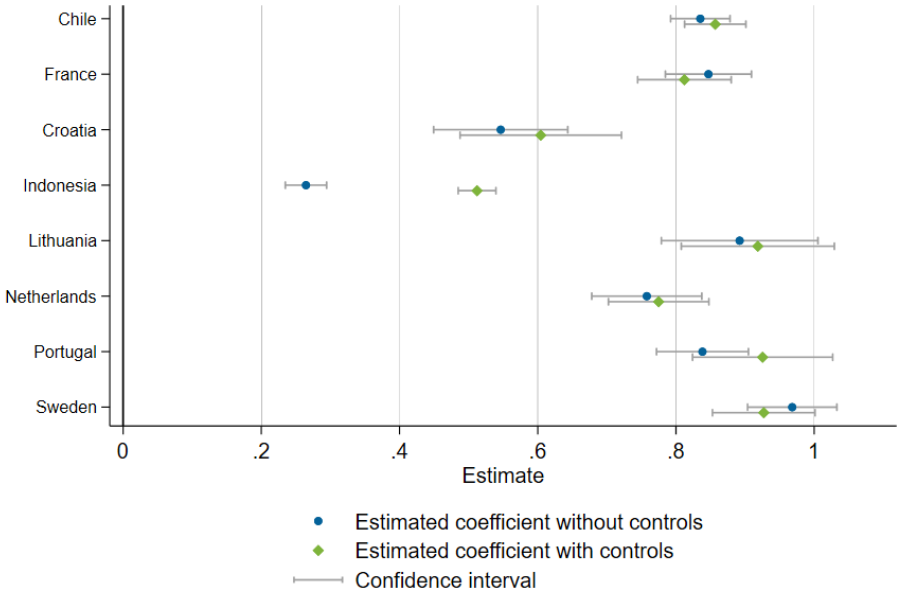
Equation 16

The regression is designed to be robust to repeated cross-sectional sampled data, where firms might not be sampled in all years. It is estimated in long changes – taking the difference between the first and last observation of the firm. The variables $\left(\frac{VA_{ict}}{ENQ_{ict}}\right)$, $\left(\frac{VA_{ict}}{L_{ict}}\right)$, and $\alpha_{k(i)t}$ are defined as before (energy productivity, labour productivity, and industry-year fixed effects). X_{ict} is a set of firm controls, including (log) capital, (log) intermediate inputs, and (log) age (set to 1 in the first year and included in levels). To account for

differences in the time periods where each firm is first observed and finally observed, first year (the first year the firm is observed), final year (the final year the firm is observed), and year span (the number of years between the first and the last year of observation) dummy variables are included. The results are robust to the exclusion of controls, to recalibrating sample weights to account for the requirement of two periods, and to alternative specifications such as shorter period changes or a specification in levels with firm fixed effects.

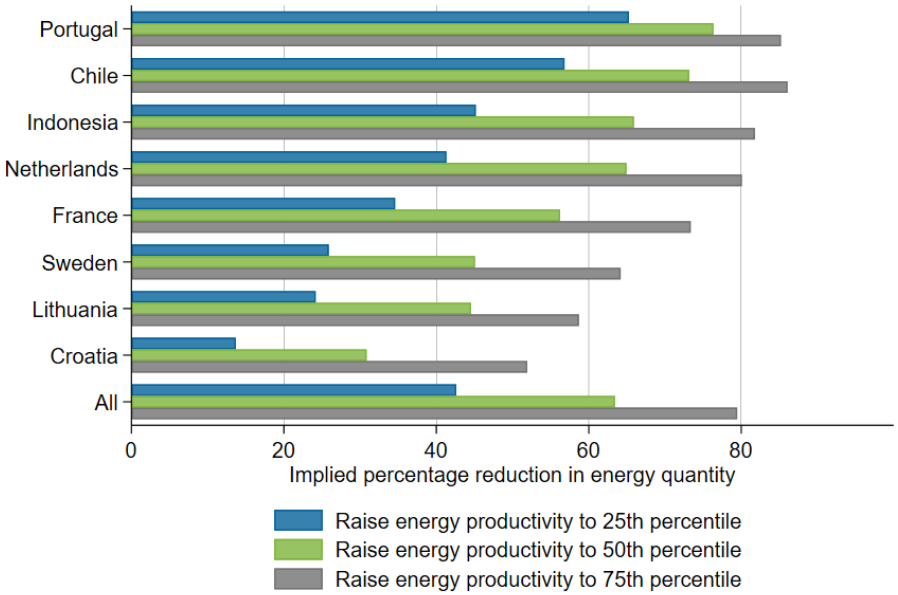
The results, with and without controls, are presented in Figure A B.18. There is a strong positive relationship between changes in labour productivity and changes in energy productivity. Put differently, as firms improve their labour productivity, they also improve their energy productivity, even controlling for changes in their overall size. The estimated coefficients are between 0.2 and 1 depending on the country, suggesting that a 10% increase in labour productivity is associated with between a 2% and 10% increase energy productivity.

Figure A B.20. Estimated coefficients from OLS regressions of change in energy productivity on change in labour productivity and controls



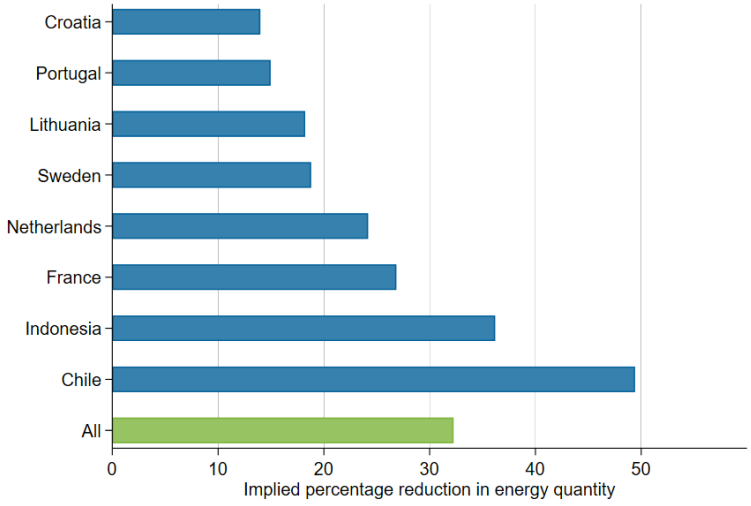
Note: The chart shows estimates from OLS regressions of the long change in log value added per energy on the long change in log labour productivity. Controls for (log) capital, (log) intermediate inputs, and (log) age are included in initial levels and, for the first two, in changes. Additional fixed effects are included for the entry and exit year of the firm, and the length of that period. The regressions also include industry-year fixed effects. They are estimated separately for each country, with country-specific coefficients reported in the plot. The y axis corresponds to the explanatory variable in each regression. The bars indicate 95% confidence intervals with standard errors that are robust to heteroskedasticity. The sample is restricted to firms with at least 20 employees in manufacturing industries.
Source: authors' calculations.

Figure A B.21. Counterfactual exercises showing implied reduction in energy to produce the same aggregate value added if all firms were at least as productive as the 25th, 50th, and 75th percentile firms in their industry



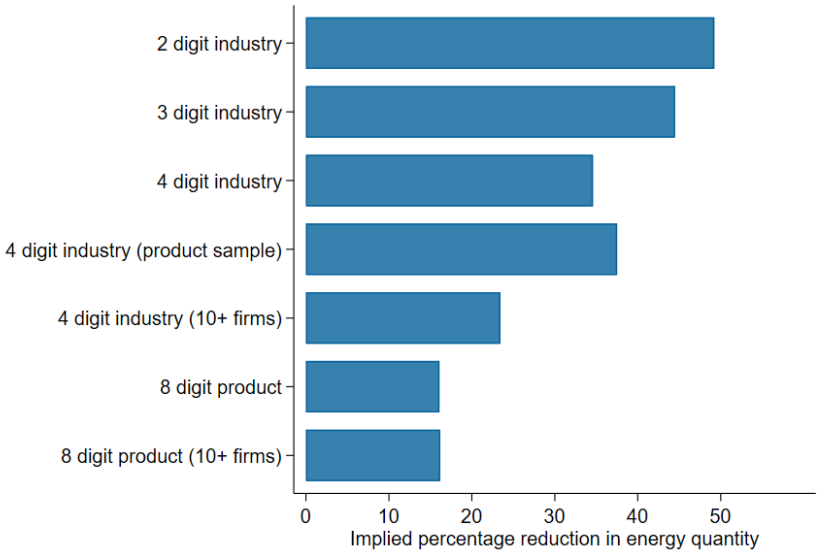
Note: The chart shows the implied percentage reduction in energy that would be achieved if all firms had at least a energy productivity equal to the 25th, 50th, or 75th percentile firm in their 4-digit industry. The data is for the latest year available in each country up to 2019 and is restricted to manufacturing sector firms with at least 20 employees.
Source: authors' calculations

Figure A B.22. Counterfactual exercises showing implied reduction in energy to produce the same aggregate value added if all firms were at least as productive as the 25th percentile firm in their industry for firms of similar employment



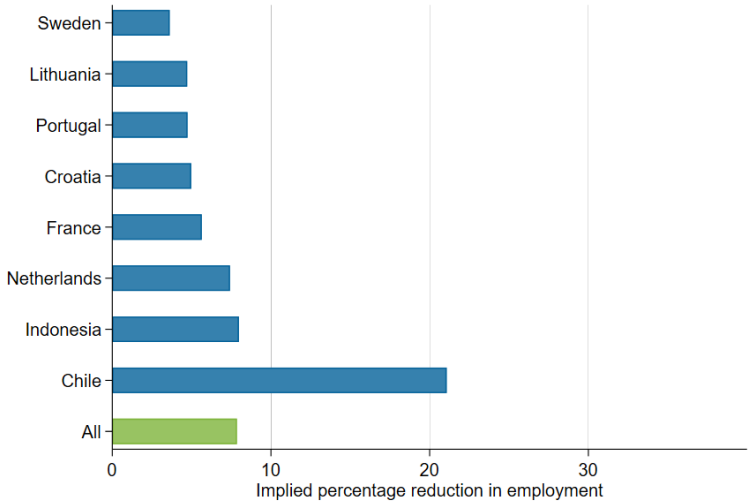
Note: The chart shows the implied percentage reduction in energy that would be achieved if all firms had at least a energy productivity equal to the 25th percentile firm in their 4-digit industry and of a similar employment size (with bands defined as employment of 20-49, 50-249, and 250+). The data is for the latest year available in each country up to 2019 and is restricted to manufacturing sector firms with at least 20 employees.
Source: authors' calculations

Figure A B.23. Counterfactual exercises showing implied reduction in energy to produce the same aggregate value added if all firms were at least as productive as the 25th percentile firms in their industry or product



Note: The chart shows the implied percentage reduction in energy that would be achieved if all firms had at least a energy productivity equal to the 25th percentile firm in their industry or product. Industries are defined at the 2, 3, or 4 digit level, while products are defined at the 8 digit level. For 4 digit industries and 8 digit products, results are presented restricting to cells with at least 10 firms because the narrowly defined product markets have many cells with few firms. The data is for France in 2019 and is restricted to manufacturing sector firms with at least 20 employees.
Source: authors' calculations

Figure A B.24. Counterfactual exercises showing implied reduction in employment to produce the same aggregate value added if all firms were at least as productive as the 25th percentile firm in their industry



Note: The chart shows the implied percentage reduction in employment that would be achieved if all firms had at least a labour productivity equal to the 25th percentile firm in their 3 digit industry. The data is for 2019 and is restricted to manufacturing sector firms with at least 20 employees.
Source: authors' calculations

Explaining dispersion in energy productivity

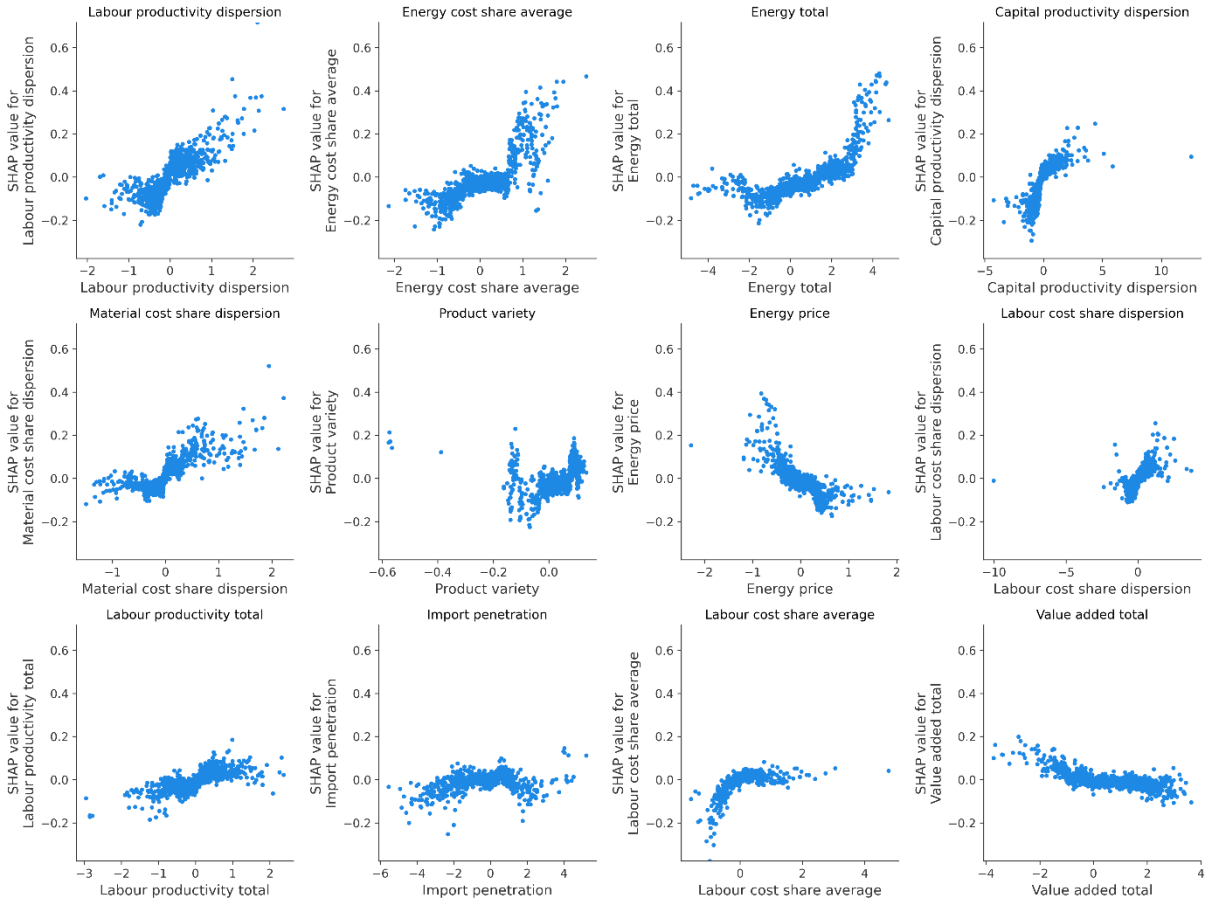
Table A B.11 shows the ML prediction model performance on the test dataset and compares with the performance of an equivalent OLS regression estimated on the test data only and on the full dataset. The ML method explains 46% of the variation in within country-year dispersion in energy productivity, compared with 34% from the OLS even on the full sample, as measured by the R-squared. The root mean square error (RMSE) – a measure of the difference between predicted and observed values – is lower in the ML model compared with the OLS. Finally, note that the ML is able to use many more observations of the OLS on the test dataset and even on the full dataset because of its robustness to missing values. As some of the variables included in the ML model and in the bivariate OLS regressions are particularly problematic for missing values, they are excluded from the multivariate regressions presented in Figure A B.27, where the sample size increases to 1 931 observations.

Table A B.11. XGBoost prediction of energy productivity dispersion: model performance

	XGBoost (test sample)	OLS (test sample)	OLS (full sample)
Observations	1 031	135	680
Root mean square error	0.53	0.61	0.54
R-squared	0.46	0.18	0.34

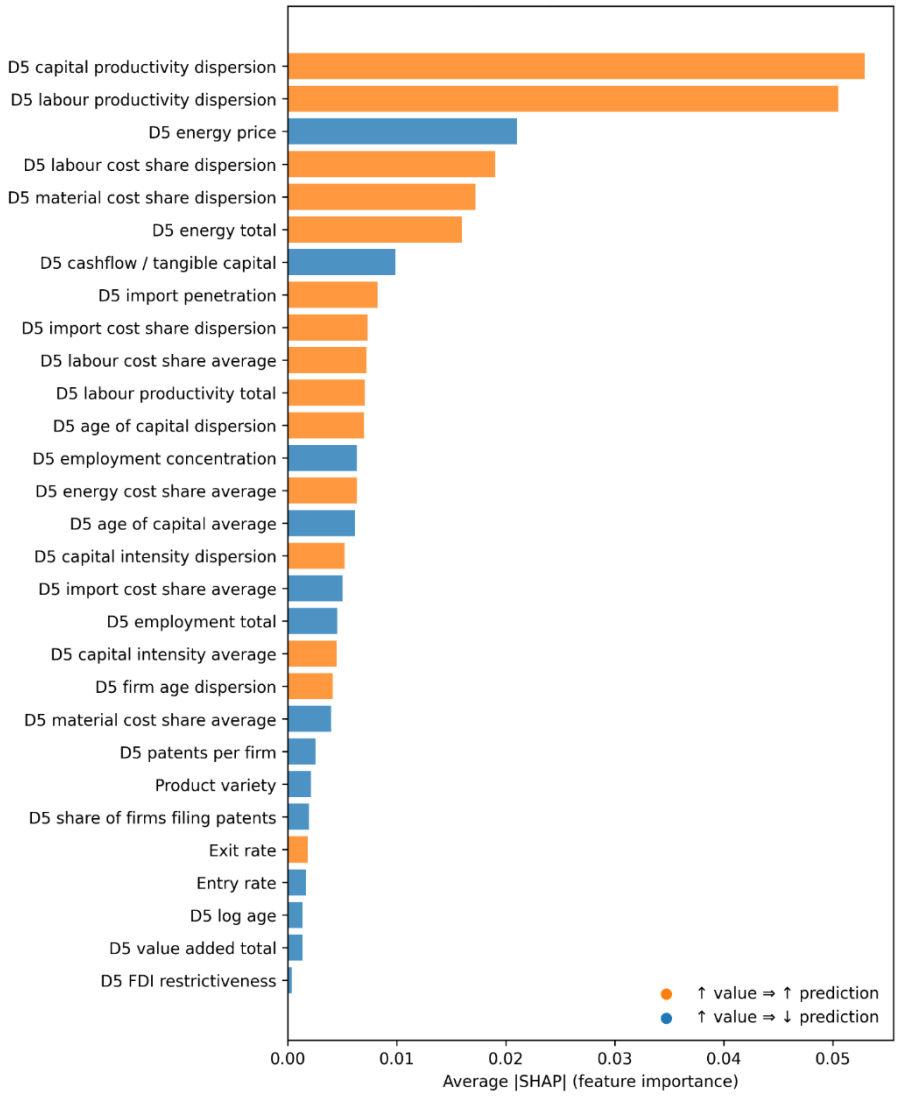
Note: The table shows model statistics for the XGBoost prediction method on the test sample (column 1), an OLS regression on the test sample (column 2), and the OLS on the full sample. The models are estimated on country-industry-year level data. The dependent variable is dispersion in energy productivity and all variables are residualised on country-year fixed effects before the estimation.
Source: authors' calculations

Figure A B.25. Relationship between SHAP and predictor values for industry dispersion in energy productivity for top 12 factors



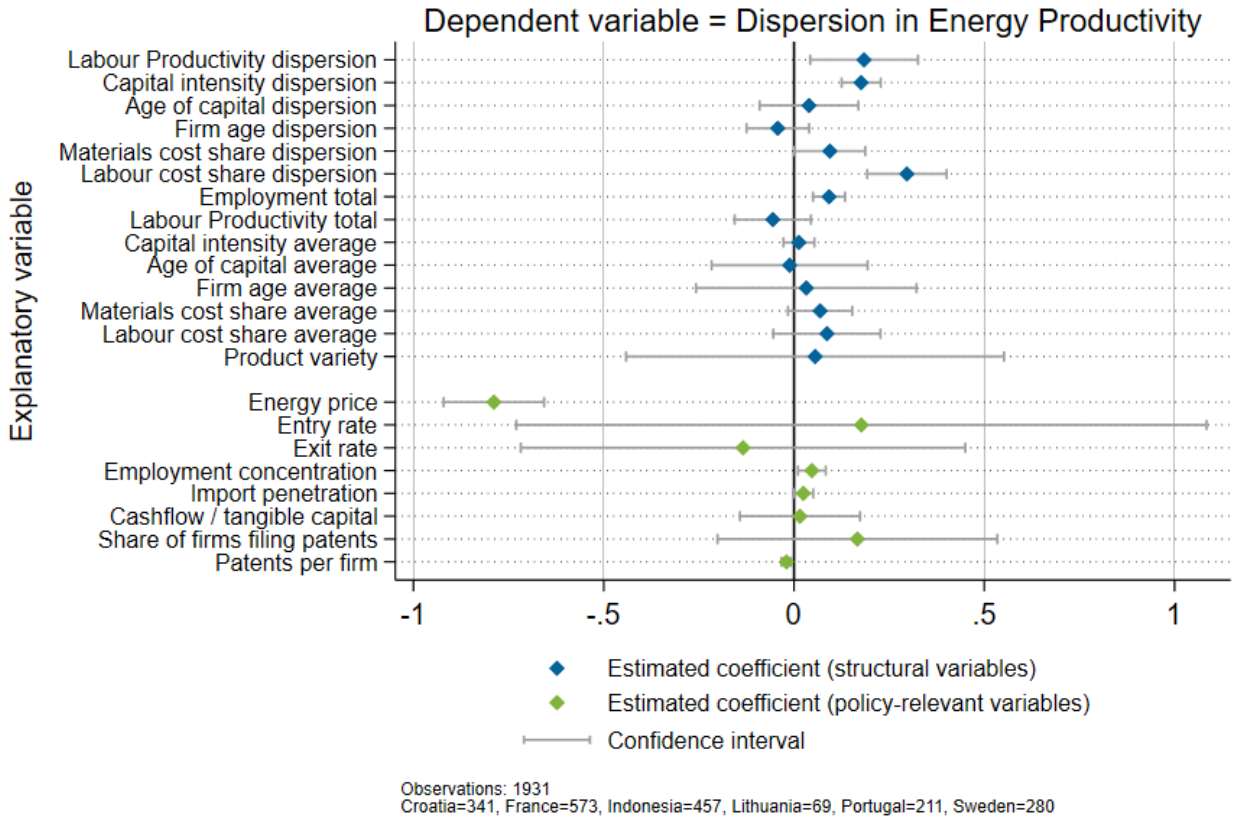
Note: The charts plot SHAP values against the value of the predictor for each observation. Each chart represents a different predictor variable. The response variable is industry dispersion in energy productivity. Both the predictor and response variable are measured subtracting the country-year average value, computed separately within the training and test datasets. The prediction model is estimated using XGBoost. The sample is restricted to manufacturing industries with at least 10 firms which have at least 20 employees.
Source: authors' calculations.

Figure A B.26. Average SHAP values of predictors of changes in dispersion in energy productivity in XGBoost model and direction of marginal effect on model prediction



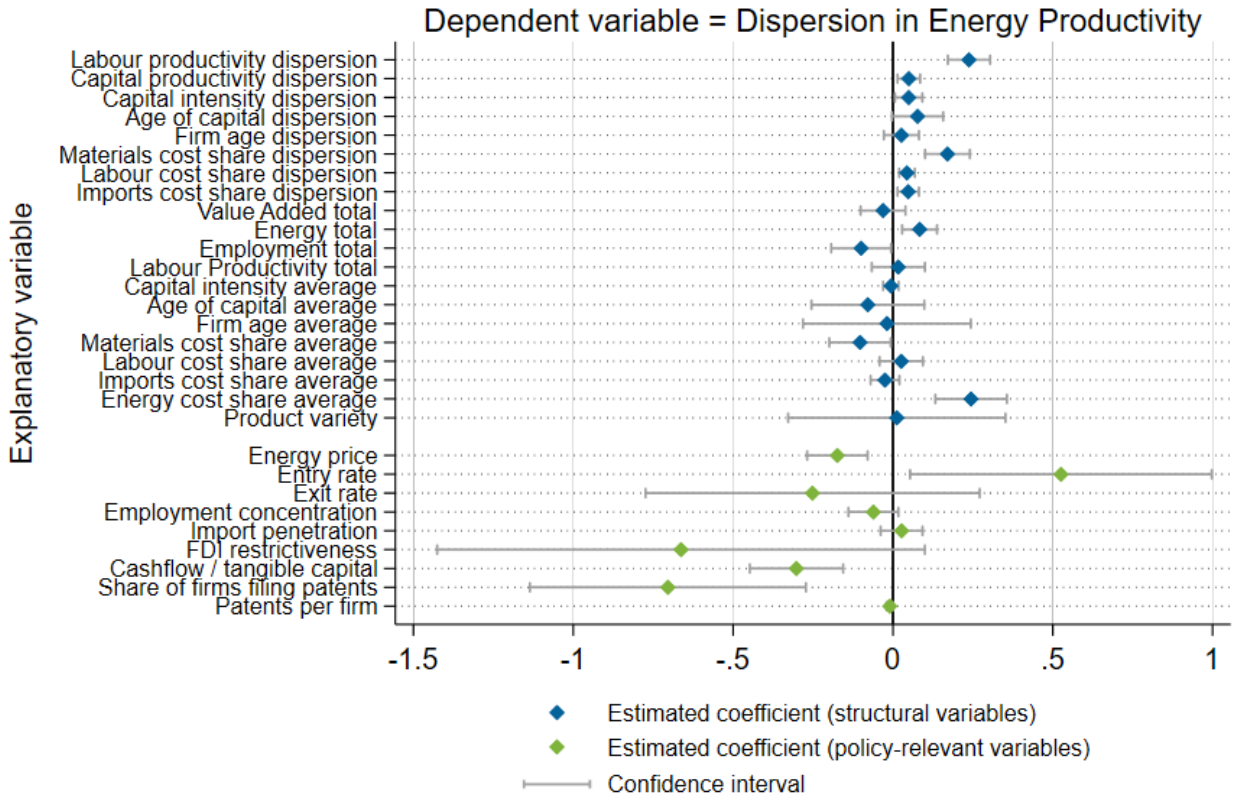
Note: The chart plots the mean of the absolute SHAP value for each predictor. The response variable is industry dispersion in energy productivity. Both the predictor and response variable are measured subtracting the country-year average value, computed separately within the training and test datasets. The prediction model is estimated using XGBoost. The sample is restricted to manufacturing industries with at least 10 firms which have at least 20 employees.
Source: authors' calculations.

Figure A B.27. Estimated coefficients from multivariate OLS regressions of energy productivity dispersion on explanatory variables



Note: The chart shows results from a multivariate regression of within-industry dispersion in energy productivity on various explanatory variables. The regressions are estimated by OLS on industry-country-year level data, with industries defined at the 3-digit level. Country-year fixed effects are included as controls. The sample is restricted to manufacturing industries and firms with at least 20 employees. Standard errors are robust to heteroskedasticity and 95% confidence intervals are plotted.
Source: authors' calculations.

Figure A B.28. Estimated coefficients from bivariate OLS regressions of energy productivity dispersion on explanatory variables estimated in 5 year changes



Note: The chart shows results from regression estimates of 5-year differences of within-industry dispersion in energy productivity on 5 year differences of various explanatory variables. The regressions are estimated by OLS on industry-country-year level data, with industries defined at the 3-digit level. Each row is estimated as a separate bivariate regression. Country-year fixed effects are included as controls. The sample is restricted to manufacturing industries and firms with at least 20 employees. Standard errors are robust to heteroskedasticity and 95% confidence intervals are plotted.
Source: authors' calculations.