INFER WORKING PAPER SERIES

INTERNATIONAL NETWORK FOR ECONOMIC RESEARCH

No. 12 | 2025

Bridging the Gap: Estimating Scope 3 Emissions at Company's Level

BARET Matilda *(University of Orléans)*
LUCOTTE Yannick *(University of Orléans)*
TOKPAVI Sessi *(University of Orléans)*

Website:
https://infer-research.eu/

Contact:
publications@infer.info

Bridging the Gap: Estimating Scope 3 Emissions at Company's Level*

Matilda Baret†1 Yannick Lucotte‡1,2,3,4 Sessi Tokpavi§1

[1]Université d'Orléans, LEO, F-45067, Orléans, France
[2]PSB Paris School of Business, Department of Economics, 59 rue Nationale, 75013 Paris, France
[3]Center of Research for Energy and Climate Change, 59 rue Nationale, 75013 Paris, France
[4]AFGES, 29 rue Lauriston, 75116 Paris, France

December 17, 2025

Abstract

The 21st century faces an accelerating climate change challenge, requiring deep, fast and sustainable reductions in greenhouse gas emissions. Among them, Scope 3 emissions covering indirect emissions across a company's entire value chain are critical but difficult to estimate due to scarce and unreliable data. This paper proposes a new empirical methodology to estimate firm-level Scope 3 emissions by integrating value chain dynamics and company-specific characteristics. Using input–output tables and sectoral emissions data, we reconstruct company value chains to capture upstream and downstream emissions. To address missing data, we apply both parametric models and machine learning techniques to estimate reported and unreported emissions. Using French firm-level data, our results suggest that company characteristics and sectoral emissions throughout the value chain strongly influence Scope 3 emissions. Machine learning models, particularly Random Forests, significantly outperform traditional models. Overall, our findings highlight the importance of improved emissions reporting and comprehensive climate policies to better manage emissions across all sectors.

JEL Classification: C51, C52, C53, Q52, Q54, Q58

Keywords: Climate Change, Climate Policy, Scope 3 Emissions, Value Chains, Machine Learning.

*We would like to warmly thank Crédit Agricole Centre Loire (CACL) for its financial support as part of the CACL-LEO Research Initiative "Energy Transition and Transformation of Economic Models".

†Université d'Orléans, LEO, F-45067, Orléans, France. Corresponding author. E-mail: matilda.baret@univ-orleans.fr.

‡ORCID ID: 0000-0001-8721-5845. Université d'Orléans, LEO, F-45067, Orléans, France & PSB Paris School of Business, Department of Economics, 59 rue Nationale, 75013 Paris, France & Center of Research for Energy and Climate Change (CRECC), 59 rue Nationale, 75013 Paris, France & AFGES, 29 rue Lauriston, 75116 Paris, France. E-mail: ylucotte@gmail.com

§Université d'Orléans, LEO, F-45067, Orléans, France. Corresponding author. E-mail: sessi.tokpavi@univ-orleans.fr

1 Introduction

One of the major challenges of the 21st century relates to the environmental crisis. The effects of climate change are intensifying rapidly and affect all components of our system (climatic, economic, and social). The consequences of climate change are alarming. Since 1900, global mean sea level has risen faster than in any previous century in at least 3000 years, and between 2010 and 2019, the melting of the Greenland and Antarctic ice sheets was four times greater than between 1992 and 1999. Global temperatures now exceed pre-industrial levels (1850–1900) by more than 1°C, and the trend continues to accelerate. In 2023, the report by the Intergovernmental Panel on Climate Change (IPCC, henceforth) highlights that human activities impact all key components of the climate system, with some responses occurring over decades and others over centuries. The IPCC's models suggest that greenhouse gas emissions involving global warming are man-made. Over the past 30 years, annual global emissions of greenhouse gases (GHG, henceforth) have surged by 50%, with more than 40% of historical emissions from human activities occurring in the last three decades alone. Experts therefore stress the urgent need to accelerate climate-change mitigation and adaptation efforts to bridge the gap between current commitments and what is required to ensure the stabilization of the climate.

Adopting targeted climate policies requires a precise understanding of GHG emissions. As described in the Greenhouse Gas Protocol[1], GHG emissions are categorized into three Scopes representing the different types of emission generated by a company throughout its operations and value chain. Scope 1 represents direct emissions produced by a company. Scope 2 emissions refer to indirect emissions associated with energy but not directly produced by a company's activities. Scope 3 emissions include indirect emissions occurring upstream and downstream from the value chain, broken down into fifteen separate categories. Companies are required to report Scopes 1 and 2, while Scope 3 reporting remains recommended but not mandatory. Reporting Scopes 1 and 2 is relatively straightforward because firms directly control or observe the necessary activity data. In contrast, Scope 3 emissions are difficult to measure due to limited visibility over indirect activities along the value chain. Various calculation methods and decision trees, offered by the Greenhouse Gas Protocol (Team (2011)), exist to quantify Scope 3 emissions, yet the process remains complex. For these reasons, the control of Scope 1 and 2 emissions has become the focus of the literature on decarbonisation and the achievement of GHG emission reduction targets for companies. Despite these challenges, Scope 3 emissions are increasingly recognized as strategically significant. They often represent the majority of a company's carbon footprint and provide insight into its dependence on fossil fuels. Managing Scope 3 emissions is therefore crucial for sustainable development, encompassing indirect impacts beyond direct operations.

[1] Ranganathan et al. (2004).

The literature on Scope 3 emissions is still limited. Most studies estimate emissions at the sectoral level, benefiting from the availability and reliability of aggregate data (see Huang et al. (2009), Hertwich and Wood (2018), Sparrevik and Utstøl (2020), Godin and Hadji-Lazaro (2022)). Estimating Scope 3 at the company level is more challenging due to data scarcity and reliability issues, which hinder the establishment of robust benchmarks. Research has predominantly focused on predicting Scope 1 and 2 emissions (e.g. Goldhammer et al. (2017), Han et al. (2021), Heurtebize et al. (2022) and Assael et al. (2022)), while only a few studies attempt to estimate Scope 3 at the company level (Nguyen et al. (2022) and Serafeim and Caicedo (2022)), and even fewer consider all three Scopes simultaneously (Nguyen et al. (2021)). Our study addresses this gap by proposing a new empirical approach to estimate Scope 3 emissions at the company level. The methodology integrates both the dynamics of value chains and company-specific characteristics, offering a robust and reliable estimation framework. This methodology can serve as a reference for companies, financial institutions, and policymakers, enabling more accurate carbon footprint assessments and designing targeted climate policies.

To achieve this, we implement a multi-step empirical strategy :

i) we tackle the issue of representing a company's value chain both upstream and downstream by reconstructing it using input-output tables, which provide a structured overview of transactions in goods and services by sector. This allows us to approximate inter-sectoral dependencies, even in the absence of confidential or non-existent company-level transaction data.

ii) we incorporate the emissions aspect into this framework as Scope 3 emissions comprise all emissions not directly controlled by the company. We assume that a company's Scope 3 emissions depend on the total emissions of its own sector as well as the total emissions of sectors upstream and downstream in its value chain, depending on the intensity of the interdependence of transactions between sectors.

iii) we establish a hierarchical structure for sectoral emissions: sectors closer to a company's own sector contribute more to its Scope 3 emissions than more distant sectors[2]. To capture variability not explained by sectoral flows, company-specific predictors are included, mitigating omitted variable bias and refining estimates.

iv) we estimate the model on the full sample, validate its robustness, and use it to predict unreported Scope 3 emissions, addressing data gaps that are particularly significant for small and medium-sized enterprises (SME, henceforth). This contributes to the establishment of a comprehensive benchmark and demonstrates its utility and effectiveness in supporting informed decision-making and policy formulation.

Our study focuses on France as a benchmark economy, using OECD Input-Output Tables

[2]Both upstream and downstream in the company's value chain.

(2021) to construct value chains and sectoral hierarchies. We focus on data ranging from 2010 to 2018, providing a sufficiently extensive study period. Sectoral emissions are also obtained from the OECD data, especially the Air Emissions Accounts database. For our dependent variable, Scope 3 emissions, we combine data from the Carbon Disclosure Project (CDP, henceforth) and Refinitiv, yielding a sample of 588 firms aligned with the Corporate GHG Emission Estimations (COGEM, henceforth) database from the Louis Bachelier Institute (ILB, henceforth). We use it as a reference for company selection and data processing. Finally, for the companies' specific predictors, data are extracted from Refinitiv to maintain consistency and avoid data fragmentation. We select fifteen variables covering descriptive, financial, and stock market metrics.

Given the potential multicollinearity between company predictors, sectoral emissions, and value chain data, we apply the Least Absolute Shrinkage and Selection Operator (LASSO, henceforth) to reduce dimensionality and improve estimation. We compare various models for predicting Scope 3 emissions: Ordinary Least Squares as a baseline, panel fixed-effects to control for unobserved heterogeneity, and various machine learning algorithms to capture complex relationships. Model performance is evaluated with multiple metrics, such as Root Mean Square Error (RMSE, henceforth), R-squared, and Mean Absolute Error (MAE, henceforth), alongside a Good-Fit measure to ensure robustness.

Our results reveal key findings for predicting Scope 3 emissions. First, company-specific characteristics such as operational and financial aspects significantly influence Scope 3 emissions, highlighting the importance of firm-level attributes. Interestingly, the LASSO selection process did not select sectoral emissions from the company's own sector as a key predictor, emphasizing the role of individual company attributes in shaping Scope 3 emissions. Then, our model suggests that upstream and downstream emissions from the value chain, regardless of their distance from the company's sector, contribute meaningfully to Scope 3 demonstrating that they cannot be overlooked in the design of empirical benchmarks for estimating these emissions. One climate policy recommendation should be that emissions reduction strategies should be applied across all sectors of the economy, not just within high-emission sectors, to promote an effective transition to a low-carbon economy. Moreover, machine learning models, particularly Random Forest, demonstrate superior predictive performance, balancing accuracy and generalization, making them suitable for estimating reported and unreported emissions. Random Forest thus appears as the best model to deliver robust and reliable predictions. Finally, a sectoral analysis was conducted to identify sectors with a balanced proportion of missing and reported data to enable reliable predictions. This demonstrates that reliable Scope 3 emissions estimates can be obtained even with relatively small datasets. However, improving prediction accuracy requires expanding the sample and enhancing data quality, which highlights the need for more comprehensive and systematic reporting as well as targeted climate policies.

The rest of the paper is structured as follows. The following section presents the literature review. Section 3 details the setup and specification of the model, while Section 4 describes the data used. Section 5 defines the estimation strategy and Section 6 provides empirical applications. The last Section concludes.

2 Review of literature

2.1 Climate Change and the Need for Immediate Action

The climate crisis stands as the most pressing challenge of the century, with unprecedented consequences. Climate data indicates rapidly rising sea levels, accelerating ice melt, and increasing global temperatures. Leading experts, such as the IPCC, attribute these changes primarily to human-driven GHG emissions. They emphasize the urgent need to reassess adaptation strategies in line with the Paris Agreement (2015) to prevent the worsening of extreme events such as droughts, floods, and biodiversity loss.

2.2 The Role of Companies in Addressing Scope 3 Emissions

Although the link between economic activities and climate change is widely recognized, several studies highlight the key role of companies in reducing GHG emissions, particularly Scope 3 emissions. Downie and Stubbs (2011) emphasize that neglecting Scope 3 can significantly under estimate a company's carbon footprint, as these emissions often represent the majority of total emissions. Their analysis of Australian firms reveals inconsistent reporting practices and calls for rigorous methodologies aligned with international standards. Similarly, Shrimali (2021) stresses the urgency of managing climate risks in the economy and financial sector through comprehensive carbon exposure assessments, including Scope 1, 2, and 3. While Scope 3 offers the greatest reduction potential, its measurement remains complex due to data limitations, methodological gaps, risks of double counting, and challenges in fairly allocating emissions across supply chains. Despite growing interest and efforts, the reliability of available methods remains a key issue. Both studies underline the urgent need for concrete climate targets and improved management of Scope 3 emissions to ensure meaningful progress toward climate goals.

Given this strategic importance and the methodological challenges surrounding Scope 3 emissions, the academic literature has evolved through several distinct phases, which we review below.

2.3 GHG Emission Estimation: From Parametric to Machine Learning

2.3.1 Early parametric approaches

Early work focused on estimating and predicting GHG emissions through linear regression models. Goldhammer et al. (2017) highlight the importance of corporate carbon footprints, often calculated using detailed internal data, that many firms do not disclose them. Their study estimates Scope 1 and 2 emissions externally for 93 European firms in high-emitting sectors[3], using predictors such as company size, vertical integration, capital intensity, production centrality, and the national energy mix's carbon intensity. Size, capital intensity, and production centrality emerged as key drivers, and the inclusion of sector-specific dummies produced an adjusted R-squared of 0.817, validating the model's external accuracy.

2.3.2 The shift toward machine learning

Due to the complexity of emissions and data limitations, machine learning approaches gradually gained prominence. Ma et al. (2021) apply Gaussian process regression to predict CO_2 emissions, demonstrating superior accuracy over traditional methods. Machine learning has been increasingly adopted for Scopes 1 and 2, where data are more available. Assael et al. (2022) develop a machine learning model to estimate Scope 1 and 2 emissions for non-reporting firms using Shapley values to enhance interpretability and identifying emission drivers at the firm level. Similarly, Han et al. (2021) and Heurtebize et al. (2022) confirm the superior predictive performance of machine learning methods for Scopes 1 and 2.

2.3.3 Current state and limitations

While these advances demonstrate the effectiveness of both parametric and machine learning approaches for Scopes 1 and 2 estimation, they share a common limitation. They focus almost exclusively on direct and energy-related emissions, leaving the more complex challenge of Scope 3 emissions largely unaddressed. This creates a significant gap, as Scope 3 emissions often represent the majority of a company's total carbon footprint.

2.4 The Scope 3 Challenge: Sectoral vs Firm-Level Approaches

2.4.1 Sectoral approaches: LCA, LCI, and EIO-LCA

Building on the recognition of Scope 3's importance, they became a major concern for managing climate change, as they often represent the majority of a company's total emissions. This opens up new opportunities for more significant, better targeted and more effective

[3] Chemicals, construction, engineering, and industrial machinery sectors.

emissions reductions. A key initial challenge lies in representing the full value chain, which is particularly complex at the firm level due to data limitations. As a result, much of the methodological literature focuses on sectoral-level approaches.

Two main frameworks stand out: Life Cycle Assessment (LCA, henceforth) and Life Cycle Inventory (LCI, henceforth). LCA evaluates the environmental impact of a product or service across its life cycle through phases like inventory analysis and impact assessment, allowing for iterative improvements in data quality[4]. LCI, the data-intensive component of LCA, quantifies inputs and outputs (resources, emissions, energy use) across numerous processes and stages of the supply chain, making it inherently complex[5].

A widely used method at the sectoral level is Economic Input-Output Life Cycle Assessment (EIO-LCA, henceforth), building on LCA with Input-Output Analysis from Leontief (Nobel Prize, 1973). EIO-LCA models inter-industry monetary flows to estimate emissions across supply chains. It links economic activities to environmental impacts through an intervention matrix, distinguishing direct (zero-order) and indirect (higher-order) environmental burdens. While EIO-LCA provides systematic, transparent, and scalable estimates, supported by freely available tools, it faces challenges such as price consistency, potential biases, and limited applicability at the firm level. Its upstream, sector-focused nature makes it less suitable to formulate company-level strategies and estimate unreported emissions (Wiedmann and Minx (2008)).

2.4.2 Firm-level applications of machine learning

As machine learning models proved effective in estimating GHG emissions, they began to be applied to Scope 3 emissions at the firm level. Nguyen et al. (2022) evaluated the reliability of Scope 3 data and machine learning's predictive capacity, revealing significant inconsistencies across data providers. Despite limited observations and varying predictor significance, machine learning improved accuracy by up to 6% for total emissions and 25% for individual categories. The authors emphasize the importance of considering data sources, quality, and prediction errors in risk assessments.

In the same vein, Serafeim and Caicedo (2022) used machine learning to predict 15 categories of Scope 3 emissions by combining financial data, Scopes 1 and 2 emissions, and industry classification. Adaptive Boosting outperformed linear regression and other supervised algorithms, showing the effectiveness of machine learning techniques. However, these models face limitations, including interpretability, sensitivity to feature selection, overfitting risks, computational demands, and reliance on high-quality data. Kerimov and Chernyshev (2022) reviews the increasing use of these models, their performance, limitations, and future

[4]See Rowley et al. (2009) for details and comparisons of the different approaches of the LCA.

[5]See Suh and Huppes (2005), Islam et al. (2016), Crawford et al. (2017) for a deeper understanding and comparison of the various approaches to LCI.

directions. Authors highlight data scarcity and model complexity as key issue when using machine learning for GHG estimation. To overcome the lack of labeled data, it suggests semi-supervised and self-supervised learning methods, emphasizing the importance of incorporating inductive bias.

Beyond academia, data providers such as Bloomberg (Quants (2022)) and CDP (CDP (2020)) also employ both machine learning and linear regression, to estimate both reported and unreported GHG emissions, illustrating the practical relevance of both advanced and linear models despite limited and uncertain data.

2.5 Research Gaps and Study Positioning

2.5.1 Synthesis of critical gaps

Despite substantial progress in GHG emission estimation, particularly for Scopes 1 and 2, several fundamental gaps remain in the literature on firm-level Scope 3 emissions:

i) A lack of methodological benchmarking and transparency, which makes it difficult to compare approaches and assess the robustness of existing Scope 3 estimation methods. Despite rapid growth in Scope 3 research, most studies focus on a single modeling approach, often machine learning, without systematically comparing it to alternative parametric methods or confront different machine learning techniques. Many also use proprietary or fragmented datasets, limiting reproducibility and hindering meaningful benchmarking. Consequently, it remains unclear which methods perform, how sensitive results are to variable selection, or how transferable models are across different firms and sectors. This absence of transparent and replicable frameworks leaves policymakers, financial institutions, and companies with little guidance on producing reliable and comparable estimates.

ii) Incomplete value chain representation, as existing approaches often fail to capture both upstream and downstream interactions at the firm level. While sectoral approaches like EIO-LCA capture upstream dependencies, they lack firm-level granularity and typically overlook downstream dynamics. Existing firm-level studies either rely on aggregate sectoral data or fail to systematically integrate both upstream and downstream value chain components into a unified framework.

iii)Absence of standardized variable selection, leading to inconsistent predictor choices and potential multicollinearity. Unlike Scopes 1 and 2, where key predictors are relatively well-established, Scope 3 estimation lacks consensus on predictor selection. Studies use varying combinations of financial, operational, and sectoral variables without systematic dimensionality reduction, leading to potential multicollinearity issues and reduced model generalizability.

iv) Limited frameworks for unreported emissions, leaving a blind spot for non-reporting firms. Most studies focus on improving estimates for firms that already report Scope 3 data. However, the majority of companies, particularly small and medium-sized enterprises, do not report these emissions. The literature lacks comprehensive methodologies to predict unreported Scope 3 emissions reliably, creating a significant blind spot for policymakers and financial institutions.

2.5.2 Our contributions and methodological positioning

In response to the growing importance of Scope 3 emissions and these identified gaps in the literature, we propose an innovative empirical methodology to estimate these emissions at the company level. Our study makes four key contributions that directly address the limitations identified above.

First, we establish a comprehensive benchmark focusing on France, providing a transparent and replicable framework for all economic stakeholders (companies, financial institutions, policymakers). To ensure data consistency, accuracy, and reproducibility, we use publicly available sector-level predictors from a single, major international source: the Organization for Economic Cooperation and Development (OECD, henceforth). Unlike previous studies that rely on proprietary or fragmented data sources, our approach ensures full replicability and can be easily extended to other economies.

Second, as highlighted in the literature, developing a methodology aligned with international protocols and able to capture the complexity of Scope 3 emissions at the company level is crucial. As Scope 3 emissions cover all GHG generated along the entire value chain, we propose an empirical framework that fully integrates the entire value chain, both upstream and downstream. Due to limited data on inter-company transactions, we adopt Input-Output Analysis, specifically the Leontief inverse matrix for upstream dependencies and the Ghosh inverse matrix for downstream dynamics, often overlooked. While traditionally applied at the sectoral level, we adapt these matrices to formally incorporate, both upstream and downstream, company level interactions into our empirical specification. This addresses one of the most critical gap by providing a complete value chain representation at the firm level. Beyond representing the value chain in line with international calculation protocols, we also account for the relative importance of Scope 3 emissions in relation to companies' commercial objectives. To this end, we formulate specific assumptions (detailed in the next section) about the structure, intensity, and influence of predictor variables on Scope 3 emissions, enhancing the accuracy and relevance of our estimates.

Third, we systematically compare parametric approaches with machine learning models. While machine learning techniques dominate recent Scope 3 literature, they face notable challenges, particularly in interpretability and model specification flexibility. Parametric methods, by contrast, remain widely used by major data providers and offer advantages in transparency and replicability. By comparing these approaches, we aim to establish a robust benchmark that delivers accurate, credible, and easily implementable results for a wide range of stakeholders. Given the large number of predictors in our dataset, we apply LASSO for variable selection to reduce dimensionality, identify the most influential variables driving Scope 3 emissions, and address multicollinearity issues. This technique contributes in two important ways: it provides a standardized approach to predictor selection and enhances estimation precision by reducing overfitting risks while enabling more parsimonious parametric specifications.

Fourth, to develop a comprehensive benchmark, we extend our analysis to predict unreported Scope 3 emissions. Beyond challenges related to data reliability, there is a major lack of Scope 3 data, particularly for SMEs. By enabling estimation of both reported and unreported Scope 3 emissions, our approach supports economic actors and regulators in making reliable predictions even when data are missing. To the best of our knowledge, we are the first to provide a detailed empirical framework to estimate both reported and unreported Scope 3 emissions step-by-step, combining value chain reconstruction, systematic variable selection, and model comparison within a unified methodology. By tackling missing data through a theoretically grounded and empirically validated framework, our approach enhances the accuracy of emission estimates and supports the development of more targeted and realistic climate policies across the entire production system.

In summary, our study addresses key gaps in the literature by providing a transparent and replicable framework for estimating Scope 3 emissions at the firm level, fully integrating both upstream and downstream value chains, systematically comparing parametric and machine learning approaches with standardized variable selection, and extending the analysis to unreported emissions. Together, these contributions establish a comprehensive benchmark that not only advances academic understanding but also offers practical tools for economic actors and policymakers to assess and manage corporate carbon footprints more effectively.

3 Setup and model's specification

This section presents our empirical framework for estimating Scope 3 emissions at the company level. We develop a progressive specification that explicitly models value chain dependencies through four key assumptions. Starting from a baseline where emissions depend only on own-sector emissions (Assumption 1), we successively incorporate upstream sectors (Assumption 2), downstream sectors (Assumption 3), and sectoral proximity weighting (Assumption 4). We then augment this specification with company-specific predictors to capture within sector heterogeneity. Finally, we detail how we operationalize value chain hierarchies using input-output matrices.

3.1 Notation and baseline setup

Formally, let i, $i = 1, ..., n$ be the index of a given company operating in the sector j, with n and K the total number of companies and sectors, respectively. We denote t the time index indicating a given year, and let $y_{i,t}$ be the level of company i Scope 3 emissions at time t which is the target variable to be predicted.

3.1.1 Sectoral dependence

We first make the following assumption :

Assumption 1 *The level of company i's Scope 3 emissions at time t depends on the emissions of its own sector j.*

The baseline that follows is:

$$y_{i,t} = \alpha_i + \gamma_0 \bar{y}_{j,t} + \varepsilon_{i,t}, \tag{1}$$

where α_i is the company-specific effect, $\bar{y}_{j,t}$ is the level of emissions of the sector j of company i and $\varepsilon_{i,t}$ is the error term.
This initial specification captures the notion that a company's Scope 3 emissions, encompassing its entire value chain, are shaped by the emissions of its sector. This reflects the fact that companies in the same sector, sharing similar suppliers and processes, are influenced by sector-level emissions. Large or leading companies in a sector can indirectly affect others due to their significant contribution to total emissions.

3.1.2 Upstream value chain effects

This baseline specification establishes own-sector emissions as the primary driver. However, Scope 3 emissions by definition encompass the entire value chain. We therefore extend the model to capture upstream dependencies.

Assumption 2 *The emissions of the sectors upstream in the production value chain of company i, also determine the level of its Scope 3 emissions at time t.*

The specification becomes:

$$y_{i,t} = \alpha_i + \gamma_0 \bar{y}_{j,t} + \sum_{k=1}^{K-1} \delta_k \bar{y}_t^{(k)} + \varepsilon_{i,t}, \tag{2}$$

where $\bar{y}_t^{(k)}$ denotes the level of emissions of the k^{th} sector of company i, upstream in its production value chain. Hence, $\bar{y}_t^{(1)}$ is the level of emissions of the first sector, upstream, of the company i, $\bar{y}_t^{(2)}$ is the counterpart from the second sector, and finally $\bar{y}_t^{(K-1)}$ is the counterpart from the last sector.

Upstream sectors provide the inputs necessary for the company's activity. Suppliers from high-emission sectors, such as metallurgy or energy, contribute to higher Scope 3 emissions, whereas less carbon-intensive sectors contribute less. Companies with similar supply chains inherit the environmental impact of these sectors, so aggregated upstream emissions reliably measure the overall carbon intensity of the upstream value chain. For example, a company in automotive depends on steel and energy sectors, which largely determine its upstream emissions.

3.1.3 Downstream value chain effects

Having incorporated upstream dependencies, we now extend the framework to include downstream value chain effects, which reflect how a company's products are used and disposed of by customers and end-users.

Assumption 3 *The emissions of the sectors downstream in the production value chain of company i, also determine the level of its Scope 3 emissions at time t.*

This assumption calls for an extension of the specification in (2), with:

$$y_{i,t} = \alpha_i + \gamma_0 \bar{y}_{j,t} + \sum_{k=1}^{K-1} \delta_k \bar{y}_t^{(k)} + \sum_{k=1}^{K-1} \theta_k \bar{z}_t^{(k)} + \varepsilon_{i,t}, \tag{3}$$

where $\bar{z}_t^{(k)}$ is the level of emissions of the k^{th} sector of company i, downstream in its production value chain. Therefore, $\bar{z}_t^{(1)}$ is the level of emissions of the first sector, downstream, of the company i, $\bar{z}_t^{(2)}$ is the counterpart from the second sector, and finally $\bar{z}_t^{(K-1)}$ is the counterpart from the last sector.

Downstream emissions capture environmental impacts from the use and end-of-life of the company's products. For instance, sectors like transport or construction shape Scope 3 emissions of companies supplying fuel or construction materials. A company transmits part of its carbon footprint to downstream sectors that use its inputs. A steel producer transmits part of its carbon footprint to downstream sectors, such as automotive or real estate reflecting the aggregated carbon intensity of the downstream value chain.

3.1.4 Proximity ranking of sectors

At this stage, based on specification (3), the model assumes that all companies within the same sector have identical value chains, both upstream and downstream. To introduce heterogeneity and enhance the model's representativeness at the company level, we now incorporate a sectoral proximity ranking mechanism.

Assumption 4 *The emissions of the closest sectors, upstream and downstream, to the sector j of company i contribute more to the level of its Scope 3 emissions, compared to the least close sectors which contribute less.*

$$y_{i,t} = \alpha_i + \gamma_0 \bar{y}_{j,t} + \sum_{k=1}^{K-1} \delta_k \bar{y}_{j,t}^{(k)} + \sum_{k=1}^{K-1} \theta_k \bar{z}_{j,t}^{(k)} + \varepsilon_{i,t}, \quad (4)$$

where $\bar{y}_{j,t}^{(k)}$ represents the level of emissions of the k^{th} closest sector to the sector j of company i, upstream in its production value chain and $\bar{z}_{j,t}^{(k)}$ the level of emissions of the k^{th} closest sector to the sector j of company i, downstream in its production value chain. In other words, this proximity assumption is embedded in specification (3) through the hierarchical ordering of sectors. The ranking $(k = 1, 2, ..., K-1)$ reflects sectoral proximity, with $k = 1$ denoting the closest sector and $k = K-1$ the most distant. This hierarchy is constructed using input-output matrices, as detailed below.

This assumption introduces variability to the value chain by emphasizing that a company's Scope 3 emissions are mainly influenced by sectors closest to its own sector in its value chain, both upstream and downstream, while more distant sectors have a limited effect. Upstream, supplier sectors providing critical inputs play a central role. Downstream, focus is on sectors that directly use the company's products, as they significantly affect Scope 3 emissions. These closest sectors are key due to dense economic and environmental interactions, which facilitate direct transmission of emissions. Their influence is reinforced by integrated supply chains and shared characteristics, whereas distant sectors, less connected and often differing in processes or carbon intensity, have a weaker impact.

3.1.5 Company-specific characteristics

It is worth noting that in specification (4) the predictive variables $\bar{y}_{j,t}^{(k)}$ and $\bar{z}_{j,t}^{(k)}$ are defined at the sector level, and are the same for two companies in the same sector. Finally, to capture variability at the company level, we consider some company's balance-sheet variables in the vector $X_{i,t}$ as additional predictors:

$$y_{i,t} = \alpha_i + \beta X_{i,t} + \gamma_0 \bar{y}_{j,t} + \sum_{k=1}^{K-1} \delta_k \bar{y}_{j,t}^{(k)} + \sum_{k=1}^{K-1} \theta_k \bar{z}_{j,t}^{(k)} + u_{i,t}, \quad (5)$$

where β is a vector of parameters.

Including individual company characteristics is crucial to ensure accuracy and account for company-specific factors, avoiding bias from omitted variables and enhancing variability. In the literature, it is standard to incorporate company-level predictors when estimating direct (Scope 1) and indirect (Scope 2 and 3) emissions at company level. Descriptive and financial data are the most widely used variables, as they capture operational and financial aspects of a company (Goldhammer et al. (2017), Assael et al. (2022), Heurtebize et al.

(2022), Nguyen et al. (2022) and Serafeim and Caicedo (2022)). These variables reflect the company's size, profitability, asset base, location, and use of assets, capturing its financial structure. Geographical variables can also be included when studies cover several countries, to account for the environment in which a company operates and international trade patterns; typically, this is done using company locations or inter-firm distances (Heurtebize et al. (2022) and Serafeim and Caicedo (2022)). Company sector is another key variable. Due to limited data on inter-company transactions, including sector information is essential. This can be done by grouping companies by sector (Han et al. (2021), Nguyen et al. (2022), and Serafeim and Caicedo (2022)) or by integrating a company's primary sector, activity, or industry classification (or the proportion) by classifying companies using classification codes (Assael et al. (2022)). Aligned with the literature and the objectives of our analysis, we use data reflecting operational and financial characteristics of companies, as detailed in Section 4. Sector emissions are incorporated using the OECD Input-Output Tables (2021) sector classification and companies' upstream and downstream sectoral value chains. Geographical variables are excluded since the analysis is limited to a single country, removing the need to account for country-specific or international trade effects.

In summary, our final specification integrates sectoral emissions throughout the entire value chain, ranked by proximity, and includes company-specific characteristics, providing a comprehensive framework for estimating Scope 3 emissions at the company level.

3.2 Sectoral emissions hierarchy construction

A natural question follows: for a company i operating in sector j, how can we determine the hierarchy of sectors according to their contribution to its Scope 3 emissions? Having set out our empirical specification, we now address the practical issue of constructing the sectoral rankings $\bar{y}_{j,t}^{(k)}$ and $\bar{z}_{j,t}^{(k)}$.

3.2.1 Upstream Value-Chain Ranking from the Leontief Inverse

A simple approach is to make the assumption that this hierarchy depends on the productive structure of the company's sector j on the demand side, or say differently, the inter-sectoral flows (inputs) from each of the $K-1$ sectors to the gross output of the company's sector j. Thus defined, this hierarchy and by extension the variables $\bar{y}_{j,t}^{(k)}$, can be inferred via the Input-Output model of Leontief and specifically the Leontief inverse matrix. As discussed above[6], with the market balance equation $X = AX + F$, where X is the vector of output, F the vector of final demand, and A the so-called technical coefficients matrix whose elements are the elements of the transactions matrix divided by the total of their corresponding

[6]Section 2 devoted to the review of literature.

column, the Leontief inverse matrix L is defined as follows:

$$X = (I - A)^{-1}F = LF. \tag{6}$$

For the sector j of company i, the elements $l_{o,j}$ of the Leontief inverse matrix L, with $o = 1, ..., K$ and $o \neq j$, represent the amounts of gross output from sector o that are produced to satisfy a unit of final demand from sector j. These are basically the off-diagonal entries of the j^{th} column of the matrix L. Their ranking from the largest value to the smallest value allows to identify the hierarchy of sectors upstream to the sector j of company i, and then the construction of the emissions' variables $\bar{y}_{j,t}^{(k)}$ by matching the corresponding level of emissions.

3.2.2 Downstream Value-Chain Ranking from the Ghosh Inverse

Similarly, the identification of the hierarchy of sectors, downstream, is necessary, and goes through the Ghosh inverse matrix derived from the Ghosh model[7]. This model is a type of supply-side Input-Output model which can be described as an alternative of the demand-side model of Leontief. Starting with the identify $X' = BX' + V'$, with V the value added or primary factors, B the distribution coefficients matrix calculated by the elements of the transactions matrix M divided by the total of their corresponding row, the Ghosh inverse matrix G is defined as:

$$X' = (I - B)^{-1}V' = GV'. \tag{7}$$

For the sector j of company i, the elements $g_{j,o}$ of the Ghosh inverse matrix G, correspond to the amounts of gross inputs from sector j that are absorbed by sectors o, $o = 1, ..., K$, $o \neq j$, and are the off-diagonal entries of the j^{th} row of the matrix G. As above, ranking these entries from the largest value to the smallest helps identifying the hierarchy of sectors downstream to the sector j of company i, and the construction of the predictive variables $\bar{z}_{j,t}^{(k)}$ by aligning with the corresponding emission levels[8].

Taken together, the two Input-Output based hierarchies provide a consistent and operational mapping of the sectors most closely connected to a company's activity, upstream and downstream. By ranking the relevant coefficients of the Leontief and Ghosh inverse matrices and matching them with sectoral emissions, we construct the ordered variables $\bar{y}_{j,t}^{(k)}$ and $\bar{z}_{j,t}^{(k)}$, capturing the carbon intensity of the sectors most economically intertwined with firm i's sector j. This ranking-based construction allows the model to account for differentiated value-chain influences, moving beyond the assumption of homogeneous sectoral structures and enabling a more granular and realistic estimation of company-level Scope 3

[7] See Section 2 for discussion.

[8] A concrete example of sectoral emissions hierarchy is given in the appendix A.1

emissions. Importantly, treating sectors through a hierarchy, rather than applying weights or aggregating contributions, avoids double counting: each sector appears only once and is handled independently, allowing a full reconstruction of the value chain without overlap. This structure also naturally integrates cross-cutting sectors such as energy, whose influence extends beyond Scope 2 and affects multiple Scope 3 categories. Overall, by combining Input-Output based proximities with sectoral emissions, the method captures the full complexity of Scope 3 emissions while ensuring a coherent and non-redundant representation of the entire value chain.

4 Data

4.1 Sectoral hierarchy structure

This analysis uses the OECD (2021) Input-Output Tables, which describe the interdependencies between producers and consumers within an economy. These tables capture flows of both final and intermediate goods and services, either by sector ("industry by industry tables") or by product ("product by product tables"). For our purposes, we adopt an industry-by-industry approach, relying on the OECD harmonized national Input-Output Tables that present matrices of intersectoral flows of domestically produced and imported goods and services.

In this framework, each sector plays a dual role: it purchases goods and services from other sectors and factors for its production, and it sells the commodities it produces to other sectors. From the OECD (2021), Input-Output Tables, we derive the $n \times n$ intersectoral transaction matrix M, where n is the number of sectors in the economy. This matrix summarizes the monetary transactions between sectors, providing a comprehensive view of the economy's structure and the interactions among sectors.

		Buying Sector				
		1	. . .	*j*	. . .	*n*
	1	m11	. . .	*m1j*	. . .	m1K
	⋮	⋮		⋮		⋮
Selling Sector	*i*	mi1	. . .	*mij*	. . .	miK
	⋮	⋮		⋮		⋮
	n	mK1	. . .	*mKj*	. . .	mKK

Figure 1: **Intersectoral transaction matrix** M

In the transaction matrix M, column entries represent the inputs required by each sector, quantifying the monetary value of inputs used, while row entries show the outputs produced by each sector, representing the value of outputs it generates. Diagonal entries capture internal dependencies, reflecting a sector's reliance on its own output. This matrix forms the basis for computing technical and distribution coefficients, which are then used to construct the Leontief and Ghosh models. From these, we derive the Leontief and Ghosh inverse matrices, representing a company's value chain both upstream and downstream, and use them to construct the variables $\bar{y}_{j,t}$, $\bar{y}_{j,t}^{(k)}$, and $\bar{z}_{j,t}^{(k)}$. This approach captures each company's upstream and downstream connections, enabling a detailed, realistic, and non-redundant estimation of company-level Scope 3 emissions.

Overall, the OECD (2021) Input-Output Tables cover 45 sectors[9] in current prices (USD million) for all OECD countries and various non-OECD economies from 1995 to 2018. For our analysis, we focus on the French economy over the period 2010–2018.

4.2 Sectoral greenhouse gas emissions, upstream and downstream greenhouse gas emissions in production value chain

In order to obtain sectoral GHG emissions data for our analysis, we extract data from the OECD specifically using the "Air Emissions Accounts" database from the sector "Air and Climate" within the category "Environment and Climate Change". This database details atmospheric emissions resulting from production and consumption processes across OECD and non-OECD countries from 1990 to 2023. Productive activities are categorized by economic sector using the ISIC Rev.4 classification[10], while household consumption activities are classified by purpose, such as transport, heating, and other activities.

Air emissions include both individual GHG as well as air pollutants. They are also aggregated to reflect environmental pressures using equivalence factors capturing global warming potential, acidifying gases, and ozone precursors. The OECD compiles this database from multiple national sources and aligns it with the System of Environmental-Economic Accounting (SEEA) framework. Air emissions are assigned to a national economy based on the residence of the economic entities involved and include bridging items to the UNFCCC national inventories, which follow the territory principle.

For our analysis, we focus on greenhouse gas emissions in tonnes of CO2-equivalent for France from 2010 to 2018, covering the six main greenhouse gases responsible for global warming. Assuming that a company's Scope 3 emissions are linked to the emissions of its sector, these data form the basis for constructing sectoral emissions variables, which are then matched to company sectors to create the variable $\bar{y}_{j,t}$. Further, we assume that a company's Scope 3 emissions depend not only on its own sector's emissions but also on

[9]Appendix A.2 provides sector details. We aggregated the "Mining and quarrying" sector, which is broken down into three separate sectors to form a single sector, to match sectoral greenhouse gas emissions data. We also do not take into account the last sector, namely "Activities of households as employers; undifferentiated goods- and services- producing activities of households for own use" since it does not emit any greenhouse gases.

[10]The same as OECD (2021), Input-Output Tables with greater disaggregation except for "Mining and quarrying" sector.

upstream and downstream sector's emissions within its production value chain. To formalize this hierarchy, we employ the Leontief and Ghosh inverse matrices derived from OECD (2021) Input-Output Tables. These matrices allow us to rank sectors by inter-sectoral intensity from largest to smallest. We extract the off-diagonal entries in the jth (or ith) column (row) of the Leontief (Ghosh) inverse matrix, representing the upstream (downstream) production value chain, and rank them accordingly. Matching GHG emissions to this sectoral hierarchy produces the variables $\bar{y}_{j,t}^{(k)}$ and $\bar{z}_{j,t}^{(k)}$, representing the emissions of the k^{th} closest sector upstream and downstream of sector j for company i. This approach captures the influence of total sectoral emissions on a company's indirect emissions, fully aligning with our assumptions outlined in Section 3.

4.3 Dependent variable: Scope 3 emissions

To achieve our objectives, we rely on Scope 3 emissions data from two major providers: CDP and Refinitiv. Using these sources, we compiled a panel of 588 French companies covering the period between 2010 and 2018, aligned with the COGEM database developed by ILB in partnership with Pladifes. This database serves as our reference dataset for selecting companies and processing our target variable, Scope 3 emissions.

We use these two providers for several reasons. They are the main sources widely used in academic literature for estimating both direct (Scope 1) and indirect (Scope 2 and 3) emissions[11], and are also used in practice, notably by ILB. Using them is particularly important because data reliability is a major challenge when working with Scope 3 emissions: these emissions require modeling the entire value chain, firms lack standardized reporting benchmarks, and disclosure remains optional. As highlighted by Downie and Stubbs (2011) and Nguyen et al. (2022), this results in inconsistent measurements and significant uncertainty. As the two leading providers of emissions data, CDP and Refinitiv offer a sufficiently reliable foundation for robust estimation. Finally, many companies do not disclose their Scope 3 emissions, creating missing values. This is due to the complexity of measuring these emissions, the lack of standardized methodologies, and the fact that reporting is not yet mandatory. To mitigate this issue, we combine CDP and Refinitiv data, filling gaps, increasing the number of observations, and reducing uncertainty.[12]

Our estimation of Scope 3 emissions relies on the assumption that a company's Scope 3 emissions depend on its sector as well as upstream and downstream sectors within its production value chain. To formalize this, we construct sectoral hierarchies structures using Leontief and Ghosh inverse matrices from OECD (2021) Input-Output Tables, and match them with sectoral greenhouse gas emissions from the OECD "Air Emissions Account" database. Company sectors are then matched to the OECD classification primarily using

[11] Assael et al. (2022), Nguyen et al. (2022), and Serafeim and Caicedo (2022)

[12] Shrimali (2021) suggests that integrating multiple sources improves predictive quality.

French activity nomenclature codes retrieved from the business register identification system, supplemented by company names, ISIN codes, and tickers, and verified via the UNSD website to assign corresponding ISIC 4 categories.

4.4 Company-specific predictors

Company-specific characteristics are sourced from Refinitiv, allowing us to rely on a single, comprehensive data provider and ensure consistency across all company-level variables. Following the variables commonly used in the literature[13], we include the following variables for each company from 2010 to 2018: Net sales, Number of employees, Net sales to gross fixed assets ratio, Total assets, Total debt, Earnings before interest and taxes (EBIT), Earnings before interest taxes depreciation, and amortization (EBITDA), Net property, plant, and equipment, Research and development ($R\&D$) expenditures, Return on equity (percentage), Cash flow to sales ratio, Market-to-book value ratio, Beta, Price and Volatility. These variables capture key dimensions of a firm's financial and operational profile, strengthening the robustness of our analysis.

All in all, our dataset contains Scope 3 emissions for 588 firms, obtained from Refinitiv and CDP and aligned with the COGEM database. Companies are assigned to sectors based on the OECD (2021) Input-Output Tables[14]. The panel covers the period 2010–2018, yielding 5,292 firm-year observations for the dependent variable. In addition to the 15 firm characteristics, the dataset includes 83 variables capturing sectoral emissions, upstream and downstream emissions derived from the Input-Output based sectoral hierarchy.

5 Estimation strategy

5.1 Data processing

Our empirical strategy begins with a data pre-processing step, where we address missing values. We focus on both the target variable and company-specific predictors, as there are no missing values for sectoral, upstream, or downstream emissions. For the target variable, we exclude all observations with missing values, without applying any imputation, to ensure the quality and reliability of our model. By relying only on raw, reported data, our regressions reflect the actual availability of Scope 3 emissions, providing a realistic benchmark and avoiding the introduction of artificial information. For the company-specific predictors, we apply K-Nearest Neighbors (KNN) imputation. Given that the proportion of missing data in these predictors is relatively small, this technique effectively completes the dataset without introducing significant bias. This step is crucial because it ensures that regressions

[13]See Section 3 for discussion.
[14]See appendix A.3 for details of the company structure.

are performed on complete data, reducing potential bias and estimation errors. Moreover, many models, especially machine learning algorithms, cannot be trained on datasets with missing values, as incomplete outcome or predictor data can hinder learning. By addressing missing values in this way, we also facilitate the interpretation of results and enable a more accurate assessment of model performance.

5.2 Automatic variable selection model

We first examined correlations among our predictors, given the dataset includes 15 company-specific variables and 83 predictors representing sectoral and value chain emissions. This raises multicollinearity concerns, particularly within value chain emissions, as overlapping upstream and downstream sector information could bias results. A correlation matrix analysis[15] confirmed substantial multicollinearity, notably between upstream and downstream emissions and some company-specific variables.
To address the curse of dimensionality and mitigate multicollinearity, we employ LASSO, introduced by Tibshirani in 1996[16], allowing to identify the key drivers of Scope 3 emissions. LASSO is a regularization method designed for variable selection in high-dimensional settings. It enhances model generalization and prevents overfitting by penalizing model complexity through an L_1 norm penalty added to the standard linear regression loss function, shrinking coefficients of irrelevant variables on the target to zero. LASSO simultaneously minimizes prediction error and performs automatic variable selection, effectively reducing dimensionality and multicollinearity while identifying the most relevant explanatory variables[17]. Applied to our 98 predictors, the LASSO loss function is:

$$L(\beta) = \sum_{i=1}^{N} \sum_{t=1}^{T} (y_{i,t} - X_{i,t}\beta)^2 + \lambda \sum_{k=1}^{98} |\beta_k| \tag{8}$$

where $y_{i,t}$ is Scope 3 emissions for company i in year t, $X_{i,t}$ the vector of explanatory variables, β the vector of coefficients to be estimated, λ the L_1 regularization parameter and β_k is the coefficient associated with each explanatory variable, N the number of companies and T the number of years.
The coefficients $\hat{\beta}$ are obtained by minimizing the loss function, which entails solving the following optimization problem:

$$\hat{\beta} = \arg\min_{\beta} L(\beta) = \arg\min_{\beta} \left(\sum_{i=1}^{N} \sum_{t=1}^{T} (y_{i,t} - X_{i,t}\beta)^2 + \lambda \sum_{k=1}^{98} |\beta_k| \right) \tag{9}$$

[15] The correlation matrix is 98x98; further details are available upon request.
[16] Tibshirani (1996).
[17] See Ranstam and Cook (2018) for details about the LASSO methodology.

The optimal λ is chosen by cross-validation, yielding a subset of explanatory variables for parametric and machine learning models.

LASSO emerged as the most suitable method for our purposes, as it simultaneously reduces dimensionality and addresses multicollinearity through automatic variable selection, thereby identifying the key drivers of Scope 3 emissions. Ridge regression employs an L_2 penalty that shrinks coefficients toward zero but does not eliminate them, limiting its ability to simplify the model through variable selection. Elastic Net combines both L_1 and L_2 penalties, often retaining groups of correlated variables, making it less stringent than LASSO in achieving parsimonious specifications. Principal Component Analysis (PCA) addresses dimensionality and multicollinearity by transforming predictors into uncorrelated linear combinations, but sacrifices interpretability by discarding the original variables and operates independently of the target variable. In contrast, LASSO optimizes variable selection based on explanatory power for the dependent variable while maintaining the interpretability of the original predictors.

5.3 Parametric and machine leaning models

Having identified the most relevant predictors through LASSO, we now turn to the estimation of Scope 3 emissions. We apply a set of parametric and machine-learning models to evaluate how different methodological choices affect predictive accuracy and robustness. This comparative approach enables us to benchmark traditional econometric specifications against more flexible algorithms and to identify the model that offers the best balance between interpretability, in-sample fit, and out-of-sample performance.

5.3.1 Ordinary Least Square

We begin with an Ordinary Least Squares[18] model, which serves as a baseline against which more advanced methods can be evaluated. In our panel setting, the Ordinary Least Squares specification is written as :

$$y_{i,t} = \alpha + \sum_{k=1}^{K} \beta_k X_{i,t}^{(k)} + \varepsilon_{i,t} \tag{10}$$

where $y_{i,t}$ denotes Scope 3 emissions for company i in year t, $X_{i,t}^{(k)}$ refers to the k^{th} explanatory variable, K is the total number of predictors, β_k the associated coefficient and $\varepsilon_{i,t}$ the error term.

Ordinary Least Square estimates the coefficients by minimizing the sum of squared residuals.

[18] For a review of the methodology, see Chumacero (2012).

In our case, this amounts to solving:

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{N} \sum_{t=1}^{T} (y_{i,t} - \alpha - \sum_{k=1}^{K} \beta_k X_{i,t}^{(k)})^2 \tag{11}$$

where N is the number of companies and T the number of years.

Ordinary Least Squares is easy to implement and offers full interpretability of the estimated coefficients. However, it is limited in its ability to capture nonlinearities, interactions, or unobserved heterogeneity. Its performance can nonetheless be substantially improved when preceded by a variable selection step, making it a robust benchmark for evaluating more sophisticated models.

5.3.2 Panel Fixed Effects Model

We also estimate a panel fixed effects model[19]. The panel structure of our dataset, which follows multiple firms over several years, naturally lends itself to this specification. The panel fixed effects model is written as:

$$y_{i,t} = \alpha_i + \sum_{k=1}^{K} \beta_k X_{i,t}^{(k)} + \varepsilon_{i,t} \tag{12}$$

where $y_{i,t}$ represents Scope 3 emissions for company i in year t, α_i is a firm-specific time-invariant effect, $X_{i,t}^{(k)}$ is the k^{th} explanatory variable, K is the total number of predictors β_k is the associated coefficient, $\varepsilon_{i,t}$ is the error term.

The coefficients are estimated by minimizing :

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{N} \sum_{t=1}^{T} (y_{i,t} - \alpha_i - \sum_{k=1}^{K} \beta_k X_{i,t}^{(k)})^2 \tag{13}$$

where N is the number of companies and T the number of years.

To eliminate the individual effects α_i,we apply the within transformation :

$$(y_{i,t} - \bar{y}_i) = \sum_{k=1}^{K} \beta_k \left(X_{i,t}^{(k)} - \bar{X}_i^{(k)} \right) + (\varepsilon_{i,t} - \bar{\varepsilon}_i) \tag{14}$$

where $\bar{y}_i$ et $\bar{X}_i$ denote company-specific time averages. This transformation removes α_i yielding consistent estimates of β.

The associated minimization problem becomes :

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(y_{i,t} - \bar{y}_i - \sum_{k=1}^{K} \beta_k \left(X_{i,t}^{(k)} - \bar{X}_i^{(k)} \right) \right)^2 \tag{15}$$

[19]Both Fixed and Random Effects models were estimated. A Hausman test ($p < 0.05$) confirms that Fixed Effects is more appropriate, ensuring that time-invariant unobserved heterogeneity across firms is properly accounted for.

By controlling for all unobserved, time-invariant company characteristics, the panel fixed effect estimator mitigates omitted-variable bias and provides more reliable causal estimates[20]. In addition to these parametric approaches, we incorporated machine learning models as traditional parametric methods may fall short in capturing the potential complex relationships between Scope 3 emissions and predictors. We selected Random Forest[21] along with a set of tree-based boosting algorithms, including Extreme Gradient Boosting, Light Gradient Boosting and Categorical Boosting[22].

5.3.3 Random Forest

Random Forest is an ensemble learning method that builds a large number of decision trees to improve predictive accuracy and reduce overfitting. It is particularly well suited to capture nonlinear patterns and complex interactions among predictors, which makes it an attractive option for modeling Scope 3 emissions.
Random Forest relies on bootstrap aggregation (bagging): each tree is trained on a bootstrap sample of the dataset, and at every split a random subset of predictors is considered. This double source of randomness decorrelates the trees, thereby reducing variance and improving generalization performance. The final prediction is obtained by averaging the predictions of the individual trees (majority voting) in the case of regression.
In our context, the Random Forest prediction for company i in year t is given by :

$$\hat{y}_{i,t} = \frac{1}{B} \sum_{b=1}^{B} T_b(X_{i,t}^{(k)}) \tag{16}$$

where $\hat{y}_{i,t}$ is the predicted Scope 3 emissions for company i in year t, B is the number of trees in the forest, $T_b(X_{i,t}^{(k)})$ is the prediction of the b^{th} tree based on the explanatory variables, $X_{i,t}^{(k)}$ is the k^{th} explanatory variable, K is the number of predictors.
Each individual tree is grown by recursively selecting the split that minimizes the mean squared error (MSE) within the bootstrap sample. The overall Random Forest estimator can thus be viewed as minimizing the aggregated prediction error :

$$\hat{T} = \arg\min_{T_b} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(y_{i,t} - \frac{1}{B} \sum_{b=1}^{B} T_b(X_{i,t}) \right)^2 \tag{17}$$

Through the aggregation of many weakly correlated trees, Random Forest offers strong predictive performance while maintaining robustness to noise, nonlinearities and multicollinearity.

[20]For details on Panel models, see Kelejian and Piras (2017).
[21]For a detailed methodology refer to Breiman (2001).
[22]For a review of the methodology of the chosen boosting algorithms, see Bentéjac et al. (2021).

5.3.4 Boosting Algorithms

Boosting algorithms are ensemble methods that sequentially build decision trees to minimize residual errors, improving predictive accuracy and model refinement. We test three state-of-the-art boosting algorithms, Extreme Gradient Boosting, Light Gradient Boosting, and Categorical Boosting, which differ in growth strategies, optimization techniques, and handling of data structure.

Extreme Gradient Boosting is known for its speed and accuracy, using gradient-based optimization combined with regularization techniques to prevent overfitting. Trees are built sequentially using a "level-wise" growth strategy, where each new tree is trained to correct the residual errors of the previous iteration. At each iteration, a new tree is trained to correct the errors of the previous model. In our application, we use the mean squared error (MSE) as the cost function for regression. Light Gradient Boosting is designed for efficiency, large datasets, and memory efficiency. Trees are also built sequentially, but Light Gradient Boosting uses a "leaf-wise" growth strategy, splitting the leaf that maximally reduces the loss at each step. The functioning and mathematical specification is identical to Extreme Gradient Boosting; the distinction lies in tree construction and growth strategy. Initially developed to handle categorical variables via target encoding with permutations, Categorical Boosting also performs well with numerical data. It builds trees using an ordered, sequential update method, which preserves the integrity of training and mitigates overfitting. Advanced regularization ensures strong generalization even on complex datasets.

All three algorithms belong to the gradient boosting family. For clarity, we adopt a single general specification :

$$\hat{y}_{i,t} = f(X_{i,t}^{(k)}; \Theta) \tag{18}$$

where $\hat{y}_{i,t}$ represents the estimated Scope 3 emissions for company i in year t, $X_{i,t}^{(k)}$ is the k^{th} explanatory variable, K is the number of explanatory variables, $f(\cdot)$ is a non-parametric function constructed from a set of optimized decision trees, and Θ is the set of model parameters including learning rate, regularization, tree depth, and sampling rate.

The associated loss function is:

$$L(\Theta) = \sum_{i=1}^{N} \sum_{t=1}^{T} (y_{i,t} - \hat{y}_{i,t})^2 + \Omega(\Theta) \tag{19}$$

where $\Omega(\Theta)$ is a regularization term controlling tree complexity. Following LASSO-based variable selection, we use L_2 regularization on the leaf weights :

$$\Omega(\Theta) = \lambda \sum_{j} w_j^2 \tag{20}$$

with w_j the leaf scores and λ the L_2 regularization parameter.
The final optimization problem is:

$$\hat{\Theta} = \arg\min_{\Theta} \sum_{i=1}^{N} \sum_{t=1}^{T} (y_{i,t} - \hat{y}_{i,t})^2 + \Omega(\Theta) \tag{21}$$

This formulation allows us to compare the predictive performance of different boosting strategies while controlling for overfitting and capturing complex relationships between Scope 3 emissions and predictors.

5.4 Evaluation metrics

The final crucial step involves evaluating the predictive performance of our models to identify which one delivers the highest accuracy in predicting Scope 3 emissions. In line with existing literature[23], we rely on a range of evaluation metrics to capture various dimensions of model effectiveness.

5.4.1 Root Mean Squared Error

First, RMSE which measures the square root of the average squared differences between actual and predicted values:

$$\text{RMSE} = \sqrt{\frac{1}{N \times T} \sum_{i=1}^{N} \sum_{t=1}^{T} (y_{i,t} - \hat{y}_{i,t})^2}$$

where $y_{i,t}$ denotes the actual Scope 3 emissions, $\hat{y}_{i,t}$ the predicted values, N the number of companies, and T the number of years and $N \times T$ the number of company-year pairs.
The RMSE is particularly sensitive to large errors due to the squaring of residuals. This characteristic is advantageous when penalizing significant prediction errors is important. Lower values indicate better predictive accuracy.

5.4.2 R-squared

We next consider the R-squared metric, which measures the proportion of variance in the target variable explained by the model :

$$R^2 = 1 - \frac{\sum_{i=1}^{N} \sum_{t=1}^{T} (y_{i,t} - \hat{y}_{i,t})^2}{\sum_{i=1}^{N} \sum_{t=1}^{T} (y_{i,t} - \bar{y})^2}$$

where $y_{i,t}$ is the actual Scope 3 emission, $\hat{y}_{i,t}$ the predicted value, $\bar{y}$ the overall mean across all companies and years, N the number of companies, and T the number of years.
R^2 ranges from 0 to 1, with values closer to 1 indicating a better model fit.

[23] Assael et al. (2022), Heurtebize et al. (2022) and Nguyen et al. (2022)

5.4.3 Mean Absolute Error

We also consider the MAE, which measures the average absolute difference between actual and predicted values :

$$\text{MAE} = \frac{1}{N \times T} \sum_{i=1}^{N} \sum_{t=1}^{T} |y_{i,t} - \hat{y}_{i,t}|$$

where $y_{i,t}$ is the actual Scope 3 emissions, $\hat{y}_{i,t}$ the predicted Scope 3 value, $|y_{i,t} - \hat{y}_{i,t}|$ the absolute error between actual and predicted value, N the number of companies, T the number of years, and $N \times T$ the number of company-year pairs.

MAE is less sensitive to outliers than RMSE since it does not square errors, providing a complementary perspective on overall predictive accuracy. Lower MAE values indicate better model performance.

5.4.4 Good-Fit

While previous metrics assess overall predictive accuracy, they do not explicitly measure generalization capability. To address this, we introduce the Good-Fit metric, which compares in-sample and out-of-sample performance :

$$\text{Good-Fit} = \text{RMSE out-of-sample } - \text{RMSE in-sample}$$

This metric quantifies the generalization gap between training and test performance. A value close to zero indicates that the model performs consistently on both training and test data, demonstrating robustness and good generalizability. Negative values suggest underfitting, meaning the model fails to capture the complexity of the data, whereas positive values indicate overfitting, where the model performs well on training data but poorly on unseen data.

Using multiple evaluation metrics provides a more nuanced and comprehensive understanding of model performance. Each metric captures different aspects, such as sensitivity to outliers, variance explained, and generalization ability, ensuring a balanced assessment and supporting the selection of the most effective model for estimating and predicting Scope 3 emissions.

6 Empirical applications

Having introduced the empirical strategy and the set of models used for prediction, we now turn to the empirical assessment of their performance. This section presents the different applications of the methodology to our Scope 3 dataset, including the LASSO-based variable selection, the estimation of observed emissions, and the prediction of missing values. The objective is to evaluate how each modeling strategy performs in practice and to highlight the relative contribution of the proposed approach to improving the measurement of value-chain emissions.

6.1 Predictor selection

We begin by splitting the data into training and test samples. The training sample is used both for LASSO-based variable selection and for estimating all models, while the test sample is reserved exclusively for evaluating out-of-sample performance. This separation prevents data leakage and ensures an unbiased assessment of predictive accuracy. Both samples are standardized using statistics computed on the training data only. This avoids leakage, guarantees consistency between the two samples, and simulates realistic prediction conditions, ensuring that performance metrics accurately reflect the models' ability to generalize to unseen data.

Then, LASSO is applied to the full set of company-level predictors as well as to sectoral, upstream and downstream sectoral emissions. This approach is well suited to our context, where Scope 3 emissions depend on a large number of potentially collinear drivers. By penalizing irrelevant coefficients, LASSO isolates the most informative predictors and reduces dimensionality. The selected variables are reported in Table 1.

Table 1: LASSO variables' selection results

Company Specific Predictors	**Upstream Emissions Predictors**	**Downstream Emissions Predictors**
EBITDA	Upstream9	Downstream5
Property Plant and Equipement	Upstream12	Downstream 12
R&D Expenditures	Upstream17	Downstream20
Employees	Upstream28	Downstream25
Net Sales	Upstream29	Downstream26
Total Assets	Upstream30	Downstream40
	Upstream34	Downstream41

Regarding company-specific predictors, the LASSO systematically selects variables capturing both operational scale and financial structure. These choices are consistent with the economic interpretation of Scope 3 emissions: companies with higher sales, larger workforces, or more complex asset structures typically operate broader and more emission-intensive value chains. The selected predictors also align with findings in the literature (e.g., Serafeim and Caicedo (2022), Nguyen et al. (2022)), which identify similar compagny characteristics as key determinants of Scope 3 emissions. By penalizing less informative variables, the LASSO isolates those with the strongest explanatory power, confirming the relevance of these operational and financial drivers. Standardizing all variables ensures that differences in scale do not distort the selection process or the interpretation of the coefficients. Overall, the LASSO, combined with cross-validation, provides a robust and economically grounded approach to selecting the most relevant predictors of Scope 3 emissions.

LASSO's selection of sectoral emission variables reveals an important insight: it does not retain the sectoral emissions of the firm's own sector. This challenges the initial assumption that a company's Scope 3 emissions should be primarily driven by the aggregate emissions of its sector. While this may seem counterintuitive, given that companies within the same sector often rely on similar suppliers, technologies, and production processes, the result highlights the strong idiosyncratic component of Scope 3 emissions. In practice, the structure of value chains varies substantially across firms, even within a single sector. Companies differ in their supplier networks, logistics choices, product mixes, and downstream markets. These company-specific linkages create heterogeneous emission profiles that cannot be fully captured by sector averages. Moreover, cross-sector interactions such as relying on suppliers from energy- or metal-intensive industries may play a larger role than intra-sector dynamics. Overall, the fact that LASSO penalizes and excludes own-sector emissions indicates that Scope 3 emissions are shaped less by broad sectoral characteristics and more by the micro-structure of each firm's value chain, underscoring the importance of company-level heterogeneity in modeling Scope 3 emissions.

Regarding upstream and downstream emissions, LASSO selects variables all along the value chain. This indicates that these emissions are key drivers of Scope 3 emissions, consistent with their classification as indirect emissions outside the company's immediate control. Accurately estimating Scope 3 emissions therefore requires considering both upstream and downstream sectors, which validates our assumptions 2 and 3. Looking in detail, the selected upstream variables range from "Upstream 9" which represents the emissions of the 9^{th} closest sector to the company's sector in its upstream value chain to "Upstream34" which represents the emissions of the 34^{th} closest sector to the company's sector in its upstream value chain. While downstream variables range from "Downstream5" which represents the emissions of the 5^{th} closest sector to the company's sector in its downstream value chain, and the last is "Downstream41" which represents the emissions of the 41^{st} closest sector to the company's sector in its downstream value chain. This pattern implies that Scope 3 emissions are influenced not only by nearby sectors in the value chain but also by more distant ones, challenging the intuition that proximate sectors matter more. While it might seem intuitive that sectors closest to a company in the value chain would have the greatest influence on its Scope 3 emissions, our results show that this is not necessarily the case. Shared production processes and integrated supply chains do facilitate direct transmission from nearby sectors, but more distant sectors, less directly linked, can also exert significant indirect effects. This highlights the structural complexity of Scope 3 emissions: they reflect the cumulative interactions along the entire value chain. Even distant upstream and downstream sectors contribute meaningfully to a firm's indirect emissions, emphasizing that the carbon footprint extends beyond immediately adjacent sectors.

This first step in our empirical analysis not only informs the specification of our models but also delivers several robust insights.

First, company-specific variables emerge as key determinants of Scope 3 emissions. Operational and financial characteristics such as size, workforce, sales, and asset structure reflect the scale and complexity of a company's activities and its value chain. Even though Scope 3 emissions are largely indirect, they are substantially shaped by these internal factors, highlighting the importance of company-level heterogeneity in understanding and predicting these emissions.

Second, upstream and downstream emissions along the entire value chain constitute critical drivers of Scope 3 emissions. This confirms that Scope 3 emissions are not confined to immediately adjacent sectors but propagate throughout the broader network of suppliers and customers. Ignoring the full value chain empirically would bias the estimation of Scope 3 emissions, as would focusing exclusively on proximate sectors, which neglects significant contributions from more distant links in the production and distribution network. Since each company's value chain is unique, shaped by organizational choices, business relationships, and sectoral positioning, a comprehensive approach is necessary to capture these structural differences accurately.

Third, from a methodological perspective, our results underscore the importance of careful predictor selection when estimating Scope 3 emissions. In particular, LASSO proves effective, as it isolates variables that are consistent with the existing literature and capture the most informative drivers. Importantly, our findings highlight that excluding parts of the value chain or relying solely on sector-level aggregates can bias Scope 3 estimates, as the full network of upstream suppliers and downstream customers contributes meaningfully to a firm's indirect emissions. To further improve estimation accuracy, more granular firm-to-firm data along the value chain would be necessary, as sectoral averages often fail to reflect the complexity of company-level operations. These results emphasize the need for systematic emissions reporting and reliable, up-to-date data.

Fourth, from a climate policy perspective, our results indicate that Scope 3 emissions are strongly driven by sectoral emissions across both upstream and downstream parts of the value chain. Effective transition policies toward a net-zero economy must therefore account for the full network of economic interactions and cannot focus solely on the most polluting sectors. Since Scope 3 emissions represent the largest share of a company's total emissions, reduction strategies must target the entire economy, reflecting the interconnected nature of value chains. Limiting policies to selected sectors risks overlooking significant contributions from other parts of the network and may undermine overall decarbonization efforts. This underscores that successful emission reduction requires coordinated action across all sectors to address the systemic nature of Scope 3 emissions.

Overall, our findings highlight that Scope 3 emissions are shaped both by company-specific characteristics and by the full upstream and downstream value chain. Accurate estimation and effective policy design therefore require accounting for micro-level heterogeneity as well as the systemic, economy-wide propagation of indirect emissions.

6.2 Estimation of reported Scope 3 emissions

We extend our analysis by estimating reported Scope 3 emissions using the set of predictors selected through LASSO. To do so, we rely on both the parametric regression models and the machine learning algorithms introduced in the previous section. Our objective is to identify the most effective approach for estimating and then predicting Scope 3 emissions, thereby establishing a clear and comparable benchmark. Model performance is evaluated on the test sample using the metrics presented in Section 5, and the results are reported in Table 2.

Table 2: Evaluation Metric Results for all models

	OLS	**Panel FE**	**RF**	**XGBoost**	**LightBoost**	**CatBoost**
RMSE	0.42	1.79	0.27	0.23	0.42	0.30
R-squared	0.83	0.01	0.96	0.98	0.86	0.95
MAE	0.25	1.48	0.12	0.09	0.15	0.12
Good-Fit	-0.05	0.12	0.03	0.21	0.01	0.19

Our results indicate that machine learning algorithms generally outperform parametric models in terms of evaluation metrics, with Light Gradient Boosting being the only exception, performing comparably to parametric approaches.

First, it is important to note that the Panel Fixed Effects (Panel FE) model performs poorly in estimating Scope 3 emissions, which is somewhat surprising given the panel structure of the dataset. A likely explanation is the low temporal variability of Scope 3 emissions, which hampers the identification of significant relationships in a fixed-effects framework, as this approach relies only on within-unit (i.e., over time) variation. Moreover, the assumption that unobserved effects remain constant over time may not hold in this context, particularly if company-specific shocks or structural changes occur from year to year. This further limits the model's ability to capture relevant dynamics. Overall, these findings indicate that panel models, at least in their standard fixed-effects specification, may be poorly suited for estimating Scope 3 emissions.

A closer look at the results shows that the Ordinary Least Squares (OLS) model is among the least effective approaches for estimating Scope 3 emissions. Despite its simplicity and strong interpretability, it exhibits relatively high RMSE and MAE, a low R-squared, and clear signs of underfitting compared with machine learning algorithms, indicating limited ability to capture the complexity of Scope 3 emission. Somewhat unexpectedly, the Light Gradient Boosting model (LightBoost), despite its reputation for predictive performance, does not perform substantially better. Although it yields a slightly improved fit on the test sample, with marginal gains in R-squared and MAE and virtually no overfitting, its overall performance remains very close to that of the Ordinary Least Squares, particularly given their similar RMSE values. The convergence of these two fundamentally different approaches suggests that neither simplicity nor model complexity guarantees strong predictive accuracy. This highlights the importance of systematically comparing multiple techniques. Consequently, even if a well-specified parametric model supported by LASSO can yield acceptable performance, both Ordinary Least Squares and Light Gradient Boosting clearly lag behind the other machine learning algorithms and can be reasonably excluded from the most effective models for estimating Scope 3 emissions.

A this stage, Random Forest (RF), Extreme Gradient Boosting (XGBoost), and Categorical Boosting (Catboost) clearly outperform the other models. This confirms that Scope 3 emissions exhibit complex relationships with their predictors, making machine learning methods particularly well suited for this task. To refine the comparison, we focus on these three models. Random Forest performs better than Categorical Boosting for a similar MAE value, justifying its selection over the latter. Extreme Gradient Boosting also outperforms Categorical Boosting across nearly all evaluation metrics (including RMSE, R-squared, and MAE) except for a slightly higher Good-Fit value. This marginal increase in overfitting (0.02) is negligible in practice and does not meaningfully impair generalization. Given its stronger overall predictive accuracy and robustness, Extreme Gradient Boosting is therefore preferred. Taken together, these results indicate that Random Forest and Extreme Gradient Boosting deliver the most reliable performance, allowing us to exclude Categorical Boosting from further consideration.

Finally, we are left with two models: Random Forest and Extreme Gradient Boosting. Both perform similarly on RMSE, R-squared, and MAE, though Extreme Gradient Boosting is slightly better overall. The key difference lies in the Good-Fit measure: Random Forest shows a value near zero, indicating minimal gap between training and test RMSE and strong generalization. In contrast, Extreme Gradient Boosting exhibits a higher Good-Fit value, reflecting overfitting. While its test RMSE remains low, the gap with training RMSE suggests the model captures idiosyncrasies in the training data, adding unnecessary complexity and potentially reducing reliability on new observations. Given their comparable predictive accuracy, robustness becomes decisive. Random Forest, combining strong performance with stable generalization, is therefore the preferred model for our analysis.

6.3 Prediction of unreported Scope 3 emissions

In the final step of our empirical application, we develop a methodology to predict unreported Scope 3 emissions. We begin by assessing prediction accuracy at the original scale, as standardizing the target would undermine interpretability and limit the practical relevance of a benchmark method designed to address missing data and methodological gaps. We choose the Weighted Average Relative Error (WARE), obtaining a value of approximately 1.83%, meaning that predicted values deviate on average by only 1.83% from true emissions. This metric is particularly suitable given the large variability of Scope 3 emissions: it scales the error to the magnitude of each observation, reduces the influence of extreme values, enables consistent comparisons across heterogeneous firms, and ensures balanced accuracy for both large and small companies. Given this variability, such a low relative error indicates strong robustness and supports the reliability of our model for predicting missing Scope 3 emissions, making it a credible benchmark.

Then, we construct the prediction sample for unreported Scope 3 emissions, which requires predicting a large number of missing values from a comparatively smaller training set. To ensure the relevance and reliability of these predictions, we implement a sector similarity analysis. For each sector, we compute the percentage of missing and observed data among the firms belonging to that sector and compare the sectoral coverage between the training sample and the firms with missing values. This allows us to identify sectors where observed data exceed, or are only marginally lower than, missing data, thus providing a sufficiently representative basis for prediction. Restricting predictions to these sectors ensures that firms are compared within comparable environments and reduces the risk of errors arising from imbalanced data coverage. The results of this analysis are presented in Table 3.

Table 3: Sector Similarity Analysis Results

Sector	**NA**	**Observed**
Accommodation and food service activities	**1.35**	**1.13**
Administrative and support services	**3.67**	**6.98**
Arts entertainment and recreation	0.61	0.23
Chemical and chemical products	**1.49**	**1.80**
Electrical equipment	**0.86**	**1.80**
Electricity gas steam and air conditioning supply	**0.68**	**3.60**
Financial and insurance activities	**17.70**	**16.22**
Human health and social work activities	0.61	0.23
It and other information services	8.27	3.38
Machinery and equipment nec	3.13	0.23
Manufacturing nec repair and installation of machinery and equipment	**1.52**	**1.58**
Mining and quarrying	**0.26**	**1.35**
Motor vehicles trailers and semitrailers	**0.65**	**1.35**
Pharmaceuticals medicinal chemical and botanical products	1.21	0.45
Professional scientific and technical activities	30.24	41.0
Publishing audiovisual and broadcasting activities	**7.24**	**4.73**
Real estate activities	**6.35**	**6.98**
Rubber and plastics products	**0.77**	**0.68**
Telecommunications	**0.63**	**2.03**
Warehousing and support activities for transportation	**0.63**	**1.8**

Water supply sewerage waste management and remediation activities	1.03	0.23
Wholesale and retail trade repair of motor vehicles	11.11	2.25

To select sectors for the Scope 3 emission prediction sample, we implement a multi-step procedure to ensure reliable estimates. We retain sectors where observed data (Observed) substantially exceed missing values (NA) or where the gap is limited, providing a solid foundation for accurate predictions. Conversely, sectors where missing values outweigh observed data or where overall data scarcity remains high are excluded, as they could compromise estimate robustness. This approach balances data sufficiency and reliability. Applying this procedure, we identify 14 sectors[24], significantly reducing the number of missing values to be predicted. The filtered sample aligns with the training data, ensuring representative and robust predictions. We then apply the trained Random Forest model and destandardize the results using the training sample's statistics, producing a comprehensive database of Scope 3 emissions at the original scale.

Through this methodology, we show that Scope 3 emissions can be accurately estimated and predicted, even when data availability is limited. To ensure robust results, we excluded outliers, which reduced the sample size but enhanced both statistical reliability and economic consistency. Data scarcity and uneven quality are challenges for all economic agents: companies, which must integrate these emissions into their transition strategies and extra-financial reporting; banks and financial institutions, which assess climate-related portfolio risks; and governments and regulators, who rely on such data to design effective climate policies. The proposed methodology offers several key advantages. It rests on rigorous econometric foundations and combines company-level financial data with sectoral emissions indicators, enabling reliable estimates of Scope 3 emissions. It also provides a reproducible framework applicable at both company and institutional levels. Beyond prediction, this approach helps fill information gaps by delivering credible estimates where data is missing, improving carbon footprint accuracy and supporting companies in setting scientifically grounded decarbonisation targets. Designed to be methodologically sound, rigorously documented, and easily adoptable, it addresses current limitations in reporting practices, standards, and data availability, and could inform the development of low-carbon transition standards. Nonetheless, performance could be further improved by expanding the initial database, which is constrained by limited and uneven reporting. Many companies are not yet required to report Scope 3 emissions comprehensively, so these data remain widely underreported. Enhancing coverage will require greater corporate transparency, the development of internal benchmarking methods, and closer dialogue with regulators to standardize measurement and reporting practices. In the context of the climate emergency, all economic agents must actively commit to controlling and reducing these diffuse but decisive

[24]Sectors with highlighted coverage in Table 3.

emissions through targeted policies grounded in accurate and reliable data.

7 Conclusion

Given the climate emergency, understanding, quantifying, and ultimately reducing GHG emissions has become a central economic and political priority. While Scope 1 and 2 emissions are increasingly well regulated, Scope 3 emissions often representing the largest share of a company's carbon footprint, remain widely underestimated due to methodological complexity and persistent data scarcity. Our study contributes to closing this gap by proposing an empirical framework capable of estimating Scope 3 emissions at the company level, even when information is incomplete or heterogeneous. By combining company-level operational and financial attributes with sectoral upstream and downstream emissions derived from input–output tables, we capture both the internal determinants of emissions and the broader value-chain interdependencies that shape each firm's indirect climate impact.

Our findings show that Scope 3 emissions, though indirect, are strongly influenced by company-specific characteristics. Operational intensity, financial structure, and activity scale all contribute to shaping the magnitude of value-chain emissions. These results highlight that a company's exposure to value-chain emissions is not only determined by its suppliers and clients, but is also shaped by its own technological, organizational, and strategic choices. At the same time, sectoral upstream and downstream emissions emerge as crucial drivers, illustrating the importance of production-network effects. This underscores a key economic insight: the transition to a low-carbon economy cannot be achieved only by targeting a handful of high-emitting sectors; it requires coordinated and economy-wide policies that address spillovers across the entire productive system.

Methodologically, our comparisons reveal clear limitations in traditional econometric models. Panel Fixed Effects regressions suffer from low within-firm variability, while Ordinary Least Squares cannot capture the complex relationship inherent to value-chain structures. Machine-learning models provide substantial improvements, though their performance varies: Light Gradient Boosting offers only marginal gains over Ordinary Least Squares, whereas Random Forest, Extreme Gradient Boosting, and Categorical Boosting deliver significantly better predictive accuracy. Among these, Random Forest achieves the most balanced trade-off between performance and generalisation, making it particularly suitable for prediction in sparse and heterogeneous datasets.

Despite the limited availability of reported Scope 3 emissions, our study demonstrates that rigorous prediction remains feasible. Through a sectoral similarity analysis and the exclusion of excessively incomplete observations, we construct a consistent prediction sample that allows the Random Forest model to generate a coherent company-level database of estimated Scope 3 emissions. By filling critical data gaps, this approach strengthens the informational foundation required by companies, financial institutions, and regulators to assess climate risks, set credible decarbonisation targets, and integrate environmental considerations into economic decisions.

Ultimately, the future improvement of such methodologies depends on deeper institutional progress, greater corporate transparency, richer firm-to-firm data on supply-chain relationships, and harmonised reporting standards. From a methodological perspective, our variable-selection procedure proves effective in identifying the key drivers of Scope 3 emissions, yet its performance could be improved with greater visibility on inter-firm connections. More granular data on these relationships would substantially enhance predictive accuracy and allow models to capture the full complexity of production networks. These elements are essential for ensuring that estimation frameworks reflect real economic interdependencies and can meaningfully support the design of effective, economy-wide climate policies. In a world shaped by dense global value chains, closing the information gap surrounding Scope 3 emissions is not only a technical challenge: it is a prerequisite for steering the transition toward a credible, efficient, and fair low-carbon economy.

Bibliography

J. Assael, T. H. andLaurent Carlier, and F. Soupé. Greenhouse gases emissions: estimating corporate non-reportedemissions using interpretable machine learning. *arXiv*, 2022.

C. Bentéjac, A. Csörgő, and G. Martínez-Muñoz. A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review*, 54:1937–1967, 2021.

L. Breiman. Random Forests. *British Journal of Surgery*, 45:5–32, 2001.

CDP. CDP Full GHG Emissions Dataset Technical Annex III: Statistical Framework. 2020.

R. Chumacero. Ordinary Least Squares: A Review. *Working Paper*, 2012.

R. H. Crawford, P.-A. Bontinck, A. Stephan, T. Wiedmann, and M. Yu. Hybrid life cycle inventory methods - A review. *Journal of Cleaner Production*, 172:1273–1288, 2017.

J. Downie and W. Stubbs. Evaluation of Australian companies'scope 3 greenhouse gasemissions assessments. *Journal of Cleaner Production*, 56:156–163, 2011.

A. Godin and P. Hadji-Lazaro. Identification des vunérabilités à la transition induites par la demande : application d'une systémique à l'Afrique du Sud. *Presses de Sciences Po, "Revue économique"*, 73:267–301, 2022.

B. Goldhammer, C. Busse, and T. Busch. Estimating corporate carbon footprints with externally available data. *Journal of Industrial Ecology*, 21:1165–1179, 2017.

Y. Han, A. Gopal, L. Ouyang, and A. Key. Estimation of Corporate Greenhouse Gas Emissions via Machine Learning. *arXiv:2109.04318*, 2021.

E. G. Hertwich and R. Wood. The growing importance of scope 3 greenhouse gas emissions from industry. *Environmental Research Letters*, 13:104013, 2018.

T. Heurtebize, F. Chen, F. Soupé, and R. L. de Carvalho. Corporate Carbon Footprint: A Machine Learning Predictive Model for Unreported Data. *The Journal of Impact and ESG Investing*, 3 (2):36–54, 2022.

Y. A. Huang, C. L. Weber, and H. S. Matthews. Categorization of Scope 3 Emissions for Streamlined Enterprise Carbon Footprinting. *Environmental Science Technology*, 43 (22):8509, 2009.

S. Islam, S. Ponnambalam, and H. L. Lam. Review on life cycle inventory: methods, examples and applications. *Journal of Cleaner Production*, 136:266–272, 2016.

H. Kelejian and G. Piras. Chapter 15 - Panel Data Models. *Spatial Econometrics*, pages 307–341, 2017.

B. Kerimov and R. Chernyshev. Review of machine learning methods in the estimation of greenhouse gas emissions. *International Conference of Young Scientists "Modern Problems of Earth Sciences. Proceedings, ISBN 978-9941-36-044-2, Publish House of Iv. Javakhishvili Tbilisi State University, Tbilisi, November 21-22*, pages 107–110, 2022.

N. Ma, W. Y. Shum, T. Han, and F. Lai. Can Machine Learning be applied to Carbon Emissions Analysis: An Application to the CO2 Emissions Analysis Using Gaussian Process Regression. *Frontiers in Energy Research*, 9:756311, 2021.

Q. Nguyen, I. Diaz-Rainey, and D. Kuruppuarachchi. Predicting corporate carbon footprints for climate finance risk analyses: a machine learning approach. *Energy Economics*, 95: 105129, 2021.

Q. Nguyen, I. Diaz-Rainey, A. Kitto, B. I. McNeil, N. A. Pittman, and R. Zhang. Scope3 Emissions : Data Quality and Machine Learning Prediction Accuracy. *PLOS Climate*, 2 (11), 2022.

B. E. Quants. Distributional greenhouse gas emissions estimates: data challenges and modeling solutions. 2022.

J. Ranganathan, L. Corbier, P. Bhatia, S. Schmitz, P. Gage, and K. Oren. The greenhouse gas protocol: A corporate accounting and reporting standard (revised edition), Washington, DC: World Resources Institute and World Business Council for Sustainable Development. 2004.

J. Ranstam and J. Cook. LASSO Regression. *British Journal of Surgery*, 105:1348, 2018.

H. V. Rowley, S. Lundie, and G. M. Peters. A hybrid life cycle assessment model for comparison with conventional methodologies in Australia. *The International Journal of Life Cycle Assessment*, 14:508–516, 2009.

G. Serafeim and G. V. Caicedo. Machine Learning Models for Prediction of Scope 3 Carbon Emissions. *Harvard Business School Accounting Management Unit Working Paper*, 22-080, 2022.

G. Shrimali. Scope 3 Emissions: Measurement and Management. *The Journal of Impact and ESG Investing*, 2021.

M. Sparrevik and S. Utstøl. Assessing life cycle greenhouse gas emissions in the Norwegian-defence sector for climate change mitigation. *Journal of Cleaner Production*, 248:119196, 2020.

S. Suh and G. Huppes. Methods for Life Cycle Inventory of a product. *Journal of Cleaner Production*, 13:687–697, 2005.

G. P. Team. Corporate value chain (scope 3) accounting and reporting standard. 2011.

R. Tibshirani. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B*, 58:267–288, 1996.

T. Wiedmann and J. Minx. A definition of ‘carbon footprint. *Ecological economics researchtrends*, 1:1–11, 2008.

A Appendix

A.1 Sectoral emissions hierarchy example

To provide a clearer illustration of *Assumption 4*, let us focus on a specific company i operating in the sector "Wood and products of wood and cork" for the year 2018. Our goal is to identify the most relevant sectors both upstream and downstream in the company's value chain.

Upstream sectors are determined using the Leontief inverse matrix, which quantifies the total output from each sector required to satisfy a unit of final demand in the company's sector. In other words, it tells us which sectors are the main suppliers of inputs to the company and therefore contribute most to its upstream Scope 3 emissions. Downstream sectors are identified using the Ghosh inverse matrix, which measures how the outputs of the company's sector are distributed and absorbed by other sectors in the economy. This highlights the sectors that consume the company's products and thus receive the indirect emissions through downstream activities.

Without applying our additional proximity-based ranking (i.e., considering only the raw Input-Output structure), the top three sectors are identical for both upstream and downstream, not only for this specific company but also for all companies in the database. This is because we are using the universal ranking derived from the OECD (2021) Input-Output Tables, which captures the general economic structure and the largest inter-sectoral flows at the national level. Table A.1 summarizes the three most relevant upstream and downstream sectors for the example company. By examining this table, one can see which sectors contribute most to the company's indirect emissions according to the raw Input-Output structure, before incorporating sectoral proximity.[25]

Table A.1: An exemple of sectors' hierarchy upstream and downstream without ranking

	Upstream
First	"Agriculture hunting forestry"
Second	"Fishing and aquaculture"
Third	"Mining and quarrying"
	Downstream
First	"Agriculture hunting forestry"
Second	"Fishing and aquaculture"
Third	"Mining and quarrying"

Note: For the sector "Wood and products of wood and cork", this table presents the top three sectors upstream (suppliers) and downstream (customers) according to the raw Input-Output structure from the OECD (2021) Input-Output Tables. No proximity ranking has been applied.

To introduce heterogeneity in the value chain and account for differences between companies,

[25] A detailed description of the data used to generate the Leontief and Ghosh inverse matrices is provided in the data section of the paper.

we extend the analysis beyond the raw Input-Output structure by applying a proximity-based ranking of sectors. For a given company i operating in sector j, the objective is to identify which upstream and downstream sectors are the most influential in determining its Scope 3 emissions. We start from the Leontief and Ghosh inverse matrices.

The Leontief inverse matrix indicates the total output from each sector required to satisfy one unit of final demand in sector j, so sectors with larger coefficients are the most important suppliers and contribute the most to upstream emissions. Conversely, the Ghosh inverse matrix shows how outputs from sector j are distributed and absorbed by other sectors, meaning that sectors with higher coefficients are the main recipients of the company's outputs and therefore most relevant for downstream emissions.

For each sector j, we rank these coefficients from largest to smallest to identify the sectors most closely connected economically and environmentally to sector j. This ranking captures the intensity of inter-sectoral transactions, which directly translates into potential contributions to Scope 3 emissions. By selecting the top three sectors upstream and the top three sectors downstream, we focus on those that are most relevant to the company's value chain, either as critical input suppliers or major output consumers.

Applying this methodology to the year 2018, Table A.2 presents the three most relevant upstream and downstream sectors for the example company in the "Wood and products of wood and cork" sector. This ranking-based approach allows us to construct a value chain that reflects the differential influence of sectors, providing a more nuanced and realistic view of indirect emissions compared to the uniform, universal Input-Output ranking.

Table A.2: An exemple of sectors' hierarchy upstream and downstream with ranking

	Upstream
First	"Agriculture hunting forestry"
Second	"Wholesale and retail trade repair of motor vehicles"
Third	"Professional scientific and technical activities"
	Downstream
First	"Construction"
Second	"Manufacturing nec repair and installation of machinery and equipment"
Third	"Wholesale and retail trade repair of motor vehicles"

Note:For the sector "Wood and products of wood and cork", the top three upstream sectors are those supplying the most inputs to the company, and the top three downstream sectors are the main recipients of the company's outputs. The ranking is based on the Leontief inverse matrix (upstream) and Ghosh inverse matrix (downstream) for the year 2018.

It is important to understand how the hierarchy of sectors translates into the variables used in our model. For the company i in the sector "Wood and products of wood and cork," the three upstream sectors identified as the closest in the value chain from its sector are "Agriculture, hunting, forestry," "Wholesale and retail trade; repair of motor vehicles," and "Professional, scientific, and technical activities." In practice, this means that the variables

$\bar{y}^{(1)}_{j,t}$, $\bar{y}^{(2)}_{j,t}$, and $\bar{y}^{(3)}_{j,t}$ in our specification represent the total emissions of these three sectors, respectively, with $\bar{y}^{(1)}_{j,t}$ corresponding to the sector most closely connected to company sector j, $\bar{y}^{(2)}_{j,t}$ to the second most connected, and $\bar{y}^{(3)}_{j,t}$ to the third. This ordering captures the intuition that sectors more directly connected in the production chain have a larger influence on the company's Scope 3 emissions.

The same logic applies downstream. The three sectors most closely linked to company sector j are "Construction," "Manufacturing, repair, and installation of machinery and equipment," and "Wholesale and retail trade; repair of motor vehicles." Accordingly, $\bar{z}^{(1)}_{j,t}$, $\bar{z}^{(2)}_{j,t}$, and $\bar{z}^{(3)}_{j,t}$ represent the emissions of these downstream sectors in order of closeness. By structuring the variables in this way, the model explicitly accounts for the fact that both upstream suppliers and downstream consumers of the company's sector contribute differently to its overall Scope 3 emissions. This hierarchical ordering thus allows us to weight the influence of sectors according to their proximity in the value chain, providing a more precise and realistic measure of the company's indirect environmental impact.

A.2 Sector details of Input-Output Tables, OECD (2021)

Table A.3

Industry Code	ISIC 4 corresponding Division	Description
D01T02	01, 02	Agriculture, hunting, forestry
D03	3	Fishing and aquaculture
D05T06	05, 06	Mining and quarrying, energy producing products
D07T08	07, 08	Mining and quarrying, non-energy producing products
D09	9	Mining support service activities
D10T12	10, 11, 12	Food products, beverages and tobacco
D13T15	13, 14, 15	Textiles, textile products, leather and footwear
D16	16	Wood and products of wood and cork
D17T18	17, 18	Paper products and printing
D19	19	Coke and refined petroleum product
D20	20	Chemical and chemical products
D21	21	Pharmaceuticals, medicinal chemical and botanical products
D22	22	Rubber and plastics products
D23	23	Other non-metallic mineral products
D24	24	Basic metals
D25	25	Fabricated metal products
D26	26	Computer, electronic and optical equipment
D27	27	Electrical equipment
D28	28	Machinery and equipment, nec
D29	29	Motor vehicles, trailers and semi-trailers
D30	30	Other transport equipment

D31T33	31, 32, 33	Manufacturing nec; repair and installation of machinery and equipment
D35	35	Electricity, gas, steam and air conditioning supply
D36T39	36, 37, 38, 39	Water supply; sewerage, waste management and remediation activities
D41T43	41, 42, 43	Construction
D45T47	45, 46, 47	Wholesale and retail trade; repair of motor vehicles
D49	49	Land transport and transport via pipelines
D50	50	Water transport
D51	51	Air transport
D52	52	Warehousing and support activities for transportation
D53	53	Postal and courier activities
D55T56	55, 56	Accommodation and food service activities
D58T60	58, 59, 60	Publishing, audiovisual and broadcasting activities
D61	61	Telecommunications
D62T63	62, 63	IT and other information services
D64T66	64, 65, 66	Financial and insurance activities
D68	68	Real estate activities
D69T75	69 to 75	Professional, scientific and technical activities
D77T82	77 to 82	Administrative and support services
D84	84	Public administration and defence; compulsory social security
D85	85	Education
D86T88	86, 87, 88	Human health and social work activities
D90T93	90, 91, 92, 93	Arts, entertainment and recreation
D94T96	94,95, 96	Other service activities
D97T98	97, 98	Activities of households as employers; undifferentiated goods- and services- producing activities of households for own use

Source: OECD.Stat

A.3 Structure of companies sample matched with OECD sectors details

Table A.4

Sector	company	Number
Agriculture, hunting, forestry		0
Fishing and aquaculture		0
Mining and quarrying	La Française de l'Energie SA, Imerys SA⋆	2
Food products, beverages and tobacco	Fleury Michon SA⋆, Paulic Meunerie SA, Malteries Franco Belges SA, Saint Jean Groupe, Lombard et Medot SA	5

Textiles, textile products, leather and footwear	Devernois SA, Barbara Bui SA	2
Wood and products of wood and cork	Cogra Quarante Huit SA, Ober SA, Tonnellerie Francois Freres SA	3
Paper products and printing	Imprimerie Chirat SA	1
Coke and refined petroleum product	Esso Societe Anonyme Française SA*	1
Chemical and chemical products	Arkema SA*, Baikowski SA, Encres Dubuit SA, Biomerieux SA*, Interparfums SA*, Orapi SA, Pcas SA*, Robertet SA*	8
Pharmaceuticals, medicinal chemical and botanical products	Biosynex SA, Theradiag SA, Boiron SA*, Guerbet SA*, Vetoquinol SA*, Virbac SA*	6
Rubber and plastics products	Mg International SA, Piscines Desjoyaux SA, Augros Cosmetic Packaging SA, Bolloré SE*	4
Other non-metallic mineral products	Hoffmann Green Cement Technologies SAS, Vicat SA*, Placoplatre SA	3
Basic metals	Poujoulat SA, Dynafond SA	2
Fabricated metal products	Signaux Girod SA, Figeac Aero SARL*, Parrot SA*, Streit Mecanique SA, Verney Carron SA	5
Computer, electronic and optical equipment	Adeunis SA, Egide SA, I2S SA, Kalray SA, Kerlink SA, Cogelec SA, Munic SA, Alpha MOS SA, Enensys Technologies SA, NSE SA, TronicS Microsystems SA, Evolis SA, Ekinops SA*, Gea Grenobloise Electronique Automatisme SA, Lumibird SA*, Memscap SA, Paragon Id SA, Soitec SA*, Made SA Novatech Industries SA	20
Electrical equipment	Prismaflex International SA, Lucibel SA, Forsee Power SA, Legrand SA*, Precia SA*	5
Machinery and equipment, nec	Enertime SAS, Mgi Digital Technology SA, Riber SA*, Bio-UV Group SA, Txcom SA, UV Germi SA, Vergnet Vsa SA, Balyo SA*, Lectra SA*, Manitou BF SA*, Orege SA, Haulotte Group SA*, Rousselet Centrifugation SA, Hydraulique PB SA, Metalliance SA	15
Motor vehicles, trailers and semi-trailers	Akwel SA*, Navya SA, Plastiques du Val de Loire SA, Renault SA*	4
Other transport equipment	Drone Volt SA, Fountaine Pajot SA, Dassault Aviation SA*, Latecoere SA*	4
Manufacturing nec; repair and installation of machinery and equipment	Hydrogen-Refueling-Solutions SA, ICeram SA, Theraclion SA, Trilogiq SA, St Dupont SA*, Essilorluxottica SA*, France Tourisme Immobilier SA, Edap Tms SA	8
Electricity, gas, steam and air conditioning supply	Mint SA, Electricite de France SA*, Engie SA*, Hydrogene De France SA, Hydro Exploitations SA	5

Water supply; sewerage, waste management and remediation activities	Derichebourg SA , Groupe Pizzorno Environnement SA, Recylex SA⋆, Revival Expansion SA, Compagnie des Eaux de Royan SA	5
Construction	AST Groupe SA⋆, Capelli SA⋆, Hexaom SA⋆, Realites SA, Uniti SA, Altareit SCA , Colas SA , Les Constructeurs du Bois SA	8
Wholesale and retail trade; repair of motor vehicles	Agrogeneration SA, AURES Technologies SA⋆, Advini SA, Cibox Inter@ctive SA, Cybergun SA, Ecomiam SA, EO2 SA, Gold by Gold SA, Implanet SA, Immersion SA, Innelec Multimedia SA, Archos SA⋆, Groupe LDLC SA⋆, Lexibook Linguistic Electronic System SA, Largo SA, Logic Instrument SA, Installux SA, Mastrad SA, Miliboo SA, MR Bricolage SA⋆, Neolife SA, Novacyt SA, Ordissimo SA, Spineguard SA, SMAIO SA, Spartoo SA, Spineway SA, Metavisio SA, Upergy, Vialife SA, Visiomed Group SA, Vente-Unique.Com SA, We. Connect SA, Avenir Telecom SA⋆, Bigben Interactive SA⋆, Societe BIC SA⋆, La Chausseria SA, Adl Partner SA⋆, Eramet SA⋆, Fashion B Air SA, Guillemot Corporation SA, Groupe JAJ SA, Bogart SA, Marie Brizard Wine and Spirits SA⋆, Maisons du Monde SA⋆, Passat SA, Remy Cointreau SA⋆, Samse SA⋆, Ses Imagotag SA , Graines Voltz SA, Vranken Pommery Monopole SA⋆, Body One Lingerie SA, EAVS Groupe Equipements Audiovisuels et Systemes SA, Union Metallurgique de la Haute Seine SA	54
Land transport and transport via pipelines	Compagnie du Mont Blanc SA	1
Water transport		0
Air transport		0
Warehousing and support activities for transportation	Aeroports de Paris SA⋆, Societe Marseillaise du Tunnel Prado Carenage SA⋆, Les Docks des Petroles d'Ambes SA, Id Logistics SAS⋆	4
Postal and courier activities		0
Accommodation and food service activities	Catering International & Services SA⋆, Bernard Loiseau SA, Les Hotels Baverez SA, Sodexo SA⋆, Financiere Immobiliere Etang Berre Medit SA, Les Hotels de Paris SA, Soc Immobiliere et Exploit Hotel Majesti SA	7
Publishing, audiovisual and broadcasting activities	Acheter Louer Fr SA, Atari SA⋆, Sidetrade SA, Dont Nod Entertainment SA, Europacorp SA, Esker SA, Focus Entertainment SA, Intrasense SA, Wallix Group SA, Energisme SA, Erold SA, Predilife SA, Prologue SA, Streamwide SA, Veom Group SA, Wedia SA, Witbe SA, Ateme SA⋆, Believe SA, Coheris SA, Dassault Systemes⋆, ESI Group SA⋆, Infotel SA⋆, Gaumont SA⋆, Linedata Services SA⋆, Solocal Group SA⋆, Metropole Television SA⋆, Nacon SAS, Television Française 1 SA⋆, Ubisoft Entertainment SA⋆, Verimatrix SA⋆, Xilam Animation SA⋆, Groupe Plus Values SA, Mondo TV France SA, Damaris SA, Editions du Signe SA, Televista SA	37
Telecommunications	Osmozis SA, Orange SA⋆, Weaccess Group SA, Neocom Multimedia SA	4

IT and other information services	Bilendi SA, Alchimie SA, Arcure SA, Freelance.com SA, Hitechpros SA, IT Link SA, Keyrus SA*, Micropole SA, 1000mercis SA, Nam.R SA, Netgem SA*, Nextedia SA, Prodware SA, Reworld Media SA, Blockchain Group SA, Travel Technology Interactive SA, Vogo SA, Aubay SA*, Cegedim SA*, Acteos SA, Equasens SA, Societe Pour l Informatique Industrielle SA*, Itesoft SA, Alten SA*, Myhotelmatch SA, Neurones SA*, Artmarket.com SA*, Proactis SA, Sopra Steria Group SA*, Sqli SA*, Uti Group SA, Wavestone SA*, Worldline SA*, FranceSoir Groupe SA, Pacte Novation SA, IDS SA, Sequans Communications SA, Cheops Technology France SA, Criteo SA, Magillem Design Services SA, Consort NT SA	41
Financial and insurance activities	Alan Allman Associates SA, Abeo SA*, Affluent Medical SAS, Air France - KLM SA*, Actia Group SA*, Audacia SA, BD Multi Media SA, Cesar SA, Cofidur SA, Damartex SA, DBT SA, Diagnostic Medical Systems SA, Euromedis Groupe SA, Gaussin SA, Gevelot SA, Groupe Tera SA, Hipay Group SA, Idsud SA, Lanson BCC SA*, Methanor SCA, Moulinvest SA, Omer Decugis & Cie SA, Groupe Parot SA, Poulaillon SA, Rougier SA, Stradim Espace Finance SA, Altur Investissement SCA, Winfarm SA, Amplitude Surgical SA*, Amundi SA*, Antin Infrastructure Partners SAS , Soc Indust Financiere Artois SA, Aurea SA*, Axa SA*, Bassac, Bleecker SA, Burelle SA, Carrefour SA*, Casino Guichard-Perrachon SA*, Catana Group SA, Unibel SA*, Compagnie du Cambodge SA*, Groupe Crit SA*, Coface SA*, Christian Dior SE*, Electricite et Eaux De Madagascar SA*, Transition Evergreen SA, Elior group SA*, Exail Technologies *, Fayenceri Sarreguemin Digoin Vitry Franc SA, Ramsay Generale de Sante SA, High Co SA*, Cie Industrielle Financiere Entreprise SA, Fonciere Inea SA*, Kaufman & Broad SA*, Finatis SA, Fonciere Euris SA, MRM SA, Compagnie de l'Odet SE , Ovh Groupe SA , Fiducial Real Estate SA, Peugeot Invest SA, Prodways Group SA*, Rubis SCA*, Rexel SA*, Fiducial Office Solutions SA, Savencia SA*, Schneider Electric SE*, Sogeclair SA*, Scor SE*, SEB SA*, Sergeferrari Group SA*, Selectirente SA, Groupe SFPI SA, Selcodis SA, SMCP SA, Spie SA*, SRP Groupe SA*, Thermador Groupe SA*, Tikehau Capital SCA , Viel et Compagnie SA*, Valeo Sa*, Verallia SA , Courbet SA, Les Motocycles Ardoin Saint Amand et Cie SA, Batla Minerals SA, Smalto SA, Troc de l'Ile SA, Locasystem International SA, Corep Lighting SA, Foraco International SA, Groupe Carnivor SA, Constellium SE, Nicolas Miguet et Associes SA	94

Real estate activities	Adomos SA, Eurasia Groupe SA, Argan SA⋆, Atland SA, Carmila SA , Soc Centrale Bois Scieries Manche SA⋆, CFI-Compagnie Fonciere Internationale SCA, Acanthe Developpement SE⋆, Vitura SA⋆, Covivio Hotels SCA, Crosswood SA, Courtois SA, Covivio SA⋆, Eurasia Fonciere Investissements SA, Immobiliere Dassault SA, F I P P SA, Societe Fonciere Lyonnaise SA , Frey SA, Galimmo SCA, Gecina SA⋆, Icade SA⋆, Altarea SCA⋆, Klepierre SA⋆, Mercialys SA⋆, Paris Realty Fund SA, Patrimoine et Commerce SCA, Sartorius Stedim Biotech SA⋆, Societe de la Tour Eiffel SA⋆, Touax Sgtr Cite Sgt Cmte Taf Slm Touage Investissements Reunies SCA⋆, Societe de Tayninh SA, Cie de Chemins de Fer Departementaux SA, Grande Armee Investissement GAI SA, Soc Nat Proprietes d'Immeubles SCA, Maison Antoine Baud SA	34

Professional, scientific and technical activities	Abionyx Pharma SA, AB Science SA⋆, Abivax SA⋆, Accor SA⋆, Adocia SA⋆, Adux SA, 2Crsi SA, E-Pango SA, Agripower France SASU, Gascogne SA⋆, Crossject SA, Cellectis SA, Biocorp Production SA, Carbios SA, ALD SA⋆, Deinove SA, Delfingen Industry SA, Delta Plus Group SA⋆, Dolfines SA, Delta Drone SA, Advicenne SA, Eurofins-Cerep SA, Emova Group SA, Entreparticuliers.com SA, Ecoslops SA, Europlasma SA, Exacompta Clairefontaine SA, Global Bioenergies SA, Genoway SA, Groupe Guillin SA, HF Company SA, Hopscotch Groupe SA, Herige SA⋆, Hybrigenics SA, Integragen SA, Klarsen SA, Medesis Pharma SA, Amoeba SA⋆, M2i SA, Montagne et Neige Developpement SA⋆, Altamir SCA, Sirius Media, Neovacs SA, Nfty SA, Makheia Group SA, Valerio Therapeutics SA, Altheora SA, Plant Advanced Technologies PAT SA, Cerinnov Group SA, Pixium Vision SA⋆, Quantum Genomics SA, Roctool SA, Safe SA, Alstom SA⋆, Theranexus SA, U10 Corp SA, Valbiotis SA, Aramis Group SAS, Artea SA, Assystem SA⋆, Claranova SE , ABC Arbitrage SA⋆, Bonduelle SA⋆, Bouygues SA⋆, Bureau Veritas SA⋆, Cafom SA⋆, Capgemini SE⋆, Mersen SA⋆, Séché Environnement SA⋆, Compagnie des Alpes SA⋆, Beneteau SA⋆, Chargeurs SA⋆, Danone SA⋆, Dbv Technologies SA⋆, Edenred SE⋆, Electricite de Strasbourg SA⋆, Soc D Explosifs Produits Chimiques SA, Forvia SE , Phaxiam Therapeutics SA, Eutelsat Communications SA⋆, Euro Ressources SA⋆, Exel Industries SA⋆, Fermentalg SA, Financiere Marjos SA, Fnac Darty SA⋆, Eiffage SA⋆, Rallye SA⋆, CGG SA⋆, Lisi SA⋆, GL Events SA⋆, Genfit SA⋆, Gerard Perrier Industrie SA, Genomic Vision SA, Hermes International SCA⋆, IDI SCA, Innate Pharma SA⋆, Ipsen SA⋆, Ipsos SA⋆, Intexa SA, Inventiva SA⋆, JCDecaux SE⋆, Jacquet Metals SA⋆, Clariane SE⋆, Lacroix Group SA, Lagardere SA⋆, Societe LDC SA, Laurent Perrier SA⋆, LVMH Moet Hennessy Louis Vuitton SE⋆, Lysogene SA, Etablissements Maurel et Prom SA⋆, Mcphy Energy SA, Media 6 SA, Medincell SA, Metabolic Explorer SA⋆, Mauna Kea Technologies SAS⋆, Wendel SE⋆, Nanobiotix SA⋆, Nicox SA,⋆ Neoen SA, Nexity SA⋆, Nhoa SA, NR 21 SA, Olympique Lyonnais Groupe SA⋆, OSE Immunotherapeutics SA, Groupe Partouche SA⋆, Pernod Ricard SA⋆, Pharnext SA, Compagnie Plastic Omnium SE⋆, Poxel SA⋆, Kering SA⋆, Publicis Groupe SA⋆, Pierre et Vacances SA⋆, Quadient SA⋆, Roche Bobois SA, Oeneo SA⋆, Sanofi SA⋆, Soditech SA, Vinci SA⋆, Compagnie de Saint Gobain SA⋆, CS Group SA⋆, NRJ Group SA⋆, Fonciere Volta SA, Stef SA⋆, Thales SA⋆, Technip Energies NV , Teleperformance SE⋆, Tipiak SA, Tarkett SA⋆, Trigano SA⋆, Transgene SA⋆, TotalEnergies SE⋆, Vantiva SA⋆, Veolia Environnement SA⋆, Vilmorin & Cie SA⋆, Vivendi SE⋆, Vallourec SA⋆, Valneva SE⋆, Voltalia SA⋆, Waga Energy SA, Eduformaction SA, Ashler et Manson SA, Speed Rabbit Pizza SA, Imalliance SA, Euroland Corporate SA, Hopening SA, Groupe Cioa Centre International Hotel Express-Golden-Bike + Ajout d'une SA, Hotelim SA	167

Administrative and support services	AbL Diagnostics SA, L'Air Liquide Societe Anonyme pour l'Etude et l'Exploitation des Procedes Georges Claude SA⋆, Aquila SA, Bluelinea SA, Compagnie Lebon SA, Bourrelier Group SA, Dlsi SA, Eurobio Scientific SA, GECI International SA, Mare Nostrum SA, Median Technologies SA, NSC Groupe SA, Voyageurs du Monde SA, Visiativ SA, Atos SE⋆, Bastide Le Confort Medical SA⋆, Eurazeo SE⋆, Compagnie Generale des Etablissements Michelin SCA , L'Oréal⋆, Synergie SE⋆, Phone Web	21
Public administration and defence; compulsory social security		0
Education	Audience Labs SA	1
Human health and social work activities	LNA Sante SA, Orpea SA⋆, Eduniversal SA	3
Arts, entertainment and recreation	Fermiere du Casino Municipal Cannes SA, La Française des Jeux SA , Societe Française de Casinos SA	3
Other service activities	Elis SA⋆	1
Activities of households as employers; undifferentiated goods- and services- producing activities of households for own use		0
Total		588

Source: Refinitiv and CDP data providers from COGEM Database (Institut Louis Bachelier, ⋆ *represents companies that are in both Refinitiv and CDP*