



UNIVERSITÀ  
DI PAVIA

**Dipartimento di Scienze Economiche e  
Aziendali**

Corso di Laurea magistrale in Finance

***ESG Drivers and Corporate Profitability:  
A Computational Experiment on ESG  
Measurement Divergence***

**Relatore:**

Chiar.mo Prof. Paolo Giudici

**Tesi di Laurea**  
di Francesca Aiello

Matr. n.556067

**Anno Accademico 2025-2026**



## **Abstract**

The empirically relevant question regarding ESG does not concern whether or not it improves corporate profitability in absolute terms, but rather the conditions of measurement, modelling and sectoral context in which it proves to be economically significant. This thesis therefore aims to analyze this issue through a controlled computational experiment on a simulated dataset comprising 1,000 firms observed over a ten-year period, breaking down the aggregate ESG scores into twelve specific drivers in line with the taxonomy developed by Berg, Kölbel and Rigobon (2022). The analysis is set within the framework of European legislation on sustainable finance (EU Taxonomy, SFDR, CSRD), which, despite its ambition, defines what to measure but not how to measure it, thus leaving the issue of measurement divergence unresolved. The results show that measurement divergence is the main source of empirical instability; indeed, even small changes in how a single ESG dimension is measured can radically alter the statistical significance of ESG factors, whereas divergence in weighting produces virtually negligible effects. Furthermore, a firm's sector, if left uncontrolled, emerges as a strong confounding factor, capable of generating spurious relationships in the absence of adequate controls. Overall, this thesis suggests that the debate on ESG should shift from the question "Is ESG important?" to the question "Using which metrics, for which dimensions, in which sectors and with what controls does ESG emerge as relevant?".



# Table of Contents

|                                                                     |    |
|---------------------------------------------------------------------|----|
| <b>Chapter 1 – Introduction</b>                                     | 9  |
| 1.1 Background: The Growing Importance of ESG Ratings               | 9  |
| 1.2 Problem Statement: The Divergence of ESG Ratings                | 10 |
| 1.3 Research Gap and Research Questions                             | 11 |
| 1.4 Methodological Approach                                         | 12 |
| 1.5 Main Findings                                                   | 13 |
| 1.6 Structure of the Thesis                                         | 14 |
| <b>Chapter 2 – Regulatory framework for Sustainable Finance</b>     | 15 |
| 2.1 Introduction: The EU Green Deal and the need for Regulation     | 15 |
| 2.2 The EU Taxonomy (Regulation EU 2020/852)                        | 17 |
| 2.2.1 <i>The situation before Taxonomy</i>                          | 17 |
| 2.2.2 <i>The structure of the Taxonomy</i>                          | 17 |
| 2.2.3 <i>Limitations and Critical issues</i>                        | 18 |
| 2.2.4 <i>Implications for Rating Divergence</i>                     | 19 |
| 2.2.5 <i>An overall assessment</i>                                  | 19 |
| 2.3 The Sustainable Finance Disclosure Regulation (EU 2019/2088)    | 20 |
| 2.3.1 <i>SFDR Role</i>                                              | 20 |
| 2.3.2 <i>The products' classification (Art.6, 8, 9)</i>             | 20 |
| 2.3.3 <i>Limitations and Implementation issues</i>                  | 21 |
| 2.3.4 <i>The Principal Adverse Impacts (PAI) and the Data Issue</i> | 22 |
| 2.3.5 <i>The Reform (SFDR 2.0) and Future Perspectives</i>          | 23 |
| 2.3.6 <i>Overall view</i>                                           | 24 |
| 2.4 The Corporate Sustainability Reporting Directive (EU 2022/2464) | 25 |
| 2.4.1 <i>The context: why the CSRD was needed</i>                   | 25 |
| 2.4.2 <i>The European Sustainability Reporting Standards (ESRS)</i> | 25 |
| 2.4.3 <i>The simplification process and the Omnibus reform</i>      | 26 |
| 2.4.4 <i>Implications for rating divergence</i>                     | 26 |
| 2.4.5 <i>An overall assessment</i>                                  | 27 |
| 2.5 Concluding remarks                                              | 28 |
| <b>Chapter 3 – Empirical Analysis: Methodology and Results</b>      | 29 |
| 3.1 Introduction to the Empirical Analysis                          | 29 |
| 3.1.1 <i>Research Questions</i>                                     | 29 |
| 3.1.2 <i>Methodology</i>                                            | 30 |
| 3.2 Dataset                                                         | 32 |
| 3.2.1 <i>Data Preparation and Descriptive Statistics</i>            | 33 |
| 3.2.2 <i>Limitations of the Dataset</i>                             | 34 |
| 3.3 Construction of ESG Drivers                                     | 35 |

|                                                                                   |           |
|-----------------------------------------------------------------------------------|-----------|
| 3.3.1 List of Drivers and Normalization.....                                      | 35        |
| 3.3.2 Simulation of Missing Drivers.....                                          | 36        |
| 3.3.3 Correlation Structure and Validation .....                                  | 37        |
| 3.4 Preprocessing and Models.....                                                 | 39        |
| 3.4.1 Feature Engineering: Interactions .....                                     | 39        |
| 3.4.2 Models Estimated .....                                                      | 40        |
| 3.5 Model Performance .....                                                       | 42        |
| 3.5.1 Actual vs Predicted and Residual Analysis .....                             | 44        |
| 3.6 Feature Importance.....                                                       | 46        |
| 3.6.1 Permutation Importance.....                                                 | 47        |
| 3.7 Within-Sector Analysis .....                                                  | 50        |
| 3.7.1 Methodology and Results .....                                               | 50        |
| 3.7.2 Implications for Materiality and Comparison with the Aggregated Model ..... | 52        |
| 3.7.3 Economic Interpretation.....                                                | 53        |
| 3.8 Divergence Experiments .....                                                  | 54        |
| 3.8.1 Measurement Divergence: The Case of Labor.....                              | 54        |
| 3.8.2 Scope Divergence: Excluding Environmental Drivers.....                      | 55        |
| 3.8.3 Weight Divergence: Artificial Ratings with different weights.....           | 55        |
| 3.8.4 False Positive: Sector Confounding.....                                     | 56        |
| 3.8.5 Cancellation .....                                                          | 58        |
| 3.9 Implications of ESG Rating Divergence.....                                    | 60        |
| 3.9.1 Evidence from the Divergence Experiments.....                               | 61        |
| 3.10 Robustness Checks.....                                                       | 63        |
| 3.10.1 Summary of Robustness Findings.....                                        | 64        |
| 3.11 Economic Interpretation.....                                                 | 66        |
| 3.11.1 Environmental and Governance Drivers.....                                  | 66        |
| 3.11.2 Economic Significance and the ESG Paradox .....                            | 67        |
| 3.12 Summary of Empirical Results.....                                            | 68        |
| <b>Chapter 4 – Discussion and Implications.....</b>                               | <b>70</b> |
| 4.1 Literature on ESG Ratings Divergence .....                                    | 70        |
| 4.2 Literature on ESG and Financial Performance.....                              | 72        |
| 4.3 Positioning of This Thesis in Literature .....                                | 73        |
| 4.4 Discussion of Empirical Findings .....                                        | 74        |
| 4.4.1 Key Findings and Their Interpretation .....                                 | 74        |
| 4.4.2 Comparison with the Existing Literature .....                               | 75        |
| 4.4.3 Implications for Stakeholders.....                                          | 76        |
| <b>Chapter 5 – Conclusions .....</b>                                              | <b>78</b> |
| 5.1 Overall Summary of Findings .....                                             | 78        |
| 5.2 Contribution of the Thesis.....                                               | 80        |

|                                                          |                                     |
|----------------------------------------------------------|-------------------------------------|
| 5.3 Limitations and Directions for Future Research ..... | 82                                  |
| 5.4 Concluding Remarks .....                             | 84                                  |
| <b>References .....</b>                                  | <b>86</b>                           |
| <b>Ringraziamenti .....</b>                              | <b>Error! Bookmark not defined.</b> |



# Chapter 1 – Introduction

## 1.1 Background: The Growing Importance of ESG Ratings

Over the past decade, ESG (Environmental, Social, Governance) ratings have become an indispensable benchmark for investors, regulators and researchers. This is demonstrated by the 2024 annual report of the Principles for Responsible Investment (PRI), which lists over 5,000 signatories and assets under management exceeding 128.4 trillion USD<sup>1</sup>, which testify to the importance these assessments have acquired in global financial markets. Consequently, academic literature has devoted increasing attention to the relationship between ESG performance and financial performance, producing a growing number of empirical studies that use ESG ratings as a key explanatory variable<sup>2</sup>.

In this context, ESG ratings are used not only as indicators of sustainability, but also as informational tools for risk assessment, portfolio construction and the evaluation of corporate resilience to environmental and regulatory shocks. However, this growing use is hampered by an equally well-documented problem: ESG ratings provided by different agencies diverge substantially; indeed, two providers assessing the same company using different methodologies may produce markedly different scores, thereby causing confusion among investors and instability in the results of empirical research<sup>3</sup>.

---

<sup>1</sup> Principles for Responsible Investment (2024). *Annual Report 2024*. <https://www.unpri.org/about-pri/annual-report>

<sup>2</sup> Friede, G., Busch, T., Bassen, A. (2015). ESG and financial performance: aggregated evidence from more than 2000 empirical studies. *Journal of Sustainable Finance & Investment*, 5(4), 210–233.

<sup>3</sup> Berg, F., Kölbel, J.F., Rigobon, R. (2022). Aggregate Confusion: The Divergence of ESG Ratings. *Review of Finance*, 26(6), 1315–1344.

## 1.2 Problem Statement: The Divergence of ESG Ratings

The phenomenon of ESG rating divergence has been widely documented in the literature; indeed, examples include Chatterji et al. (2016)<sup>4</sup>, which highlighted the lack of agreement between different providers, documenting a low comparability between scores even after adjusting for explicit differences in the definitions used. Furthermore, Berg, Kölbel and Rigobon (2022)<sup>5</sup> expanded the analysis by examining the phenomenon across a sample of six ESG rating agencies, which revealed correlations ranging from 0.38 to 0.71, showing a surprisingly high level of disagreement compared with credit ratings, where correlations exceed 0.99<sup>6</sup>.

It is therefore evident that this heterogeneity has direct implications for the efficiency of capital markets, as it compromises the comparability of ESG information and may influence capital allocation decisions in the presence of inconsistent information signals. Therefore, the authors themselves have broken down this divergence into three main components<sup>7</sup>:

- **Measurement divergence (56%)**: different ways of measuring the same attributes;
- **Scope divergence (38%)**: different sets of attributes considered;
- **Weight divergence (6%)**: different weights assigned to attributes.

This breakdown is more significant than it might appear at first glance; in fact, it is precisely because of this that the focus shifts from the debate on weights, that is, which pillars should carry greater weight and therefore how to balance E against S and G, to the crucial issue of measurement. That said, the existing literature has focused primarily on documenting the existence of divergence, rather than on its implications for empirical analysis; in other words, we know that ratings diverge, but we do not yet know to what extent this divergence influences the conclusions that researchers and investors draw from the ratings. Hence, this thesis aims to fill this existing gap.

---

<sup>4</sup> Chatterji, A.K., Durand, R., Levine, D.I., Touboul, S. (2016). Do Ratings of Firms Converge? Implications for Managers, Investors and Strategy Researchers. *Strategic Management Journal*, 37(8), 1597–1614.

<sup>5</sup> Berg, F., Kölbel, J.F., Rigobon, R. (2022) Aggregate Confusion: The Divergence of ESG Ratings. *Review of Finance*, 26(6), 1315–1344.

<sup>6</sup> Ibid.

<sup>7</sup> Ibid.

### 1.3 Research Gap and Research Questions

The key question underpinning this thesis is the following: to what extent do measurement choices, in particular *measurement divergence*, influence the empirical results of analyses on the relationship between ESG and corporate profitability? It is crucial to understand this mechanism not only for academic literature but also for the growing European regulatory framework, which is implicitly based on the idea that ESG ratings are comparable and consistent in terms of the information they provide. Consequently, four specific research questions arise from this overarching question:

**1. Which specific ESG dimensions are most relevant to corporate profitability?**

By breaking down aggregate ESG scores into specific drivers (environmental, social and governance), the analysis aims to identify which components drive the observed relationship and whether there are therefore systematic differences between the three pillars.

**2. How does the sector in which a firm operates influence the relationship between ESG and profitability?**

A review of the existing literature indicates that the sector has a strong influence on profitability<sup>8</sup>; consequently, the analysis examines whether ESG effects remain robust when sector controls are introduced and whether they vary depending on the sector under consideration.

**3. To what extent do measurement choices influence the empirical results?**

Using the framework of Berg et al. (2022) as a basis<sup>9</sup>, the analysis tests the impact of *measurement divergence*, *scope divergence* and *weight divergence* on the stability of the results, the statistical significance of the drivers and the predictive power of the models.

**4. What is the economic significance of ESG drivers compared with other firm characteristics?**

In addition to statistical significance, the analysis also examines the relative predictive importance of ESG drivers compared with sectoral and financial variables, thereby distinguishing between statistical and economic significance.

---

<sup>8</sup> Eccles, R.G., Ioannou, I., Serafeim, G. (2014). The Impact of Corporate Sustainability on Organizational Processes and Performance. *Management Science*, 60(11), 2835–2857.

<sup>9</sup>Berg, F., Kölbel, J.F., Rigobon, R. (2022). Aggregate Confusion: The Divergence of ESG Ratings. *Review of Finance*, 26(6), 1315–1344.

## 1.4 Methodological Approach

In order to answer these questions, this thesis presents the results of a controlled computational experiment performed on a simulated dataset. Although this approach cannot be generalized to real-world contexts in quantitative terms, it does allow us to isolate the effect of measurement choices, thereby avoiding the problems of endogeneity and missing data that characterize studies using real data. Furthermore, the use of a simulated dataset enables the systematic variation of methodological choices, namely, measurement, the set of attributes and aggregation weights, and the observation of their consequences for empirical inference in a controlled environment. Details of the dataset, the construction of the drivers and the estimated models are presented and discussed in Chapter 3.

## 1.5 Main Findings

The analysis yields four main findings, which will be discussed in detail in Chapter 3. Firstly, with regard to the relative importance of ESG dimensions, it will be demonstrated that the Random Forest model offers the best predictive performance and, although ESG drivers contribute to the model's predictive power, they contribute less than sectoral variables. In fact, it will be observed that environmental and governance drivers are more informative than social ones, suggesting that ESG factors act as complementary information signals rather than as fundamental causal drivers of profitability.

Regarding the role of sector affiliation, ESG materiality varies significantly across sectors; this finding suggests that some of the ESG literature may have overestimated the direct effect of ESG factors by failing to distinguish adequately between structural sectoral effects and genuine effects linked to corporate sustainability.

Subsequently the third finding, which is also the most methodologically significant, with regard to the impact of measurement choices, shows that measurement divergence emerges as the main source of empirical instability; indeed, it will become clear that by altering the measurement of a single driver (e.g. Labor), the same concept shifts from being statistically insignificant to significant, which will prove that the divergence in ratings is not merely a disagreement between agencies in the abstract but it is a concrete demonstration that measurement choices propagate directly into empirical conclusions. Weight divergence, in contrast, turns out to be almost irrelevant; varying the weights assigned to drivers produces negligible differences in model performance.

Finally, in terms of economic significance, the twelve ESG drivers account for around 31% of the model's total predictive power, compared with approximately 38.5% attributable to industry sector dummies. Hence, ESG is not noise, but neither is it the main driver of profitability: it acts as a source of additional information rather than as a dominant explanatory factor.

## 1.6 Structure of the Thesis

The thesis is structured as follows: Chapter 2 presents the European regulatory framework for sustainable finance (EU Taxonomy, SFDR, CSRD), which illustrates how current regulations define what to measure but not how to measure it, thus leaving the issue of divergence unresolved. Chapter 3 forms the empirical core, as it describes the dataset, the construction of the ESG drivers, the estimated models and the results of the divergence experiments. Chapter 4 frames the results within the broader academic debate on rating divergence and the relationship between ESG and financial performance, drawing implications for investors, regulators and rating agencies. Finally, Chapter 5 summarizes the thesis' contributions and outlines directions for future research.

## Chapter 2 – Regulatory framework for Sustainable Finance

### 2.1 Introduction: The EU Green Deal and the need for Regulation

In recent years, the concept of promoting finance that is as sustainable as possible has gained increasing importance, evolving from a niche topic to the focus of European economic policy. This shift has been largely driven by the so-called European Green Deal, launched in 2019, which marked a decisive turning point by setting a highly ambitious goal: to make Europe the first climate-neutral continent by 2050.<sup>10</sup> Consequently, the focus has gradually shifted to the importance of sustainability in every area, requiring intersectoral action across all sectors. To achieve this, the European Commission has emphasized the need to channel substantial private financial resources into sustainable projects.<sup>11</sup> In this context, financial markets play a key role in redirecting investment flows; however, to function effectively, the markets require clarity, transparency and common standards, and it is precisely because of this need that the 2018 Action Plan on Financing Sustainable Growth was established, a document that laid the foundations for the entire regulatory framework that followed.<sup>12</sup> This regulatory framework is particularly relevant to the study of ESG ratings, as it helps shape the institutional and informational context in which such ratings are produced and used. In the past, the main limitations of ESG ratings were due precisely to the significant differences in the methodologies used by the individual agencies and to the lack of transparency and availability of comparable data.<sup>13</sup> In other words, because of these limitations, two similar companies could receive very different ESG scores depending on the method adopted by the various rating agencies, which has led to greater confusion amongst investors. Consequently, the various initiatives launched by the European Union aim to address these issues by introducing common criteria to define what is meant by “truly sustainable” in economic activities and, at the same time, by strengthening transparency obligations for both companies and financial operators.<sup>1415</sup> The connection with the divergence in ratings is straightforward: if regulations require greater sustainability, companies must be assessed based on how sustainable they actually are, and this can only happen if rating agencies adopt common methodologies. Without such alignment, scores will continue to diverge, making it impossible to obtain a clear picture of Europe’s progress towards its 2050 targets.

---

<sup>10</sup> European Commission, *The European Green Deal*, COM(2019) 640 final, 11 December 2019

<sup>11</sup> Ibid.

<sup>12</sup> European Commission, *Action Plan: Financing Sustainable Growth*, COM(2018) 97 final, 8 March 2018

<sup>13</sup> Berg, F., Kölbel, J. F., & Rigobon, R. (2022). Aggregate Confusion: The Divergence of ESG Ratings. *Review of Finance*, 26(6), 1315–1344.

<sup>14</sup> European Securities and Markets Authority (ESMA), Final Report on Greenwashing, 4 June 2024, p. 8.

<sup>15</sup> Regulation (EU) 2020/852.

This chapter therefore sets out to outline the three fundamental pillars of the regulatory framework created to increase the quantity and quality of ESG information available on the market: the European Taxonomy, the SFDR Regulation and the CSRD. The aim is to demonstrate how these principles interact with one another and how, together, they form an ecosystem in which ESG ratings can be based on more robust and comparable information, thereby becoming more useful tools for sustainable investment decisions.

## 2.2 The EU Taxonomy (Regulation EU 2020/852)

### 2.2.1 *The situation before Taxonomy*

Before the EU Taxonomy was introduced, there was not a common definition which identified what truly makes an economic activity “green”. A firm could be considered sustainable mainly based on vague and biased criteria, such as simple waste sorting, without any actual commitment towards broader environmental purposes. Therefore, this lack of clear guidelines generated confusion among market operators while leaving room for superficial statements, making it complex to distinguish between authentic initiatives and simple greenwashing operations.<sup>16</sup>

### 2.2.2 *The structure of the Taxonomy*

The EU 2020/852 Taxonomy Regulation introduced an innovative system of classification into the European Union, which identifies six environmental objectives<sup>17</sup>:

1. Climate change mitigation
2. Climate change adaptation
3. Sustainable use and protection of water and marine resources
4. Transition to a circular economy
5. Pollution prevention and control
6. Protection and restoration of biodiversity and ecosystems

According to Article 3 of this regulation, to ensure that an economic activity can be considered in line with the Taxonomy and so environmentally sustainable, it must meet four cumulative conditions<sup>18</sup>:

1. Substantial contribution: contribute substantially to at least one of the six environmental objectives, as defined in the technical selection criteria established in the delegated acts<sup>19</sup>;
2. Do No Significant Harm (DNSH): not causing a significant harm to any of the other objectives, according to Article 17 of the Regulation<sup>20</sup>;
3. Minimum safeguards: to comply with the minimum safeguards specified in Article 18, including the alignment with the OECD Guidelines for

---

<sup>16</sup> ESMA, *Final Report on Greenwashing*, cit., p. 8.

<sup>17</sup> Regulation (EU) 2020/852, Art. 9.

<sup>18</sup> Ibid., Art. 3

<sup>19</sup> Commission Delegated Regulation (EU) 2021/2139 of 4 June 2021 supplementing Regulation (EU) 2020/852 (Climate Delegated Act) [2021] OJ L442/1.

<sup>20</sup> Regulation (EU) 2020/852, Art. 17.

Multinational Enterprises and the UN Guiding Principles on Business and Human Rights<sup>21</sup>;

4. Technical screening criteria: comply with the detailed technical criteria set out by the European Commission through delegated acts, specifying the requirements for both substantial contribution and DNSH for each economic activity.<sup>22</sup>

In formal terms, such structure represents a significant step forward. However, the absence of measurement tools leaves room for interpretation.<sup>2324</sup> Some companies could formally meet the criteria without turning them into concrete outcomes, while others, particularly those in sectors where it is more challenging to align with the criteria, face the risk of being penalized regardless of their commitment.<sup>25</sup>

### *2.2.3 Limitations and Critical issues*

One of the main limitations of Taxonomy lies in its rigid application. Indeed, having to set identical parameters for totally different sectors, such as the fashion industry compared to the arms industry, might cause distortions; or, in the case of an industry with high energy consumption, such as steelworks, could still be classified as non-sustainable even while investing in lower-emission technologies, simply because its manufacturing processes remain, by their own nature, a source of emissions. In the absence of sector-specific criteria, there is a risk of discouraging exactly those sectors that would most benefit from a gradual transition.<sup>26</sup>

A second element of concern relates to the scope of the Taxonomy, which focuses almost entirely on environmental dimensions, whilst ignoring both social and governance aspects; however, given that the notion of ESG is, by definition, composed of three parts, this choice risks producing an incomplete representation of corporate sustainability.<sup>27</sup> As will be seen in the following chapters, this asymmetry contributes to divergences in ESG ratings: two agencies that share the same definition of “green” may still arrive at different assessments if they use different metrics to measure.<sup>28</sup>

---

<sup>21</sup> Ibid., Art.18

<sup>22</sup> Commission Delegated Regulation (EU) 2021/2139, cit.

<sup>23</sup> ESMA, *Final Report on Greenwashing*, cit., p. 12

<sup>24</sup> Berg et al. (2022), cit.

<sup>25</sup> Berg et al. (2022), cit.

<sup>26</sup> Platform on Sustainable Finance, *Technical Working Group – Part A: Methodological Report*, March 2022, p. 15.

<sup>27</sup> Platform on Sustainable Finance, *Final Report on Social Taxonomy*, February 2022, Section 2.1

<sup>28</sup> Berg et al. (2022), cit.

#### 2.2.4 Implications for Rating Divergence

The Taxonomy defines what is “green” but not how to measure it, and this is a central issue in understanding the phenomenon of divergence.<sup>29</sup> In fact, two rating agencies may be using the same definition of environmental sustainability but employ different metrics to measure it. One might use parameters tailored to the core business of the company being assessed, whilst the other might use generic benchmarks that do not capture the specific characteristics of the sector.<sup>30</sup> The consequence is that, even under a common regulatory framework, ESG scores continue to change, causing confusion among investors and reducing the usefulness of these scores as a means of allocating capital.<sup>31</sup>

#### 2.2.5 An overall assessment

In consideration of the previous points, Taxonomy can be seen as a necessary but insufficient starting point.<sup>3233</sup> It provides a common vocabulary and establishes clear parameters for what can be considered sustainable but on the other hand, it leaves unanswered the issues of measurement and comparability, which lie at the core of the divergence in ratings. For the regulatory framework to be fully effective, it will thus need to be supported by instruments that harmonize not only definitions, but also metrics and assessment methodology.<sup>34</sup>

---

<sup>29</sup> Christensen, D.M., Serafeim, G., Sikochi, A. (2022). Why is Corporate Virtue in the Eye of the Beholder? The Case of ESG Ratings. *The Accounting Review*, 97(1), 147–175.

<sup>30</sup> Chatterji, A.K. et al. (2016), *Do Ratings of Firms Converge?*, *Strategic Management Journal*, 37(8), pp. 1597–1614.

<sup>31</sup> ESMA, *Final Report on Greenwashing*, cit., p. 22

<sup>32</sup> Platform on Sustainable Finance, Report on Advancing Sustainable Finance, April 2025

<sup>33</sup> Berg et al. (2022), cit.

<sup>34</sup> Berg et al. (2022), cit.

## 2.3 The Sustainable Finance Disclosure Regulation (EU 2019/2088)

### 2.3.1 SFDR Role

While Taxonomy defines *what* is “green”, the SFDR regulation explains *how* financial operators should behave. This regulation is addressed mainly to fund managers, financial advisers and institutional investors, forcing them to set out their strategies for sustainability to increase transparency. Financial operators must make it clear how they integrate the ESG factors into their decisions, but also, as importantly, they must be honest about what they are not doing or what they are still not able to do in order to make their products more sustainable. In short, the SFDR regulation seeks to help redirect private capital towards initiatives that contribute to the green transition.

### 2.3.2 The products’ classification (Art.6, 8, 9)

SFDR classifies financial products in three main categories, which have now become part of the everyday vocabulary:

- Art.6 products: funds that do not integrate any sustainability considerations;
- Art.8 products (called “light green”): funds that promote environmental or social characteristics;
- Art.9 products (called “dark green”): funds that have sustainable investments as their explicit objectives.

On paper, this could seem to have a clear structure. However, in practice, such distinction has turned out to be much more complicated; indeed, already in 2022, the European Commission itself, in its review of the regulation, had found evidence showing that Articles 8 and 9 were being used as marketing labels instead of just indicators of transparency, creating confusion among investors.

The distinction between these categories is vague: a fund may fit in more than one category and the parameters used for its classification are not always clear. Consequently, this ambiguity leaves room for interpretation, hence, the risk is tangible. Independent organizations such as Facing Finance, whose job is to monitor the markets, have reported that many funds categorized in Article 8 or 9 continue to hold significant investments in companies involved in controversial sectors, such as military production or harmful activities for the environment. Investor confidence and the credibility of the entire system can be undermined if a behavior of this kind becomes widespread.

### 2.3.3 Limitations and Implementation issues

The actual implementation of the SFDR has brought to light a number of issues, which have been recognized by both supervisory authorities and market participants and have been the subject of in-depth analysis in the subsequent reform.<sup>35</sup>

The first issue concerns the clarity of the concepts set out in the reform. Very often, in fact, key definitions in the regulation have been deemed unclear or inconsistent, creating uncertainty in their interpretation and leaving room for inconsistent application, which has made it difficult for individual investors to compare financial products claiming similar sustainability objectives, thereby undermining the very objective of transparency.<sup>3637</sup> An example of this is the very definition of ‘sustainable investment’, which has been criticized for being too vague and has led to widely differing interpretations.<sup>3839</sup>

A second problem, which emerged from the analysis, relates to the lack of reliable and comparable ESG data. This is because the effectiveness of the regulation depends precisely on the quality of the underlying data provided by companies; consequently, the lack of such data has made it difficult for financial market participants to effectively implement reporting obligations, which have been fulfilled on the basis of incomplete, estimated or data collected using different methodologies, thereby limiting the comparability of the overall information available to investors.<sup>4041</sup> These issues have had tangible and significant consequences, as in the case of the widespread use of Articles 8 and 9 as de facto product labels, even though the SFDR was not designed as a labelling system.<sup>424344</sup> Consequently, this practice has caused confusion and increased the risk of greenwashing and misleading marketing. Confirming this, an analysis conducted by the European Securities and Markets Authority (ESMA) found that statements on adverse impacts (PAIs) lacked detail and were generally inadequate in demonstrating how investments

---

<sup>35</sup> European Commission, *Call for Evidence: Simplification and Improved Effectiveness of the SFDR*, 2 May 2025.

<sup>36</sup> Ibid.

<sup>37</sup> Cochran, I., Mackenzie, C., Brander, M. (2025). EU's sustainable finance disclosure regulation: Does the hybrid reporting regime undermine the goal to reorient capital to climate action? *Climate Policy*, 25(1), 76–88. <https://doi.org/10.1080/14693062.2024.2353115>

<sup>38</sup> Advokatfirmaet BAHR, *SFDR 2.0: From Paper Exercise to Impact?*, 5 June 2025.

<sup>39</sup> International Capital Markets Association (ICMA), *Paper on Simplification of EU Sustainable Finance Legislation*, 5 February 2025.

<sup>40</sup> European Commission, *Call for Evidence*, cit.

<sup>41</sup> European Securities and Markets Authority (ESMA). *Opinion on the functioning of the Sustainable Finance Framework*. ESMA36-1079078717-2587, 24 July 2024.

<sup>42</sup> BAHR, *SFDR 2.0*, cit.

<sup>43</sup> European Commission. *Proposal for a Regulation of the European Parliament and of the Council amending Regulation (EU) 2019/2088 on sustainability-related disclosures in the financial services sector (SFDR 2.0)*. COM(2025) 841 final, 20 November 2025.

<sup>44</sup> Finance Watch, *Response to Consultation*, 26 March 2025.

complied with the Do No Significant Harm (DNSH) principle.<sup>45</sup>

Finally, there is also the complexity and fragmentation of the regulatory framework itself: the SFDR interacts with other key European legislation, such as the Taxonomy and the CSRD, and therefore does not operate in isolation. However, it has been observed that these regulatory interactions are often characterized by overlaps, inconsistencies and interpretative ambiguities which, paradoxically, contribute to the very divergence that the regulations aim to reduce, making the system even more complex to manage.<sup>4647</sup>

In summary, these issues with the application of the SFDR have fueled market confusion and made the regulatory framework less effective than anticipated.

#### *2.3.4 The Principal Adverse Impacts (PAI) and the Data Issue*

Among the SFDR's objectives, the most ambitious is the requirement to report on Principal Adverse Impacts (PAI), i.e. the negative effects that investment decisions may have on the environment and society.<sup>48</sup> This is, in principle, a solid and widely accepted underlying principle, as an investor should not only know how 'green' a fund is, but also what its negative impacts might be, such as in terms of emissions, water consumption, and respect for human rights.

The point is that, in this case too, the main problem remains the information, as the data needed to calculate PAI indicators are often:

- Incomplete: not all companies report in a uniform way;
- Estimated: in the absence of reliable data, proxies and estimates are used;
- Inconsistent: collected using different methodologies and by different operators.<sup>49</sup>

This was highlighted in the 2025 joint report by the European Supervisory Authorities (ESAs)<sup>50</sup>; while recognizing some improvements in the quality of information, it points out that many disclosures have not been properly quantified. In particular, data coverage for certain key indicators, such as water use, waste management or the gender pay gap, remains incomplete.

Therefore, in this case as well, two managers might apply the same PAI criteria but, by using data from different sources or different computational methods, reach opposite conclusions regarding the same investment.<sup>51</sup> This situation is

---

<sup>45</sup> ESMA, *Final Report on Greenwashing*, cit., pp. 22–25.

<sup>46</sup> European Commission, *Call for Evidence*, cit.

<sup>47</sup> Cochran et al. (2025), cit.

<sup>48</sup> Regulation (EU) 2019/2088 (SFDR), Arts. 4–7.

<sup>49</sup> European Supervisory Authorities (ESAs). *Principal Adverse Impact Disclosures under the Sustainable Finance Disclosure Regulation — Annual Report to the Commission under Article 18 of Regulation (EU) 2019/2088*. JC 2025 26, September 2025.

<sup>50</sup> ESAs, *Principal Adverse Impact Disclosures under the Sustainable Finance Disclosure Regulation*, cit., Section 3.2.

<sup>51</sup> Berg et al. (2022), cit.

a prime example of the problem of measurement divergence, which is the central focus of this thesis. Consequently, if the underlying data is inconsistent, so also will be the resulting indicators; and if the indicators are inconsistent, then an investor cannot blindly trust the ESG ratings assigned to funds.

The problem, once again, lies not only with the funds or the managers, but in the absence of a common and shared basis for measuring sustainability.

### *2.3.5 The Reform (SFDR 2.0) and Future Perspectives*

In response to the various shortcomings identified in the SFDR, at the end of November 2025, following a consultation process, the European Commission presented a proposal to revise the regulation, also known as SFDR 2.0.

This reform represents a significant step forward, as it marks a shift from a system based purely on transparency to one based on distinct product categories<sup>52</sup>. The previous categories under Articles 8 and 9 are therefore replaced by new classes designed to provide more explicit information. In particular, each category sets out clear minimum requirements, such as an investment threshold (around 70% of the portfolio must be in line with the stated sustainability strategy) and mandatory exclusions for harmful sectors (such as tobacco, coal, etc.)<sup>53</sup>. On the one hand, the aim is to reduce greenwashing and improve investor protection; on the other, to simplify obligations and reduce administrative costs for financial operators, in line with the simplification and competitiveness objectives of the Draghi agenda<sup>54</sup>.

The changes made are therefore consistent, as they acknowledge the confusion generated by the current framework and respond to the various calls for greater clarity in the market. Nevertheless, organizations such as Frank Bold<sup>55</sup>, having analyzed the proposal, state that certain risks remain, including the fact that fund managers still have the option to continue using subjective criteria to select investments, and the lack of stringent obligations regarding negative impacts (PAI) across all categories, which could still leave a degree of ambiguity and make products difficult to compare.

In conclusion, SFDR 2.0 helps to reduce confusion and represents an important step forward, but one that requires further formalization, particularly regarding the measurement and comparability of data.

---

<sup>52</sup> European Commission. *Proposal for a Regulation of the European Parliament and of the Council amending Regulation (EU) 2019/2088 on sustainability-related disclosures in the financial services sector (SFDR 2.0)*. COM(2025) 841 final, 20 November 2025.

<sup>53</sup> European Commission, Proposal, cit., Annex II.

<sup>54</sup> European Commission. *Omnibus I Package*. COM(2025) 80 final, 26 February 2025.

<sup>55</sup> Frank Bold, Analysis of the SFDR 2.0 Proposal, February 2026.

### *2.3.6 Overall view*

In conclusion, both Taxonomy and SFDR represent a genuine effort to build a coherent regulatory framework aimed at achieving the most sustainable finance possible. In particular, the SFDR, in both its current and reformed form, remains a useful tool for guiding investors towards more informed choices. However, its effectiveness is still limited by the same weaknesses that characterize the entire European sustainable finance system: a lack of robust data and common measurement methodologies. Despite this, the SFDR addresses important issues and imposes transparency on them, but in doing so, it highlights the weaknesses of the system itself. It is clear that the divergence in ratings is therefore a direct consequence of these underlying weaknesses, and the only way to overcome them is to get to the root of the problem by creating a common and shared language for measuring sustainability; until this happens, confusion and skepticism will continue to reign supreme.<sup>5657</sup>

---

<sup>56</sup> European Commission, Call for Evidence, cit.

<sup>57</sup> Cochran et al. (2025), cit.

## 2.4 The Corporate Sustainability Reporting Directive (EU 2022/2464)

### 2.4.1 The context: why the CSRD was needed

The CSRD (Directive (EU) 2022/2464) was introduced to replace the previous Non-Financial Reporting Directive (NFRD), which had been in force since 2014.<sup>58</sup> This was because it turned out to be too weak; in essence, it gave companies too much freedom over what to disclose, with the result that the information published was often fragmented, inconsistent and frequently not comparable. In fact, in many cases, sustainability reports, produced according to the NFRD criteria, were nothing more than marketing exercises, with little substantive data to support the claims.<sup>59</sup> Consequently, the European Commission decided to radically change the approach by introducing the CSRD, with the new aim of bringing sustainability reporting up to the same level of rigor and comparability as financial reporting.<sup>60</sup>

The directive introduces mandatory reporting standards, expands the number of companies involved and requires an external review of the information reported; in other words, the goal was for ESG information to become a tool as robust as financial statements for investors.<sup>61</sup>

### 2.4.2 The European Sustainability Reporting Standards (ESRS)

At the heart of the CSRD lie the European Sustainability Reporting Standards (ESRS), developed by the European Financial Reporting Advisory Group (EFRAG). These are mandatory standards designed to spell out exactly what information companies must disclose on various environmental, social and governance subjects. The first set of ESRS, adopted by the European Commission in July 2023, comprises 12 standards, including two transversal standards (ESRS 1 and 2) and ten thematic standards, which cover several areas such as climate change, pollution, the workforce and business conduct.<sup>62</sup> The main innovation introduced by these standards is the principle of double materiality<sup>63</sup>, according to which companies must report not only on how sustainability factors, such as climate or social issues, affect their financial performance (financial materiality) but also on how their activities impact the environment and society (impact materiality). It is precisely this dual

---

<sup>58</sup> European Commission, *Proposal for a Corporate Sustainability Reporting Directive*, COM(2021) 189 final, 21 April 2021, p. 2.

<sup>59</sup> Directive (EU) 2022/2464 (CSRD), Recital 6.

<sup>60</sup> European Commission, *Proposal for a Corporate Sustainability Reporting Directive*, COM(2021) 189 final, 21 April 2021, Section 1.1.

<sup>61</sup> CSRD, Recital 8.

<sup>62</sup> European Commission, *Delegated Regulation (EU) 2023/2772 (ESRS)*, 31 July 2023.

<sup>63</sup> EFRAG, *ESRS Explanatory Note: Double Materiality*, November 2023, p. 3.

perspective that distinguishes the European framework from other reporting regimes, such as the ISSB, which focuses almost exclusively on financial materiality.<sup>64</sup><sup>65</sup>

#### *2.4.3 The simplification process and the Omnibus reform*

The practical implementation of the CSRD and the ESRS has, once again, proved to be challenging. The first difficulties arose as soon as companies began adopting the ESRS: the standards were too complex, the number of data points required was excessively high, and the costs of compliance turned out to be much higher than anticipated. For this reason, the European Commission started a simplification process, included in the so-called Omnibus I Package, which was announced in February 2025<sup>66</sup>, with the aim of reducing the bureaucratic load without compromising the underlying ambition. In practice, the package, which was finally adopted in November 2025<sup>67</sup>, cut the number of mandatory data points by around 70% and granted small and medium-sized enterprises (SMEs) an additional two years to adapt. The aim, therefore, was to lighten the administrative burden while preserving the competitiveness of European companies.<sup>68</sup>

#### *2.4.4 Implications for rating divergence*

On the surface, the CSRD appears to be an optimal solution, as it generates more data, greater transparency and improved comparability. The problem, however, is that it fails to specify how this additional data should be used for ratings. In practical terms, two rating agencies might receive the same detailed sustainability report but apply different metrics to convert it into a score. For example, one might focus more on rewarding companies with ambitious environmental policies, whereas the other might focus more on the impacts actually measured. The result is that, once again, the divergence between ratings may persist. Consequently, the problem is not only a lack of data, but also is the absence of shared measurement methodologies.<sup>69</sup>

---

<sup>64</sup> Ibid.

<sup>65</sup> International Sustainability Standards Board (ISSB), *IFRS S1*, June 2023.

<sup>66</sup> European Commission, *Omnibus I Package*, COM(2025) 80 final, 26 February 2025.

<sup>67</sup> Ibid.

<sup>68</sup> Ibid.

<sup>69</sup> Berg et al. (2022), cit.

#### 2.4.5 *An overall assessment*

In light of these considerations, the CSRD represents a decisive step forward compared to the NFRD, as it provides a common framework, imposes detailed standards and brings sustainability reporting closer to financial accounting standards. Nevertheless, it is not sufficient on its own to resolve the issue of divergence; as with the Taxonomy and the SFDR, the CSRD defines *what* must be measured, but still leaves open the choices regarding *how* to measure it. For the system to be truly effective, it will be necessary to go further, namely by complementing reporting obligations with a harmonization of metrics and calculation methodologies; a step that this thesis aims to highlight.<sup>70</sup>

---

<sup>70</sup> Berg et al. (2022), cit.

## 2.5 Concluding remarks

In conclusion, it has been observed that, about financial sustainability, the European regulatory framework is essentially based on three fundamental and interlinked pillars: the Taxonomy, the SFDR and the CSRD. Together, these aim to establish a common language for sustainability, seeking to ensure transparency in the decisions made by financial operators and improving the quality and comparability of corporate data. Although it appears to be an ambitious and articulated system, it also highlights an underlying fragility, namely the fact that all three regulations deal with defining what should be measured, but none of them harmonizes how to measure it. Consequently, the various choices regarding metrics, indicators and calculation methodologies remain, in most cases, delegated to market operators or individual companies. It is precisely this lack of harmonization that lies at the root of the divergence in ESG ratings and, consequently, of the instability of empirical results.

The following thesis takes its starting point from this very finding; indeed, if the data are more numerous but not comparable, if the metrics are different but not aligned, then any analysis of the relationship between ESG and corporate performance risks being fragile, as it depends on the measurement choices adopted. The following chapters will demonstrate, through a computational experiment on a panel of simulated data, how much these choices can influence the results and why the measurement divergence is, in fact, the key to interpreting the empirical debate on ESG.<sup>7172</sup>

---

<sup>71</sup> Platform on Sustainable Finance, *Report on Advancing Sustainable Finance*, April 2025

<sup>72</sup> Berg et al. (2022), cit.

## Chapter 3 – Empirical Analysis: Methodology and Results

### 3.1 Introduction to the Empirical Analysis

This chapter constitutes the core of this thesis and presents the empirical analysis developed to investigate how measurement choices influence the relationship between ESG factors and corporate profitability. The study, which is based on the framework developed by Berg, Kölbel, and Rigobon (2022)<sup>73</sup>, that divides the divergence in ESG ratings into three main components: scope, measurement, and weight; employs a computational experiment that tests the effect of different measurement choices on the outcomes of ESG profitability research. Following the discussion in Chapter 1, this thesis focuses on examining the consequences of leaving measurement methodologies open to interpretation; in fact, by systematically varying measurement choices, the fragility of empirical results and the critical gap in current regulation are brought to light.

#### 3.1.1 Research Questions

The empirical analysis was structured around four main research questions:

***1. Which ESG dimensions have the most significant impact on corporate profitability?***

The analysis aims to identify which environmental, social, or governance components drive the observed relationship with profitability, and does so by breaking down aggregate ESG scores into twelve specific drivers.

***2. How does industry sector influence the relationship between ESG and profitability?***

The literature indicates a strong influence of the industry sector on profitability; consequently, the analysis examines whether, by introducing sector controls, the ESG effects are robust and whether they vary depending on the sector being considered.

***3. To what extent do measurement choices influence the empirical results?***

The analysis, following the Berg et al. (2022)<sup>74</sup> paper, tests the impact of the three components of divergence on the stability of the results: measurement divergence, i.e., the different ways of measuring the same concept; scope divergence, referring to the different sets of attributes; and finally, weight

---

<sup>73</sup> Berg et al. (2022), cit.

<sup>74</sup> Ibid.

divergence, i.e., the different aggregation weights.

#### ***4. What is the economic significance of ESG drivers versus other firm characteristics?***

The analysis measures the relative predictive power of ESG drivers, in addition to their statistical significance, over sector and financial variables; and interprets the results in terms of economic relevance.

#### ***3.1.2 Methodology***

The analysis follows the structure of a controlled experiment on a simulated panel dataset; consequently, this approach makes it possible to clearly separate the construction of ESG drivers from the subsequent estimation of their relationship with profitability, and thus avoid the confounding effects that occur with real data, such as missing data, measurement errors, or unobserved heterogeneity. The simulation also allows for the systematic manipulation of measurement choices to test their impact on the results. Therefore, the aim is not to estimate a real-world causal ESG-profitability relationship but rather to examine how alternative measurement choices affect empirical inference.

The empirical strategy unfolds in four sequential stages:

##### ***1. Construction of ESG drivers***

Following the taxonomy of Berg et al. (2022)<sup>75</sup>, twelve drivers are constructed: four environmental (Carbon, Water, Energy, Waste), four social (Labor, Diversity, Community, Supply Chain), and four governance (Board, Corruption, Remuneration, Taxes). Since the environmental drivers are already present in the dataset, they are simply normalized, while the remaining drivers are simulated to replicate the correlation patterns found in the literature.

##### ***2. Preprocessing and model estimation***

The dataset is cleaned, standardized and encoded; next, ten models are estimated, ranging from simple linear regression to tree-based models and neural networks. The models are evaluated on a time split, with training from 2016 to 2024 and testing in 2025, to avoid look-ahead bias and in order to identify the most suitable benchmark specification.

##### ***3. Within-sector analysis and divergence analysis***

The analysis then investigates whether ESG effects vary across sectors and conducts controlled experiments on measurement, scope and weighting

---

<sup>75</sup> Ibid.

divergence to assess how methodological choices alter empirical conclusions.

#### ***4. Robustness and Economic Interpretation***

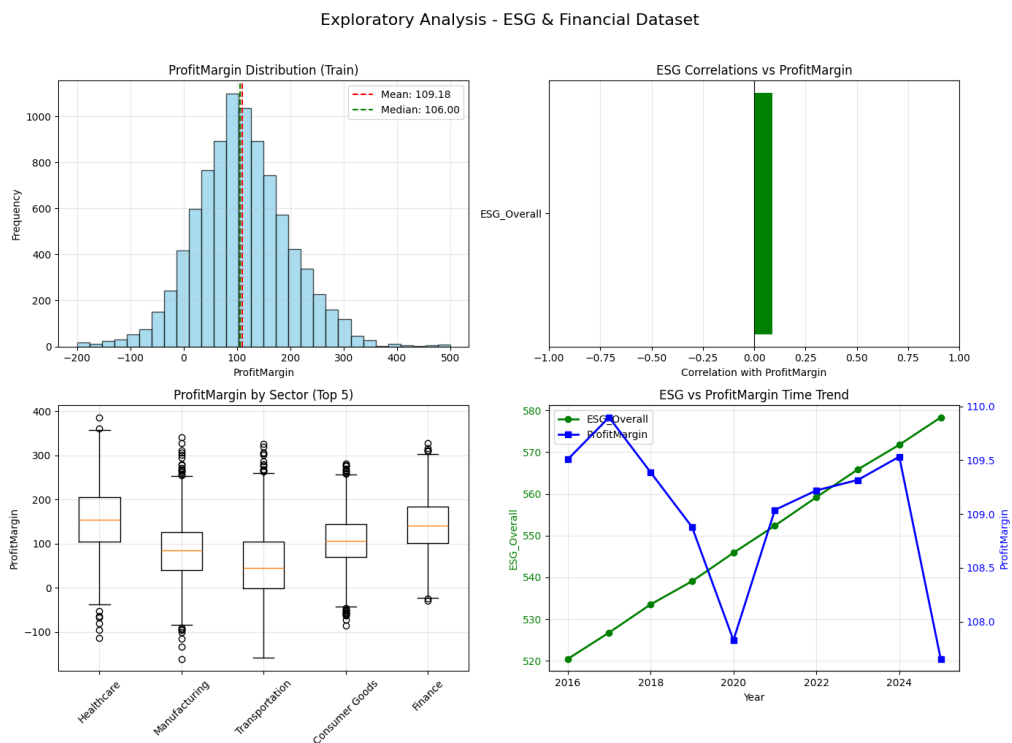
In conclusion, robustness tests are conducted to evaluate the stability of the results and then followed by an economic interpretation of the empirical findings.

### 3.2 Dataset

The analysis starts by considering a simulated dataset containing financial and ESG data at the firm level. Despite the fact that a simulated dataset has its limitations, it offers a significant advantage for this type of analysis, as it allows for precise control over the data structure and enables the isolation of the impact of measurement without interference from missing data or unobservable heterogeneity. The dataset covers a 10-year period (2016–2025) and contains annual observations for 1,000 companies, totaling 10,000 observations. The following data are available for each company:

- **Identifiers:** CompanyID, CompanyName
- **Temporal and sectoral variables:** Year, Industry, Region
- **Financial variables:** Revenue, GrowthRate, MarketCap, ProfitMargin
- **Aggregated ESG scores:** ESG\_overall, ESG\_Environmental, ESG\_Social, ESG\_Governance
- **Specific environmental indicators:** CarbonEmissions, WaterUsage, EnergyConsumption
- **Simulated ESG drivers:** twelve additional variables created specifically for the analysis and named Driver\_E, Driver\_S, Driver\_G

The target variable is ProfitMargin, i.e. a continuous and directly interpretable measure of operating profitability, expressed in percentage terms.



**Figure 1 – Exploratory analysis of the dataset.** Panel (a) shows the distribution of ProfitMargin in the training set, which ranges from -200% to 500% with a mean of 109%. Panel (b) displays the correlations between the aggregate ESG scores and ProfitMargin, which are positive but modest. Panel (c) illustrates the variation in ProfitMargin across the top five sectors, confirming that industry affiliation is a major determinant of profitability. Finally, panel (d) plots the temporal trends of ESG\_Overall and ProfitMargin over the 2016–2025 period, showing a positive co-movement.

### 3.2.1 Data Preparation and Descriptive Statistics

To avoid look-ahead bias, that is using future information during model training, the data has been divided into time periods:

- **Training set:** years 2016–2024 (9.000 observations)
- **Test set:** year 2025 (1.000 observations)

This separation ensures that the model is trained exclusively on historical data and evaluated on future data, thereby enabling simulations to be carried out in a realistic forecasting context.

Furthermore, all pre-processing operations, such as imputation of missing values, scaling and encoding, were performed exclusively on the training set and then applied to the test set, thereby avoiding any form of data leakage.

Afterwards, before proceeding with the actual analysis, the dataset went through some initial cleaning:

- Columns with very high percentages of missing values (over 99%) have been removed, in particular those named ‘Unnamed’, which were automatically generated when the Excel file was saved.
- Variables not relevant to the analysis, such as CompanyID and CompanyName, have been removed.
- Categorical variables (Industry, Region) have been retained for subsequent encoding.

At the end of this phase, the dataset contains 40 columns, 12 of which correspond to the ESG driver variables (nine simulated and three normalized from real environmental indicators).

Finally, from the table below, the descriptive statistics for the main variables are shown, which reveal considerable variability across all variables, particularly for Revenue and GrowthRate, which exhibit very pronounced tails, aspect that will be addressed in the robustness analysis, where the sensitivity of the results to outliers will be assessed.

| <b>VARIABLE</b> | <b>MEAN</b> | <b>DEV. STD</b> | <b>MIN</b> | <b>MAX</b> |
|-----------------|-------------|-----------------|------------|------------|
| ProfitMargin    | 109.18      | 87.63           | -200       | 500        |
| Revenue         | 46,101.53   | 95,849.09       | 0          | 1,000,000  |
| GrowthRate      | 48.55       | 94.40           | -100       | 500        |
| ESG_Overall     | 549.33      | 158.70          | 63         | 988        |

### *3.2.2 Limitations of the Dataset*

The use of a simulated dataset comes with certain limitations that it is important to recognize, including:

- The ESG drivers are synthetic variables: although constructed in accordance with the taxonomy of Berg et al. (2022)<sup>76</sup>, they do not correspond to actual indicators collected in the field; consequently, the results should be interpreted as illustrating mechanisms, not as estimates of real-world relationships.
- Correlation structure: the high correlations between the real environmental drivers (Carbon, Water, Energy) likely reflect the structure of the original dataset, in which these indicators derive from a common factor. This aspect, as realistic as it may be, makes it more difficult to isolate the effect of each driver.
- Absence of causality: the analysis is purely associative; despite the time-series split and controls, the possibility of omitted variables or reverse causality remains unexcluded.

---

<sup>76</sup> Ibid.

### 3.3 Construction of ESG Drivers

The construction of ESG drivers represents the methodological core of this analysis, as well as a critical step; given that the original dataset contains only aggregate scores for the three pillars (Environmental, Social, Governance), it is not possible to directly observe which specific indicators contribute to their formation. Consequently, to overcome this limitation, the aggregate pillars were decomposed into specific drivers, following the taxonomy proposed by Berg, Kölbel and Rigobon (2022)<sup>77</sup> in *Aggregate Confusion*. Specifically, Berg et al. (2022) analyzed 709 indicators from six different ESG rating agencies and classified them into 64 categories (Table IV)<sup>78</sup>. This taxonomy provides a comprehensive overview of the ESG dimensions actually measured by the agencies, divided by pillar; and based on this, twelve drivers were selected, four for each pillar, to construct a more granular and economically interpretable ESG structure.

#### 3.3.1 List of Drivers and Normalization

The selected drivers are the following:

| PILLAR | DRIVER                                           |
|--------|--------------------------------------------------|
| E      | <i>Carbon, Water, Energy, Waste</i>              |
| S      | <i>Labor, Diversity, Community, Supply Chain</i> |
| G      | <i>Board, Corruption, Remuneration, Taxes</i>    |

Of these, three environmental indicators were already provided by the original dataset, namely: CarbonEmissions, WaterUsage and EnergyConsumption. These were normalized on a scale of 0–100 and used directly as drivers for the Environmental pillar. Consequently, to maintain a symmetrical structure across the pillars, it was necessary to simulate:

- A fourth environmental driver (Waste)
- All four social drivers
- All four governance drivers

Furthermore, as the three actual environmental drivers (CarbonEmissions,

---

<sup>77</sup> Ibid.

<sup>78</sup> Ibid., Table IV, pp. 12-13.

WaterUsage and EnergyConsumption) are expressed in different units (tonnes, cubic meters, megawatt-hours), in order to make them comparable with the simulated drivers and with the original ESG scores (0–100 scale), they were normalized using the following formula:

$$\text{Driver\_norm} = 100 \times \left(1 - \frac{\text{value} - \text{min}}{\text{max} - \text{min}}\right)$$

Where:

- ‘value’ is the company’s original data
- Where min and max denote the minimum and maximum of the variable calculated on the training set

In this way, the company with the lowest value (e.g. minimum emissions) receives a score of 100 and the company with the highest value receives a score of 0, whilst all others receive a score between 0 and 100 in linear proportion. It should also be noted that the raw WaterUsage indicator was retained in the final feature set alongside its normalized counterpart Driver\_E\_Water\_norm. Given that Driver\_E\_Water\_norm is a deterministic linear transformation of WaterUsage (as defined in Section 3.3.1), the two variables are perfectly correlated ( $r = -1.00$ ), constituting a redundant feature pair; consequently, their individual feature importances (5.3% and 5.4% respectively) should be interpreted jointly as reflecting a single underlying signal rather than as two independent contributions

### 3.3.2 Simulation of Missing Drivers

Drivers not present in the original dataset were therefore simulated to meet two main requirements:

1. Moderate correlation with the respective ESG pillar (0.45–0.60), so as to reflect the fact that each driver contributes in part to the aggregate score.
2. Low correlation between drivers within the same pillar (0.20–0.40), to replicate the measurement divergence documented in the literature.

The simulation procedure therefore followed these steps:

- **Standardization of the reference ESG pillar** (e.g. ESG\_Social) with a mean of 0 and a standard deviation of 1.
- **Generation of random noise with a standard normal distribution** (mean

0, standard deviation 1).

- **Construction of the driver as a linear combination of the standardized pillar and the noise**, with weights determined by the target correlation:

$$\text{Driver}_{\text{temp}} = \rho \cdot \text{ESG\_std} + \sqrt{1 - \rho^2} \cdot \text{noise}$$

where  $\rho$  is the target correlation.

1. **Transformation to 0–100 scale** with mean 50 and standard deviation 15:

$$\text{Driver}_{\text{final}} = 50 + 15 \cdot \text{Driver}_{\text{temp}}$$

- **Cut-off** at 0 and 100 to avoid out-of-range values.

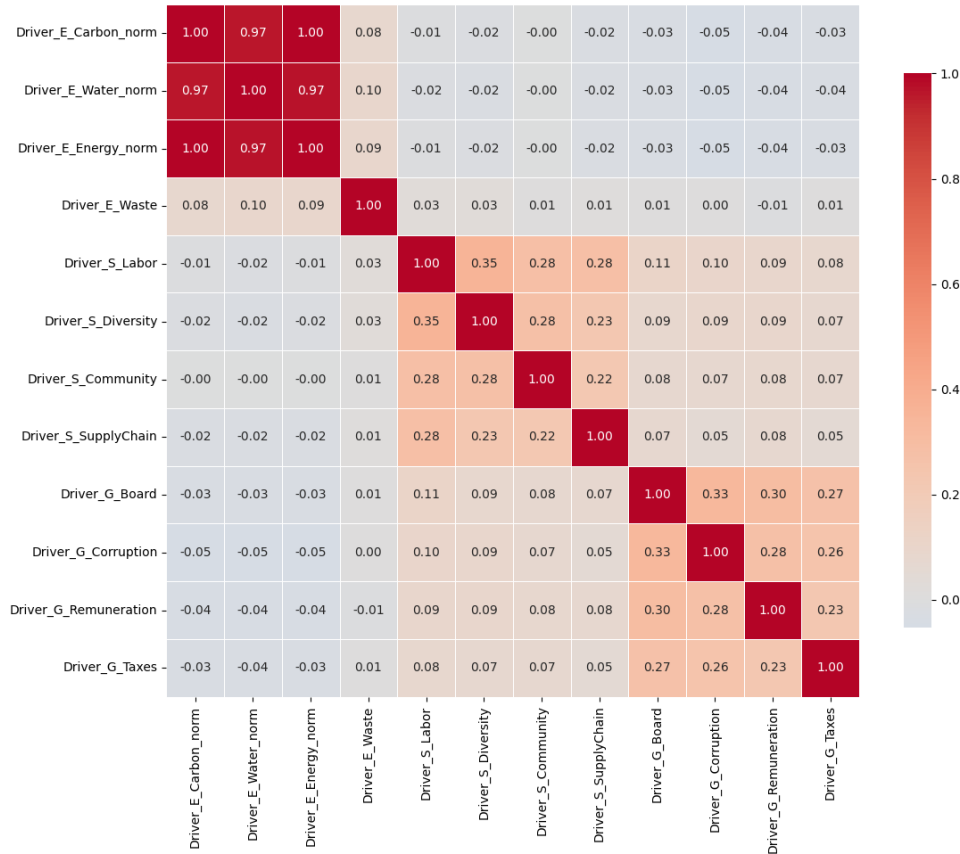
Finally, the following table summarizes the drivers, their source (actual or simulated) and the target correlations with the respective ESG pillar.

| PILLAR | DRIVER       | SOURCE     | TARGET CORRELATION |
|--------|--------------|------------|--------------------|
| E      | Carbon_norm  | Normalised | –                  |
| E      | Water_norm   | Normalised | –                  |
| E      | Energy_norm  | Normalised | –                  |
| E      | Waste_norm   | Simulated  | 0.55               |
| S      | Labor        | Simulated  | 0.60               |
| S      | Diversity    | Simulated  | 0.55               |
| S      | Community    | Simulated  | 0.50               |
| S      | SupplyChain  | Simulated  | 0.45               |
| G      | Board        | Simulated  | 0.60               |
| G      | Corruption   | Simulated  | 0.55               |
| G      | Remuneration | Simulated  | 0.50               |
| G      | Taxes        | Simulated  | 0.45               |

### 3.3.3 Correlation Structure and Validation

From **Figure 2**, we can see the correlation matrix for the twelve drivers; the intra-pillar correlations are as follows:

- **E vs E:** mean 0.53 (with Waste being largely independent)
- **S vs S:** mean 0.28
- **G vs G:** mean 0.28



**Figure 2** – Correlation matrix for the twelve ESG drivers. The intra-pillar correlations are highlighted along the diagonal blocks.

Meanwhile, the correlations between different pillars are close to zero, with mean values ranging from -0.03 to 0.08; this, overall, proves to be a structure consistent with the aims of the simulation and ensures that the drivers are sufficiently independent to allow for the analysis of measurement divergence. In addition, to further verify that the simulated drivers possessed the desired properties, the following correlations were calculated:

- **With the reference ESG pillar:** the values obtained ranged between 0.44 and 0.61, in line with the objectives.
- **Between drivers within the same pillar:** the correlations ranged between 0.22 and 0.35, faithfully replicating the measurement divergence observed in the literature.
- **Between drivers from different pillars:** the correlations were close to zero, which made sure the independence required for the analysis.

### 3.4 Preprocessing and Models

At this stage, before proceeding with model estimation, the data underwent a meticulous preprocessing phase, with the aim, on the one hand, of ensuring the quality and comparability of the variables and, on the other, of avoiding any form of data leakage that might compromise the validity of the results. Moreover, there were no missing values for the main variables in the dataset; consequently, the few columns with a high percentage of missing values (the 'Unnamed' columns generated when saving the Excel file) had already been removed during the initial cleaning phase. Therefore, it was not necessary to apply imputation.

Afterwards, the categorical variables Industry and Region were transformed using one-hot encoding; this created dummy variables for each category, allowing the models to utilize sectoral and geographical information without imposing an arbitrary order; meanwhile, to avoid perfect multicollinearity, the first category for each variable was omitted (*drop='first'*). As a result, the process generated:

- 8 dummy variables for the sector (out of 9 original sectors)
- 6 dummy variables for the region (out of 7 original regions)

Regarding all numerical variables (ESG drivers, financial variables), these were standardized using z-score normalisation:

$$z = \frac{x - \mu}{\sigma}$$

where  $\mu$  is the mean and  $\sigma$  is the standard deviation calculated on the training set. Standardization was applied to ensure that all variables contributed equally to the models, preventing variables with different scales (e.g. revenue in the thousands) from dominating those with smaller scales (e.g. ESG scores of 0–100).

Finally, the standardization parameters were calculated exclusively on the training set and then applied to the test set, in order to preserve the separation between the two sets.

#### 3.4.1 Feature Engineering: Interactions

In order to capture potential synergies between ESG and financial variables, interactions were created between the ESG drivers and the Revenue and GrowthRate variables. Specifically:

- ESG\_Overall\_x\_Revenue
- ESG\_Overall\_x\_GrowthRate
- WaterUsage\_x\_Revenue
- WaterUsage\_x\_GrowthRate

These interactions were computed and tested in a separate exploratory specification to examine whether the effect of ESG on profitability varies with firm size (Revenue) or growth rate (GrowthRate); however, they were not included in the final benchmark models (30 features) used for the results reported in Sections 3.5–3.11.

### 3.4.2 Models Estimated

At this point, ten models were estimated and selected to cover different algorithmic families and provide a comprehensive overview of performance:

| FAMILY                       | MODEL                                           |
|------------------------------|-------------------------------------------------|
| <b>Linear</b>                | Linear Regression, Ridge (L2), Lasso (L1)       |
| <b>Parametric non-linear</b> | Polynomial Regression (degree 2), GAM           |
| <b>Tree-based</b>            | Decision Tree, Random Forest, Gradient Boosting |
| <b>Kernel</b>                | Support Vector Regression (SVR)                 |
| <b>Neural network</b>        | Deep Learning (4 layer)                         |

For models that require hyperparameter tuning (Ridge, Lasso, Random Forest, Gradient Boosting), the parameters were optimized using grid search with 5-fold cross-validation on the training set, and the optimization metric was  $R^2$ . Specifically: for Ridge and Lasso, the tested alpha range was {0.001, 0.01, 0.1, 1, 10, 100}; for Random Forest and Gradient Boosting, combinations of `n_estimators`, `max_depth`, `learning_rate` and other parameters were tested; whilst the neural network was trained using early stopping (patience of 10 epochs) to prevent overfitting.

Furthermore, all models were evaluated using regression metrics:

- **$R^2$** : coefficient of determination, a measure of explained variance
- **MSE**: mean squared error
- **MAE**: mean absolute error

The evaluation protocol was strictly temporal:

- **Training:** 2016–2024 (9.000 observations)
- **Test:** 2025 (1.000 observations)

In this way, this division simulates a realistic context in which the model is trained on past data and evaluated on future data, avoiding look-ahead bias.

By the end of the preprocessing phase, the final dataset consisted of:

- **30 features:** 16 standardized numerical features (12 ESG driver variables, two financial variables, Revenue and GrowthRate, the aggregate ESG\_Overall score, and the raw WaterUsage indicator)
- **9000 observations** in the training set
- **1000 observations** in the test set

The following table summarizes the dataset dimensions at each stage.

| Stage                      | Number of features | Number of observations (train) |
|----------------------------|--------------------|--------------------------------|
| After initial cleaning     | 40                 | 9000                           |
| After variable selection   | 18                 | 9000                           |
| After encoding and scaling | 30                 | 9000                           |

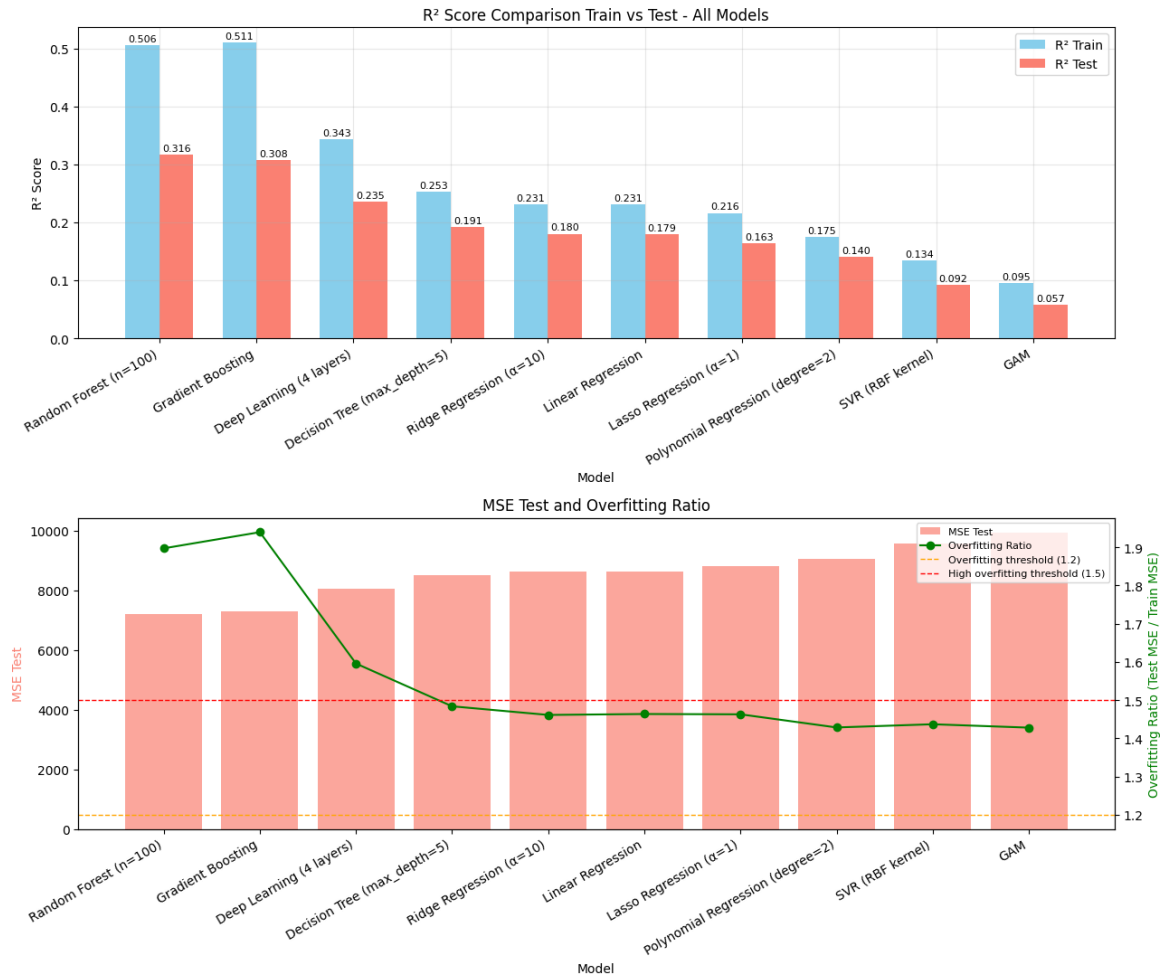
### 3.5 Model Performance

This section presents the results obtained by estimating the ten models and comparing their performance in terms of  $R^2$ , MSE and overfitting. The main objective is, therefore, to identify the model with the best predictive ability and to assess the robustness of the estimates.

The following table shows the performance of the ten models on the test set, sorted in descending order of  $R^2$ .

| <b>Model</b>       | <b><math>R^2</math> Test</b> | <b>MSE Test</b> | <b>Overfitting Ratio (Test MSE / Train MSE)</b> |
|--------------------|------------------------------|-----------------|-------------------------------------------------|
| Random Forest      | 0.316                        | 7199.2          | 1.90                                            |
| Gradient Boosting  | 0.308                        | 7287.7          | 1.94                                            |
| Deep Learning      | 0.235                        | 8051.6          | 1.60                                            |
| Decision Tree      | 0.191                        | 8512.8          | 1.48                                            |
| Ridge              | 0.180                        | 8629.2          | 1.46                                            |
| Linear Regression  | 0.179                        | 8638.8          | 1.46                                            |
| Lasso              | 0.163                        | 8806.0          | 1.46                                            |
| Polynomial (deg 2) | 0.140                        | 9050.8          | 1.43                                            |
| SVR                | 0.092                        | 9556.0          | 1.44                                            |
| GAM                | 0.057                        | 9921.1          | 1.43                                            |

The figure below provides a graphical comparison of the training and test  $R^2$  values and the overfitting ratio for all models.



**Figure 3** – Top: comparison of training  $R^2$  and test  $R^2$  for all models. Bottom: Test MSE and Overfitting Ratio (Test MSE / Training MSE).

It can be seen that linear models yield  $R^2$  values of around 0.18, whilst tree-based models perform better. Random Forest and Gradient Boosting achieve  $R^2$  values of 0.316 and 0.308 respectively. Deep Learning ranks third with an  $R^2$  of 0.235, whilst SVR and GAM record the lowest performance.

Furthermore, as a benchmark, a naive model was considered which predicts the mean ProfitMargin from the training set for all observations; this model produces an  $R^2$  of  $-0.0002$ , indicating that it performs worse than a simple mean, and it is the Random Forest that significantly improves upon this baseline, explaining approximately 31.6% of the variability in ProfitMargin on the test set.

As for overfitting, this was assessed by calculating the ratio of test MSE to training MSE, where values above 1.2 generally indicate a certain degree of overfitting; however, for complex models on simulated data, levels up to 1.5–2.0 are considered acceptable. This shows that all models have a certain degree of overfitting, with Random Forest and Gradient Boosting showing the highest ratios (1.90 and 1.94); this is to be expected, as tree-based models are more

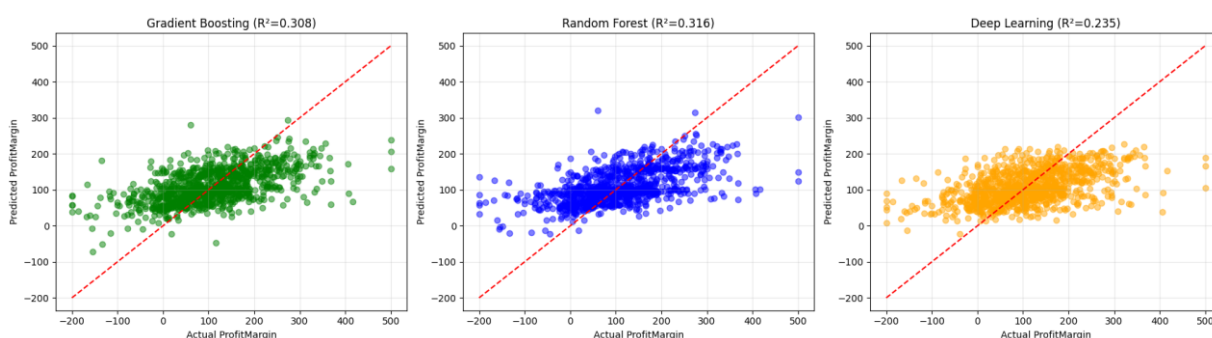
flexible and tend to capture some of the noise in the data. Linear models, on the other hand, being more rigid, exhibit lower levels of overfitting (1.46–1.48).

Finally, because of these results, Random Forest was selected as the benchmark model for the subsequent analyses, a choice which is motivated primarily by:

- Having the highest  $R^2$  on the test set (0.316);
- Its ability to capture non-linear relationships;
- And lastly, the ability to extract feature importance, which is useful for interpreting the role of ESG drivers.

### 3.5.1 Actual vs Predicted and Residual Analysis

The scatter plots in Figure 4 show the actual values plotted against the predicted values for the three best-performing models: Random Forest, Gradient Boosting and Deep Learning. It can be seen that for Random Forest and Gradient Boosting, the data points cluster along the diagonal, suggesting good predictive ability, despite moderate dispersion. Deep Learning, on the other hand, shows greater dispersion, particularly for extreme values of ProfitMargin (below -100% or above 300%), suggesting lower reliability at the tails of the distribution. As for the correlation between the predictions of Random Forest and Gradient Boosting, this is 0.933, indicating that the two models converge on very similar forecasts.



**Figure 4** – Scatter plot of actual versus predicted values for Random Forest, Gradient Boosting and Deep Learning.

This is followed by a residual analysis for the Random Forest model, i.e. the difference between the actual and predicted values. As it can be seen in Figure 5, the residuals exhibit the following characteristics:

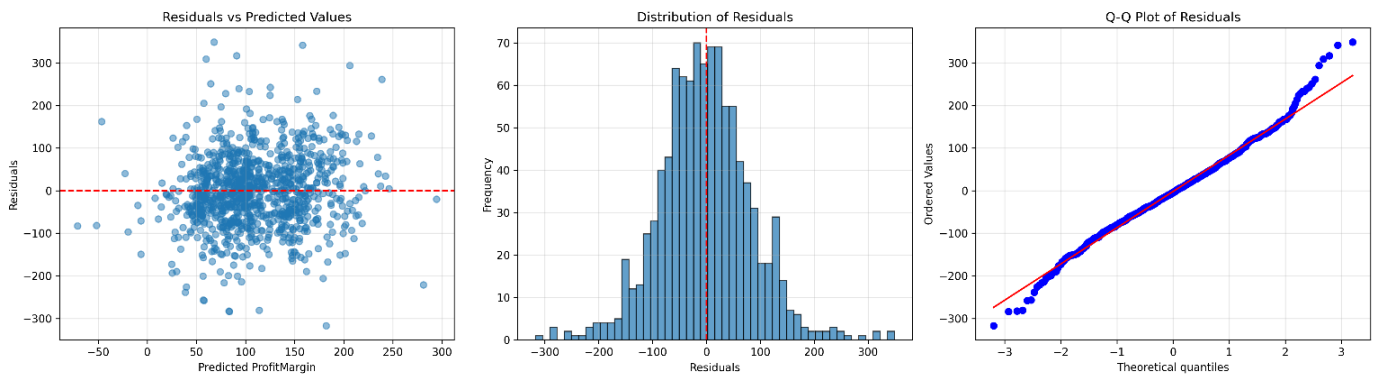
- Mean close to zero (-1.15), which indicates that the model is not affected

by systematic bias;

- Standard deviation of 84.88, reflecting the wide variability of the ProfitMargin;
- Skewness of 0.165 and kurtosis of 1.55.

An analysis of the Shapiro-Wilk test yields a p-value of less than 0.001, indicating that the residuals are not normally distributed; this is an expected result for financial data, as they often exhibit heavier tails than a Gaussian distribution.

As shown in Figure 5, the residuals-versus-predicted plot displays a relatively constant dispersion, with no obvious funnel-shaped patterns or systematic curvature, suggesting that heteroscedasticity (non-constant variance of the residuals) is not a significant issue; whilst the Q-Q plot confirms the presence of heavier tails compared to a normal distribution, consistent with the presence of outliers in the ProfitMargin.



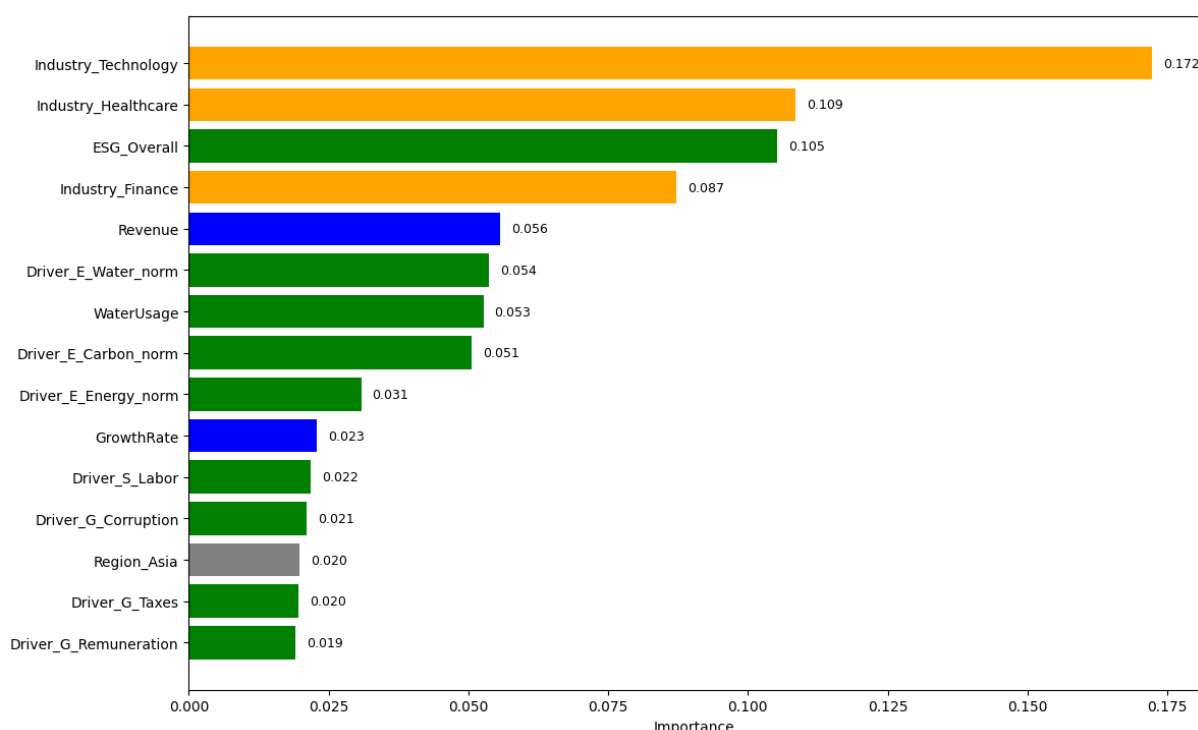
**Figure 5** – Residual analysis for the Random Forest model: (a) residuals versus predicted values, (b) residual histogram, (c) Q-Q plot.

Overall, the residual analysis confirms that the Random Forest model is well-specified and does not exhibit systematic biases, although the non-normality of the residuals, which is common in financial contexts, suggests caution in inferential interpretation.

### 3.6 Feature Importance

One of the advantages of tree-based models is the ability to extract feature importance, i.e. a measure of each variable's contribution to the model's predictive power. This section analyses feature importance in the Random Forest model, which proved to be the best-performing model, in order to understand which ESG drivers, financial variables and sector controls have the greatest influence on profitability.

Figure 6 shows the top 15 features' importance in the Random Forest, and the results highlight a clear predominance of sectoral variables; indeed, the sector dummies occupy the top three positions.



**Figure 6** – Top 15 feature importance in Random Forest

Consequently, the overall importance of the twelve ESG drivers (shown in green in the chart) stands at around 31%, whilst industry sector dummies (orange) account for approximately 38.5%, region dummies for 6.9%, and the two financial variables, Revenue and GrowthRate, for around 7.9%; the remaining share is attributable to the aggregate ESG\_Overall score (10.5%) and the raw WaterUsage indicator (5.3%). Analyzing the aggregate importance by pillar, environmental (E) drivers contribute 15.4%, social (S) drivers 7.8% and governance (G) drivers 7.8%; this estimate will be tested in Section 3.6.1 using methods that are more robust to multicollinearity between variables.

Furthermore, among the individual drivers, environmental drivers stand out in particular:

Driver\_E\_Water\_norm (5,4%), Driver\_E\_Carbon\_norm (5,1%) and Driver\_E\_Energy\_norm (3.1%); followed by the Labor driver (2.2%) and the Corruption driver (2.1%). As for social drivers, they show lower importance (all < 2.5%), in line with the results of the subsequent within-sector analysis; indeed, the dominance of sector dummies confirms that sector affiliation is the main determinant of profitability in the model, which is not surprising given the wide variability in profit margins across sectors observed in the exploratory analysis. Meanwhile, the importance of environmental drivers (15.4%) suggests that, within the same sector, differences in terms of operational efficiency (water, energy, emissions) have a significant impact on profitability. Finally, the significance of the Corruption driver suggests that the quality of governance, particularly transparency and risk management, is a key factor; conversely, the lower significance of social drivers (7.8%) may be interpreted as an indication of the lesser economic materiality of these dimensions, or as a reflection of the greater difficulty in measuring them in a comparable manner. Conversely, it is observed that linear models have produced unstable coefficients and conflicting signs among ESG factors, which is primarily due to the high multicollinearity among environmental indicators. Whereas decision-tree models naturally mitigate this problem, thereby making the importance of the variables a more reliable measure of economic relevance than the mere significance of the coefficients. In this context, ESG emerges as a significant but complementary source of predictive information, rather than as the primary determinant of profitability.

### *3.6.1 Permutation Importance*

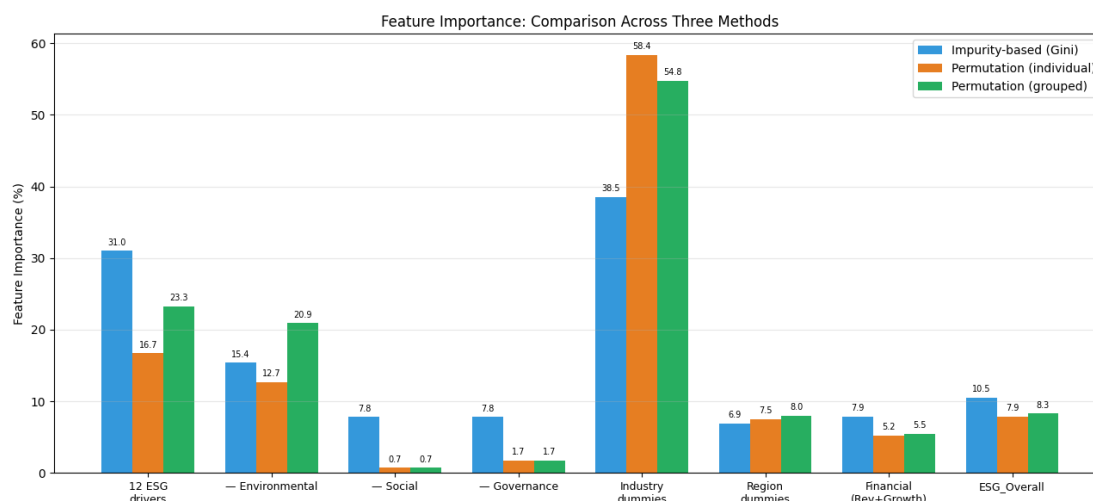
The results described above are based on the Random Forest's default importance measure, which is the importance based on impurity reduction (Gini importance); however, before drawing definitive conclusions on the relative weight of ESG drivers compared with sectoral variables, it is recommended to consider the extent to which this estimate is sensitive to the calculation method used, since it is well known that impurity-based importance can be distorted in the presence of correlated predictors; this is a significant consideration in this case, given that the raw WaterUsage indicator and its normalised version Driver\_E\_Water\_norm are, by construction, perfectly correlated ( $r = -1.00$ ). Therefore, this part of the work tests the 31% estimate obtained previously by comparing it with two alternative measures of feature importance. The permutation importance, which measures the reduction in  $R^2$  when the values of a feature are randomly permuted, confirms the dominance

of sector dummies but significantly reduces the estimated contribution of ESG drivers, from 31.0% (impurity-based) to 16.9% (individual permutation) and this discrepancy reflects a well-known limitation regarding the importance of individual permutations; in particular, when a feature is permuted, the model can partially reconstruct the lost signal by drawing on other correlated features that remain unchanged, which leads to an underestimation of the importance of each individual variable.

In this regard, to correct for this effect, a group permutation importance has been calculated, in which the correlated features (the four environmental indicators, including WaterUsage and its normalised counterpart) are permuted jointly. This yields a total ESG importance of 23.3%, falling between the individual permutation estimate and the impurity-based estimate, with the environmental pillar at 20.9%, governance at 1.7% and the social pillar at 0.7%. It can be seen, with this adjustment, that the sector dummies account for 54.8% of the total importance, which is an even higher proportion than that of the impurity-based measure (38.5%).

The following table and Figure 6bis summarise the three measures in comparison.

| <b>Category</b>        | <b>Impurity-based</b> | <b>Permutation (individual)</b> | <b>Permutation (group)</b> |
|------------------------|-----------------------|---------------------------------|----------------------------|
| 12 drivers ESG         | 31,0%                 | 16,9%                           | 23,3%                      |
| Environmental          | 15,4%                 | 14,9%                           | 20,9%                      |
| Social                 | 7,8%                  | 0,4%                            | 0,7%                       |
| Governance             | 7,8%                  | 1,6%                            | 1,7%                       |
| Sectoral Dummy         | 38,5%                 | 58,4%                           | 54,8%                      |
| Regional Dummy         | 6,9%                  | 7,5%                            | 8,0%                       |
| Financial (Rev+Growth) | 7,9%                  | 5,2%                            | 5,5%                       |
| ESG_Overall            | 10,5%                 | 7,9%                            | 8,3%                       |



**Figure 6bis** – Comparison of feature importance across three methods: impurity-based (Gini), individual permutation and group permutation

It is interesting to note that this discrepancy between methods of measuring importance mirrors, at the level of model interpretation, the same phenomenon documented at the level of data measurement in Section 3.8; namely the conclusion that ‘the sector takes precedence over ESG’ is robust across all three specifications, but the precise magnitude of the ESG contribution is itself sensitive to methodological choices; in this case, the choice of the importance metric rather than the ESG indicator. This therefore reinforces the central argument of the thesis that measurement divergence, understood in a broad sense, extends beyond the definition of ESG concepts to include the way in which their empirical significance is assessed.

### 3.7 Within-Sector Analysis

The previous analysis showed that sector dummies dominate the feature importance of the Random Forest, which suggests that the sector to which a company belongs is the main determinant of profitability. Consequently, to understand whether and how the effect of ESG varies within individual sectors, a within-sector analysis was conducted. The objective here is dual: to isolate the contribution of ESG drivers net of sectoral differences, and to identify any sector-specific patterns of materiality.

#### 3.7.1 Methodology and Results

The following analysis was conducted on the three sectors with the highest number of observations: Technology (855 observations), Healthcare (1,089 observations) and Finance (1,017 observations). For each sector, separate models were estimated using Random Forest, which proved to be the model with the best overall performance; therefore, sector dummies were removed to avoid statistical leakage (within a single sector, these variables would be constant and provide spurious information). Furthermore, the importance of ESG drivers was calculated as a percentage of the total importance of features in the sectoral model; in this way, it is possible to assess the predictive contribution of ESG when companies are compared with others in the same sector. The results of the analysis are summarized in the following table:

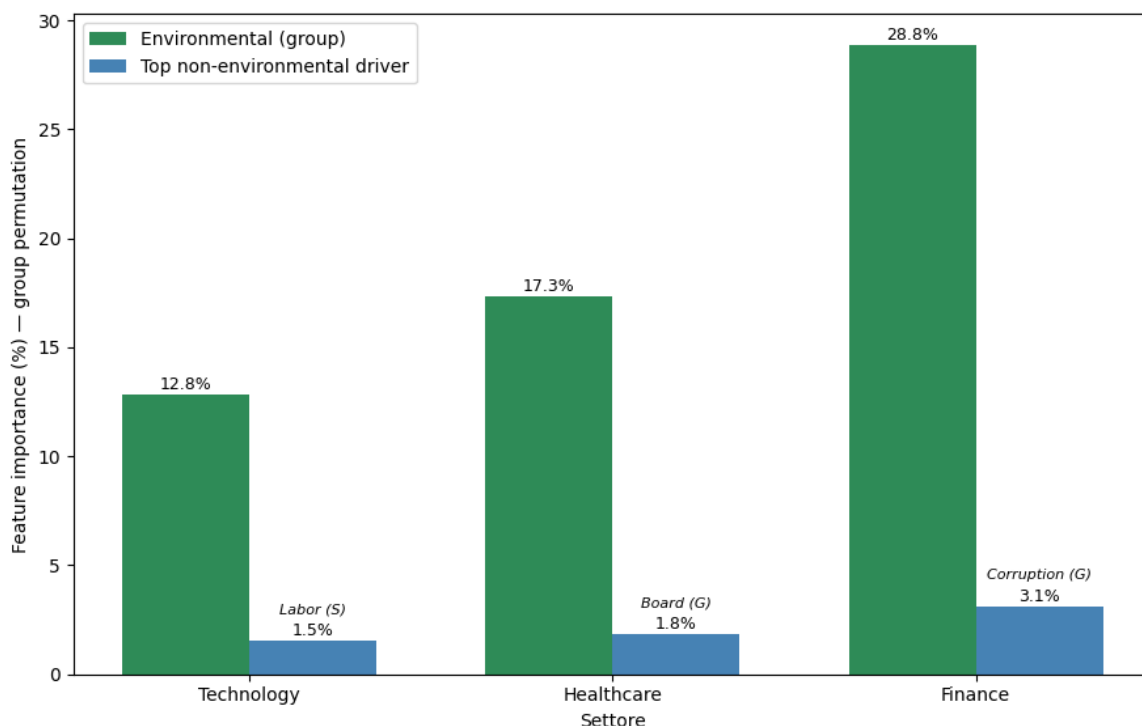
| Sector     | ESG importance | Top driver ESG                                                            |
|------------|----------------|---------------------------------------------------------------------------|
| Technology | 20.4%          | Environmental (group), Labor (S), Corruption (G), Board (G): closely tied |
| Healthcare | 28.4%          | Environmental (group), Board (G)                                          |
| Finance    | 36.1%          | Environmental (group), Corruption (G)                                     |

Thus, the importance of ESG factors ranges between 20% and 36%, confirming that even within individual sectors, ESG drivers make a significant contribution to predicting profitability; however, distinct patterns emerge:

- **Technology:** the environmental group is the largest contributor, while among individual drivers Labor (S), Corruption (G) and Board (G) contribute comparably, with no single dominant non-environmental driver.
- **Healthcare:** the environmental group dominates again; among individual

drivers, Board (G) emerges as the most relevant, ahead of Remuneration (G) and SupplyChain (S).

- **Finance:** the Corruption (G) driver remains clearly the most important individual driver next to a very large contribution from the environmental group as a whole. Hence, in finance, transparency and the management of legal risks are critical factors and in turn, the importance of environmental drivers are likely to reflect the growing focus on climate risks within the financial sector.



**Figure 7** – Group permutation importance by sector (Technology, Healthcare, Finance). The environmental group (Carbon, Water, Energy and WaterUsage, permuted jointly to account for their strong mutual correlation) is compared against the single most important non-environmental driver in each sector. Importance scores are calculated on the test set (2025), averaged over 5 random seeds.

These results invite an economic reading on two distinct levels, being, firstly the magnitude of the environmental group's contribution varies substantially across sectors, ranging from 12.8% in Technology to 28.8% in Finance, which is consistent with the differing resource intensity and regulatory exposure of these industries; indeed, financial institutions are increasingly assessed on their exposure to climate-related and transition risk through their lending and investment portfolios (financed emissions), whereas software- and service-based technology firms have comparatively limited direct environmental footprints, so that operational environmental indicators carry less discriminating information for profitability within that sector. Regarding Healthcare, it falls between the two, reflecting the sector's meaningful but less capital-intensive environmental exposure (e.g. energy use in clinical facilities)

relative to Finance. Secondly, the identity of the leading non-environmental driver differs meaningfully across sectors and is informative in its own right; indeed, in Finance, the clear dominance of Corruption is consistent with the centrality of institutional integrity and internal controls to firm-level risk in a heavily regulated, trust-dependent industry, while, in Healthcare, the prominence of Board suggests that governance oversight, for instance concerning clinical risk management and regulatory compliance, is the most material non-environmental factor. Moreover, in Technology, by contrast, no single driver dominates, as both Labor, Corruption and Board contribute comparably, which is consistent with a sector where profitability is shaped simultaneously by competition for talent, reputational sensitivity to internal conduct and investor scrutiny of governance structures, without any one of these factors being unambiguously more material than the others.

### *3.7.2 Implications for Materiality and Comparison with the Aggregated Model*

The results suggest that ESG materiality varies significantly across sectors; in particular, the environmental group (jointly permuted to account for multicollinearity) consistently emerges as the largest ESG contributor across all three sectors, though its magnitude varies substantially (12.8% in Technology vs 28.8% in Finance). Conversely, non-environmental drivers show more sector-specific patterns: Corruption (G) is clearly dominant in Finance, Board (G) is the leading individual driver in Healthcare, while in Technology no single non-environmental driver dominates. Thus, the within-sector analysis confirms that aggregating different sectors risks obscuring important relationships. Indeed, the patterns that emerged are consistent with the literature on ESG materiality and with the framework of the Sustainability Accounting Standards Board (SASB), which identifies financially material ESG issues on an industry-by-industry basis through its Materiality Finder<sup>79</sup>. It can be observed that, whilst in the aggregated analysis (all sectors combined), the overall ESG importance was 31%, with environmental drivers dominating (15.4% of the total), the within-sector analysis reveals that this aggregated result is partly determined by the strong presence of environmental drivers in Healthcare and Finance, whilst governance and social drivers emerge in Technology; therefore, aggregation tends to smooth out sectoral differences, i.e. by giving greater weight to drivers that are relevant to the sectors most numerous in the dataset.

---

<sup>79</sup> Sustainability Accounting Standards Board (SASB), Materiality Finder.

### *3.7.3 Economic Interpretation*

This variation in the importance of ESG across sectors can be interpreted in economic terms:

- In the technology sector, the absence of a single, non-environmental determining factor is in itself economically significant, as it suggests that no single governance or social practice is, on its own, sufficient to explain differences in profitability between technology firms; rather, it is a combination of factors, such as talent retention, internal conduct and governance structure, that collectively determine firm-level outcomes; which is essentially consistent with the sector's reliance on human capital and intangible assets, as risk is spread across multiple organizational dimensions rather than concentrated in a single area.
- In the healthcare sector, the leading role of Board complements the substantial contribution made by the environmental group; indeed, effective oversight by the Board is plausibly linked to factors such as a hospital's or healthcare provider's ability to manage clinical risk, regulatory compliance and resource allocation; that is to say, all factors with direct financial implications in a heavily regulated sector. Therefore, this suggests that, in the healthcare sector, environmental efficiency and the quality of governance act as complementary rather than competing sources of ESG-related information.
- In Finance, instead, the clear predominance of corruption over other non-environmental factors reinforces the interpretation that anti-corruption measures serve as an indicator of institutional strength and the quality of internal controls; i.e. factors that are particularly critical in a sector where legal and reputational risk is fundamental to corporate value. Furthermore, the very high contribution from the environmental group (28.8%) further indicates that the financial sector is increasingly being assessed through the lens of climate and transition risks inherent in its lending and investment activities, and is therefore in line with the broader materiality frameworks documented in the ESG literature (for example, the SASB Materiality Finder).

### 3.8 Divergence Experiments

This section presents a series of controlled experiments designed to test the impact of measurement divergence on empirical results. Following the framework of Berg, Kölbel and Rigobon (2022)<sup>80</sup>, three types of divergence were analyzed: measurement, i.e. different ways of measuring the same concept; scope, relating to different sets of attributes; and, finally, weight, concerning different aggregation weights. In addition to these, an experiment on false positives was conducted to verify the effect of sectoral confounding.

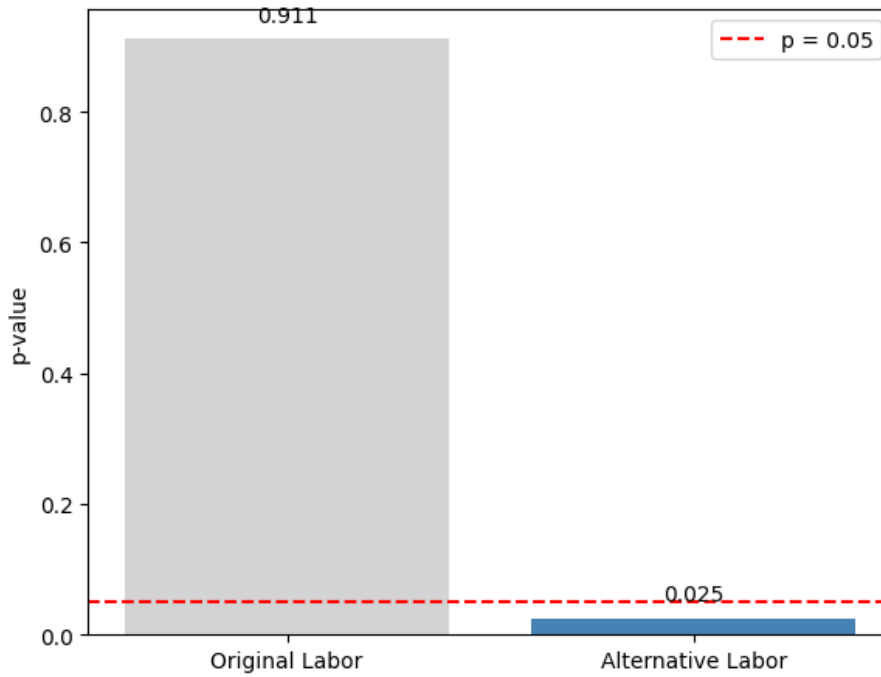
#### *3.8.1 Measurement Divergence: The Case of Labor*

In order to test the extent to which the measurement might influence the results, the first experiment conducted focuses on the `Driver_S_Labor` driver. It had previously been observed that, in the original linear model, this driver was not statistically significant, with a p-value of 0.91; therefore, an alternative version was created by adding Gaussian noise to maintain a correlation of approximately 0.5 with the original version, and both versions were estimated using simple linear models. As a result, the alternative version produced a p-value of 0.025, i.e. significant at the 5% level. Thus, the same ESG dimension, namely labour practices, measured in a slightly different way completely changes the statistical conclusion, from non-significant to significant, demonstrating that measurement divergence can therefore create false positives: an apparent effect that disappears or appears depending on how the underlying concept is measured.

The figure below shows a comparison of the p-values for the original and alternative Labour drivers, where the red dotted line indicates the significance threshold ( $p = 0.05$ ).

---

<sup>80</sup> Berg et al. (2022), cit.



**Figure 8** – Measurement divergence: p-value Labor original vs alternative

### 3.8.2 Scope Divergence: Excluding Environmental Drivers

The second experiment conducted relates to scope divergence, which examines what happens when an entire group of drivers is removed. Starting from estimating a Random Forest model without the four environmental drivers (Carbon, Water, Energy, Waste), and comparing its performance with those of the complete model. As a result, it was observed that, whilst the complete model achieves an  $R^2$  of 0.316, without the environmental drivers, the  $R^2$  drops to 0.306, representing a loss of 3.2%. Consequently, the effect in question is moderate; that is, environmental drivers contribute to predictive power but are not the dominant factor. It can therefore be concluded that excluding an entire pillar reduces performance, but to a limited extent, suggesting that the ESG signal is distributed across multiple dimensions.

### 3.8.3 Weight Divergence: Artificial Ratings with different weights

Subsequently, in order to test the impact of weight divergence, four artificial ratings were constructed by linearly combining the drivers with different weights:

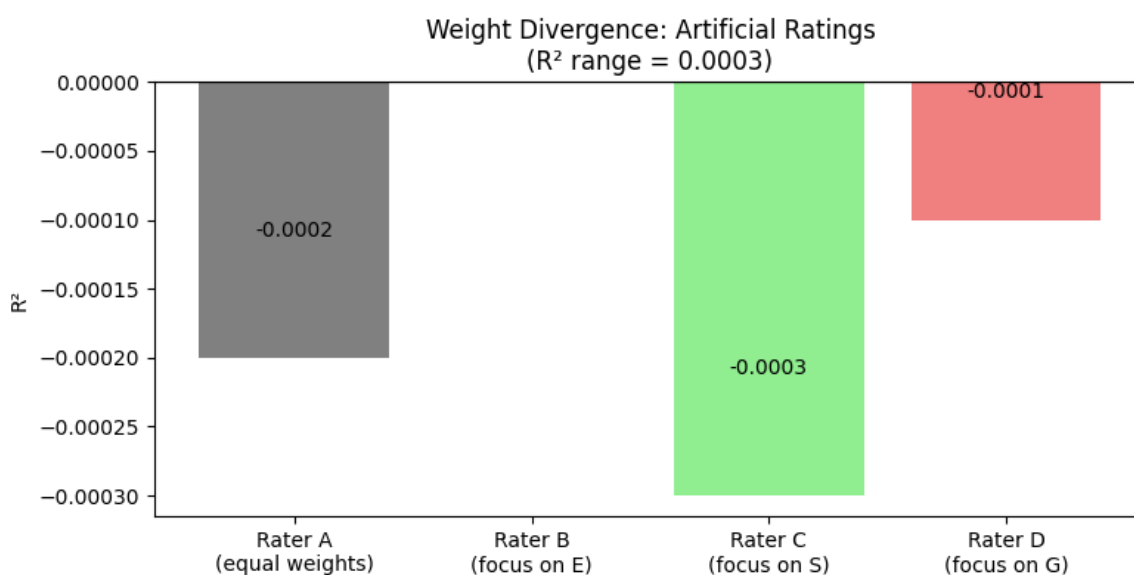
- **Rater A:** balanced weights (all drivers given equal weight)
- **Rater B:** emphasis on E (50% on environmental drivers, 25% on S and G)

- **Rater C:** emphasis on S
- **Rater D:** emphasis on G

For each rating, a linear model was estimated to predict the ProfitMargin, and the results show almost identical performance:

| <b>Rater</b> | <b>R<sup>2</sup> Test</b> |
|--------------|---------------------------|
| Rater A      | -0.0002                   |
| Rater B      | 0.0000                    |
| Rater C      | -0.0003                   |
| Rater D      | -0.0001                   |

The range of R<sup>2</sup> is just 0.0003, and is thus a negligible difference. This suggests, then, that the choice of weights has a very limited impact on predictive power, consistent with Berg et al. (2022)<sup>81</sup>, who attribute only 6% of the total divergence between ratings to weight divergence.



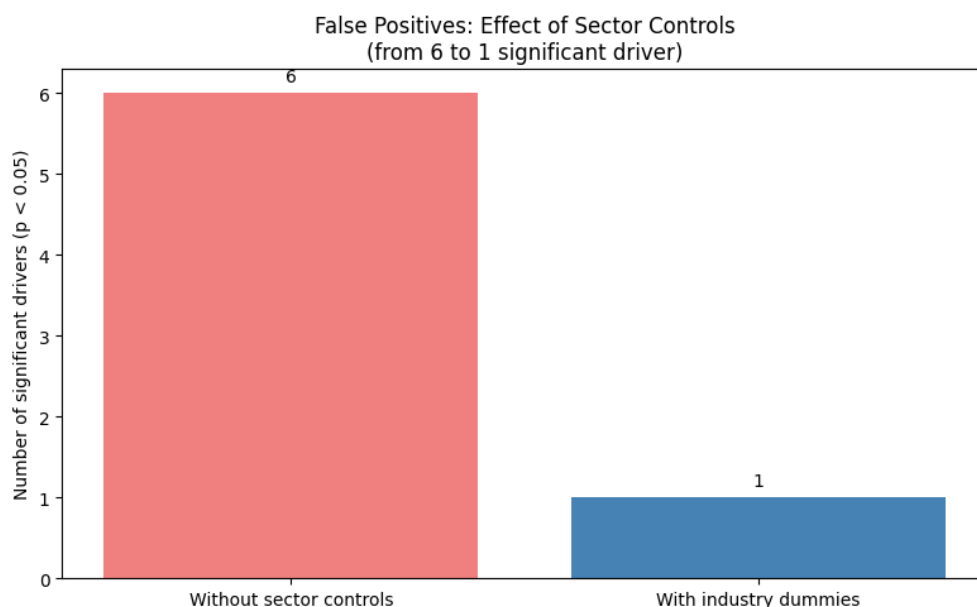
**Figure 9** – Weight divergence: R<sup>2</sup> of the four artificial ratings

### 3.8.4 False Positive: Sector Confounding

Subsequently, a linear model including only the drivers (without controls) was compared with a model that also adds industry dummies, in order to verify the effect of sectoral confounding. It can be seen that in the model without controls, 6 out of 12 drivers are statistically significant, whereas when sector

<sup>81</sup> Ibid., p.1324.

dummies are introduced, the number of significant drivers drops to 1 (only Driver\_E\_Waste). In other words, 5 drivers that appeared to be correlated with profitability were in fact false positives, and therefore their correlation was mediated by the sector to which they belonged.



**Figure 10** – Number of significant drivers ( $p < 0.05$ ) in the model without sector controls (6) and in the model with industry dummies (1).

The following table shows the p-values in the two models for the drivers that lose statistical significance.

| Driver    | p-value without Industry | p-value with Industry |
|-----------|--------------------------|-----------------------|
| Carbon    | 0,000                    | 0,644                 |
| Water     | 0,000                    | 0,532                 |
| Energy    | 0,000                    | 0,484                 |
| Diversity | 0,018                    | 0,084                 |
| Taxes     | 0,009                    | 0,138                 |

This result highlights the need to include sectoral controls in ESG analyses, as without them there is a risk of attributing effects to ESG that actually depend on the industrial structure in which companies operate.

Subsequently, to verify that the six drivers found to be significant in the model without sectoral controls were not an artefact of multiple testing (12 drivers tested simultaneously), a Benjamini-Hochberg (FDR) correction was applied

with a threshold of  $\alpha = 0.05$ . Hence, it can be seen that all six drivers remain significant even after the correction (see table below), confirming that their initial significance is not due to statistical chance linked to the number of tests carried out. Consequently, this reinforces the interpretation that the loss of significance of five of these six drivers, once sector-specific controls were introduced, reflects a genuine sector-specific confounding effect rather than a Type I error due to multiple testing.

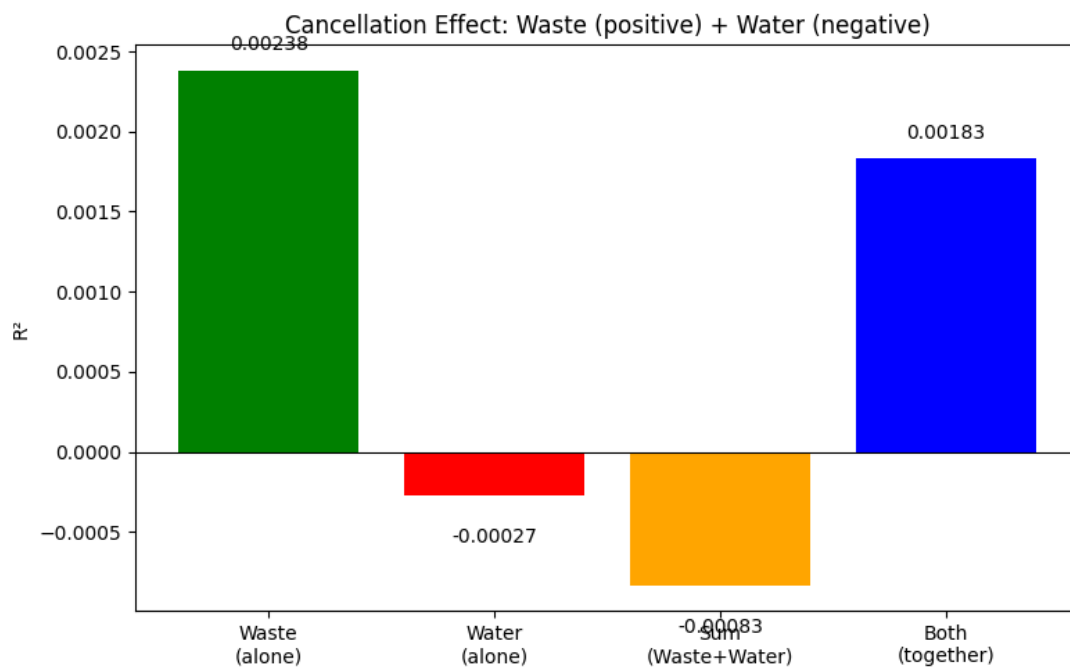
| <b>Driver</b> | <b>p-value</b> | <b>p-value FDR-correct</b> | <b>Significant after FDR</b> |
|---------------|----------------|----------------------------|------------------------------|
| Waste         | <0,001         | <0,001                     | Yes                          |
| Energy        | <0,001         | <0,001                     | Yes                          |
| Carbon        | <0,001         | <0,001                     | Yes                          |
| Water         | <0,001         | <0,001                     | Yes                          |
| Taxes         | 0,009          | 0,021                      | Yes                          |
| Diversity     | 0,018          | 0,037                      | Yes                          |
| Labor         | 0,218          | 0,374                      | No                           |
| Community     | 0,286          | 0,428                      | No                           |
| SupplyChain   | 0,347          | 0,462                      | No                           |
| Corruption    | 0,716          | 0,860                      | No                           |
| Remuneration  | 0,828          | 0,903                      | No                           |
| Board         | 0,912          | 0,912                      | No                           |

### 3.8.5 Cancellation

Finally, a cancellation experiment is carried out to test what happens when two drivers with opposite signs are combined. In this regard, Driver\_E\_Waste (positive coefficient) and Driver\_E\_Water\_norm (negative coefficient) were taken into account. It can be seen that, on their own, the two drivers have  $R^2$  values of 0.0024 and -0.0003 respectively, whilst their sum produces an  $R^2$  of -0.0008, which is worse than either of the individual drivers.

Figure 11 illustrates precisely this cancellation effect, where the sum of the two drivers not only fails to add information but even destroys the signal present in the individual drivers. This result is an example of how drivers with opposite signs can cancel each other out, a phenomenon that may explain why the aggregated ESG sometimes fails to show significant relationships even when the individual components do. Figure 11 also reports a fourth value, which is "Both (together)" and that corresponds to a bivariate model with the two drivers included as separate regressors rather than summed into one variable. This specification recovers part of the predictive power lost by the simple sum, showing that aggregation choices, not only the underlying signals, affect the

results.

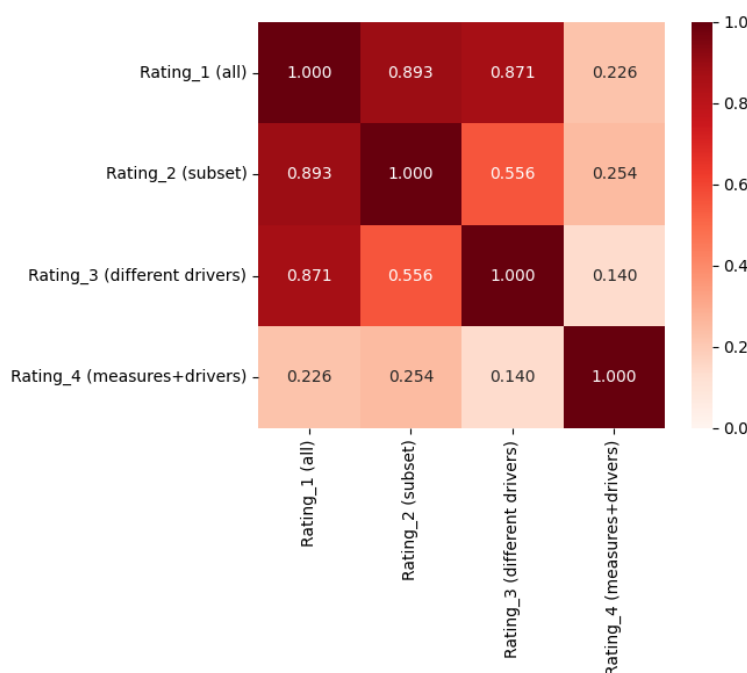


**Figure 11** – Cancellation effect:  $R^2$  Waste, Water.

In conclusion, the experiments conducted confirm the findings of Berg et al. (2022) and extend them to the context of corporate profitability; these results suggest that the fragility of ESG analyses is not due to the absence of a signal, but to its instability with respect to measurement choices.

### 3.9 Implications of ESG Rating Divergence

The experiments conducted in the previous sections have shown that the empirical findings on the relationship between ESG and profitability are sensitive to three main types of divergence: measurement, scope and weight; therefore, this section explores the implications of these findings for the debate on ESG rating divergence, linking the evidence emerging from the analysis with the theoretical framework of Berg, Kölbel and Rigobon (2022)<sup>82</sup>. One of the objectives of the analysis was to verify whether, in a simulated context, it was possible to reproduce the phenomenon of rating divergence documented by Berg et al. (2022). To this end, artificial ratings were constructed by linearly combining drivers with different weights and sets of variables in order to simulate the variability between different rating agencies. Figure 12 shows the correlation matrix between four artificial ratings, obtained by varying both the weights and the selection of drivers; the results show a range of correlations between 0.140 and 0.893, with a mean of 0.490, i.e. a range that fully covers that observed by Berg et al. (2022), which ranged from 0.38 to 0.71.



**Figure 12** – Correlation matrix between four artificial ratings obtained by varying the set of drivers considered and the measurement methods.

The lowest correlation (0.140) is observed between Rating 3 (constructed using a different subset of drivers) and Rating 4 (constructed using different measurement methods), simulating a true measurement divergence, whilst the

<sup>82</sup> Ibid.

highest correlations (above 0.85) are observed between Rating 1 and its two closest variants, namely Rating 2 and Rating 3; indeed, Rating 2 retains most of the same drivers but assigns near-zero weight to a subset of them, whilst Rating 3 retains a different, partially overlapping subset of drivers relative to Rating 1. Consequently, in both cases the underlying informational content overlaps substantially with Rating 1, even though the two ratings diverge from it to a different degree and through a different combination of drivers. Hence, this result confirms that the divergence observed in the original paper is reproducible in a simulated context and, above all, that measurement divergence produces a much greater divergence than weight divergence.

### *3.9.1 Evidence from the Divergence Experiments*

The four experiments conducted provide further insights into the causes of rating divergence. Specifically:

- **Measurement divergence:** the measurement divergence experiment on the Labour driver showed that a change in measurement can transform a driver from non-significant to significant (p-value from 0.91 to 0.02) and, extending this result, it is reasonable to assume that two rating agencies using different indicators to measure the same concept may reach opposite conclusions about the same company, contributing significantly to rating divergence. Furthermore, in the simulated ratings, the lowest correlation (0.140) is obtained precisely when both the drivers used and the measures themselves are changed, which is consistent with Berg et al. (2022), who attribute 56%<sup>83</sup> of the total divergence between ratings to measurement divergence. In other words, it is not enough to share a definition of sustainability; it is also necessary to share the measurement metrics.
- **Weight divergence:** the experiment showed that varying the weights of the drivers produced negligible differences in rating performance ( $R^2$  range = 0.0003); this result is fully consistent with Berg et al. (2022), as they attribute only 6%<sup>84</sup> of the total divergence to weight divergence. In this case, the practical implication is clear: differences between ESG ratings do not depend significantly on how different agencies weight individual dimensions, but rather on which dimensions they consider (scope) and, above all, on how they measure them (measurement). All this suggests that efforts to harmonize ratings should focus on standardizing metrics and indicators, rather than on defining common weights.

---

<sup>83</sup> Ibid., p.1324.

<sup>84</sup> Ibid., p.1324.

- **Sectoral confounding and false positives:** as regards the experiment on false positives, this showed that 5 out of 12 drivers lose statistical significance when controlling for sector, which has direct implications for the construction of ratings; indeed, if two agencies adopt different methodologies to account for sector, they may arrive at very different ratings for the same company. For example, an agency that does not adjust for sector might assign high scores to a company operating in a highly profitable sector, regardless of its actual ESG performance; whereas, another agency that does adjust for sector might assess the same company more accurately, producing a divergent score. This phenomenon contributes significantly to the divergence in ratings observed in practice.
- **Cancellation:** this, on the other hand, has shown that two drivers with opposite signs can cancel each other out when added together. Thus, this phenomenon may explain why aggregated ESG sometimes shows no significant relationship with profitability, even when the individual components (e.g., good environmental practices and governance) have positive and negative effects that offset one another. In fact, rating agencies that aggregate ESG dimensions into a single score risk hiding these opposite relationships, causing confusion among investors trying to interpret the final score.

The practical implications of these findings for investors, regulators and credit rating agencies are discussed in detail in Section 4.4.3

### 3.10 Robustness Checks

At this phase, in order to assess the stability of the results and identify any potential weaknesses, a number of robustness tests were conducted, with the aim of verifying whether the conclusions drawn from the analysis remained valid when changes were made to the estimation methodology, the composition of the sample and the model specifications.

The first test was 5-fold cross-validation on the training set, used to assess the stability of the model's performance, where the regularised Random Forest (with parameters optimised to reduce overfitting) produced an average  $R^2$  of 0.163, with a standard deviation of 0.038, whilst the relative coefficient of variation was 23.3%; which is lower than the performance on the test set ( $R^2 = 0.316$ ), but the difference is expected; indeed, cross-validation mixes observations from different years and sectors, breaking down the temporal and sectoral structure of the data. Consequently, it provides a more conservative estimate of the model's predictive ability and such difference is also intuitive from an economic perspective, as temporal holdout preserves the structural persistence of firms' profitability over time, whereas random cross-validation partially disrupts this persistence by shuffling the observations across years and sectors. Subsequently, in order to assess the relative contribution of the different categories of variables, models were estimated using only specific groups of features:

| <b>Group</b>             | <b>Test <math>R^2</math></b> |
|--------------------------|------------------------------|
| Industry dummies only    | 0.170                        |
| Financial variables only | 0.060                        |
| ESG drivers only         | 0.012                        |

The results show that:

- Industry dummies alone account for 17% of the variance, which reaffirms the dominant role of the sector.
- Financial variables alone have limited predictive power ( $R^2$  0.06).
- ESG drivers alone account for only 1.2% of the variance, suggesting that their contribution only becomes apparent when combined with other variables.

Taken together, these results reinforce the interpretation of ESG as a complementary source of information, rather than an independent driver.

Afterwards, sensitivity to outliers was tested on the Gradient Boosting model by removing the 1% of observations with extreme profit margins (both positive and negative), whilst the sample was reduced from 9,000 to 8,820 observations (98% of the original sample).

| <b>Model (Gradient Boosting)</b> | <b>R<sup>2</sup> Test</b> |
|----------------------------------|---------------------------|
| Original model                   | 0.308                     |
| Without outliers (1%)            | 0.208                     |

The fact that  $R^2$  has decreased by 32% indicates significant sensitivity to outliers, suggesting that part of the predictive signal is driven by the tails of the ProfitMargin distribution. Therefore, the results should be interpreted with caution and the magnitude of the effects should not be overestimated.

Finally, a 2% significance threshold was applied to select the most relevant features in the Gradient Boosting model, and of the initial 30 features, only 12 passed the threshold. Among these:

- **Industry dummies:** 3
- **Driver ESG:** 1 (ESG\_Overall)
- **Financial variables:** 2
- **Others:** 6 (including WaterUsage and interactions)

The model estimated using only 12 features yielded an  $R^2$  of 0.193, representing a 37% reduction compared to the full model ( $R^2$  0.308), which indicates that, although some features have limited individual importance, they collectively contribute to the model's predictive power; therefore, drastically reducing the number of features significantly impairs performance.

### 3.10.1 Summary of Robustness Findings

In summary, the robustness tests revealed:

| <b>Test</b>            | <b>Result</b>                                            | <b>Implication</b>                                                                                                                         |
|------------------------|----------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------|
| Cross-validation       | Average $R^2$ 0.163, variability 23%                     | Moderate stability, conservative CV                                                                                                        |
| Feature groups         | Industry dominates, ESG complementary                    | ESG contributes in combination with other variables                                                                                        |
| Outlier sensitivity    | 32% loss of $R^2$ (Gradient Boosting)                    | Caution regarding the magnitude of effects                                                                                                 |
| Feature selection      | 37% loss of $R^2$ (Gradient Boosting)                    | Features with low importance contribute collectively                                                                                       |
| Permutation Importance | ESG falls from 31% to 23.3% (group) / 16.9% (individual) | It confirms the robustness of the quality measure (sector > ESG), but quantifies the sensitivity to the method used to measure materiality |

Overall, the results show moderate robustness, i.e. the ESG signal is detectable in various specifications, but its magnitude is sensitive to the composition of the sample and the selection of variables.

### 3.11 Economic Interpretation

Even though statistical significance plays a key role, it is also important and necessary to assess the economic relevance of ESG drivers; this section therefore interprets the results of the analysis in economic terms, distinguishing between the predictive contribution of the various ESG dimensions and their practical significance for companies and investors.

#### *3.11.1 Environmental and Governance Drivers*

It has been observed that, in the Random Forest model, environmental drivers account for 15.4% of the total importance, with *Water\_norm* (5.4%), *Carbon\_norm* (5.1%) and *Energy\_norm* (3.1%) among the most significant; therefore, these results suggest that environmental drivers should not be interpreted solely as measures of environmental sustainability but also as proxies for operational efficiency. Indeed, for example, lower water and energy consumption and lower emissions often reflect more efficient production processes, more advanced technologies and/or more rational resource management. In other words, companies that score highly on these dimensions tend to be those that operate most efficiently, which is a factor that has a positive impact on profitability. The overall pattern observed is consistent with the empirical ESG literature, as it often finds a positive relationship between environmental performance and operational efficiency.

Regarding governance drivers, *Corruption* (2.1%) is the most significant of all, and its importance can be interpreted as an indicator of institutional risk; indeed, companies with low levels of perceived corruption tend to have better internal controls, lower legal risk and more transparent, long-term oriented management. Consequently, these factors have a positive impact on stability and profitability. Therefore, the inclusion of *Corruption* among the most important drivers is economically plausible, as the quality of governance is not an ‘optional extra’ but a factor that has a tangible impact on corporate performance.

Finally, the social drivers (*Labor*, *Diversity*, *Community*, *Supply Chain*) account for a combined 7.8% of the total weighting, with individual values all below 2.5%. This weakness of social drivers is a recurring finding in the literature; indeed, social dimensions are more difficult to measure objectively and comparably, and are also characterized by more indirect effects on profitability and longer lead times compared to operational variables. Consequently, it is plausible that, in medium-term quantitative models (the time horizon of this analysis is 10 years), these drivers exhibit lower economic materiality compared to operational (environmental) and governance variables.

### *3.11.2 Economic Significance and the ESG Paradox*

To put it all together, in the Random Forest model, the twelve ESG drivers accounted for a total of around 31% of feature importance, compared with 38.5% attributable to industry dummies, 7.9% to financial variables (Revenue, GrowthRate), and 6.9% to regions; the remaining share is attributable to the aggregate ESG\_Overall score and the raw WaterUsage indicator; this suggests that:

1. ESG contains economically relevant information and is therefore not mere noise.
2. Its contribution remains lower than that of structural factors such as the sector to which a company belongs.
3. ESG acts as a complementary source of information and therefore not as a dominant driver of profitability.

Thus, in economic terms, a company that improves its ESG performance can gain a competitive advantage, but this advantage is modulated by the sector in which it operates and its underlying financial characteristics. That said, the results of the analysis also reveal an apparent paradox; namely, on the one hand, ESG drivers contain significant predictive information (31% of the total significance), but on the other hand, the significance of individual drivers proves to be extremely unstable: 5 out of 12 drivers lose significance when sector controls are applied, and the Labor driver shifts from non-significant to significant simply by altering its measurement.

This paradox highlights the difficulty of isolating ESG effects consistently, particularly when methodological choices substantially alter empirical conclusions.

### 3.12 Summary of Empirical Results

The empirical analysis presented in this chapter yields three main findings regarding the relationship between ESG factors and corporate profitability:

- Firstly, it is found that although ESG factors contain economically relevant information, their contribution appears to be complementary rather than dominant. Indeed, in all model specifications, sector affiliation consistently emerged as the strongest predictor of profitability, accounting for the largest share of explanatory power, while ESG factors collectively added an additional but not negligible contribution. Furthermore, within the ESG dimensions, environmental factors showed the strongest predictive relevance, followed by governance variables, whilst social factors exhibited relatively weaker effects; a result that was confirmed by alternative methods of calculating importance (individual and group permutation importance, Section 3.6.1), which, while reducing the point estimate of the ESG contribution (from 31% to 23.3%), confirm the qualitative conclusion that the sector remains the dominant predictor.
- Secondly, the relevance of ESG factors varies significantly by sector; indeed, analysis within individual sectors revealed substantial heterogeneity in the relevance of individual ESG factors, with Labor, Corruption and Board contributing comparably in the technology sector, environmental factors combined with governance quality (Board) dominating in healthcare, and governance quality (Corruption) next to environmental factors playing the most significant role in the financial sector.
- Finally, the analysis demonstrates that the empirical findings are highly sensitive to measurement choices; consequently, even small variations in how a single ESG dimension is measured are sufficient to reverse statistical significance. As for differences in scope, these altered model performance, and aggregation generated offsetting effects between factors with opposing relationships to profitability. Conversely, changing the aggregation weights produced only marginal differences. Therefore, taken together, these results indicate that measurement divergence, rather than weighting divergence, is the main source of instability in ESG analysis.

In conclusion, robustness checks have largely confirmed these results; although the ESG signal remained detectable across all alternative specifications, its estimated magnitude proved sensitive to sample composition, outliers and feature selection, suggesting that ESG effects are

present but not easily isolated in a stable manner. Therefore, overall, the evidence supports a central conclusion; namely, that the main empirical challenge is not to determine whether ESG is relevant to corporate profitability, but rather to determine how ESG should be measured in a consistent and economically meaningful way. Consequently, in this sense, methodological choices become central to empirical inference, helping to explain why the literature continues to produce conflicting and often contradictory results.

## Chapter 4 – Discussion and Implications

### 4.1 Literature on ESG Ratings Divergence

The phenomenon of ESG rating divergence has been systematically documented since the mid-2010s; indeed, among the earliest studies on this topic are those by Dorfleitner, Halbritter and Nguyen (2015)<sup>85</sup> who demonstrated that ratings assigned by different providers exhibit low correlations and significantly different distributions, with only a small percentage of companies falling within the same score quartile depending on the provider used; this was an early indication that divergence was not a marginal phenomenon, but a structural characteristic of the ESG rating sector. Subsequently, Chatterji et al. (2016)<sup>86</sup> analyzed six providers, finding low comparability between scores even after adjusting for explicit differences in the definitions of corporate social responsibility. Therefore, their conclusion is that the divergence cannot be attributed solely to differences in evaluative judgement but, rather, reflects structural differences in the way each provider defines and measures sustainability; this conclusion directly anticipates the breakdown subsequently formalized by Berg et al.

Undoubtedly, the most influential contribution on this subject remains that of Berg, Kölbel and Rigobon (2022)<sup>87</sup>, who analyzed data from six agencies (KLD, Sustainalytics, Moody's ESG, S&P Global, Refinitiv and MSCI) and, using a common taxonomy of 64 categories, broke down the divergence into measurement divergence (56%), scope divergence (38%) and weight divergence (6%). This finding is therefore crucial as it shifts the focus from the debate on weightings, which proves to be relatively insignificant, to the crucial issue of how the underlying concepts are measured.

Furthermore, Christensen, Serafeim and Sikochi (2022)<sup>88</sup> have also added a counterintuitive element to this debate, demonstrating that greater corporate disclosure does not reduce the divergence between ratings, but may even widen it, as it provides more material for divergent interpretations by the agencies; this finding has direct implications for the effectiveness of the European regulations analysed in Chapter 2; namely, if the mandatory disclosure requirements introduced by the CSRD generate more data without

---

<sup>85</sup> Dorfleitner, G., Halbritter, G., Nguyen, M. (2015). Measuring the level and risk of corporate responsibility – An empirical comparison of different ESG rating approaches. *Journal of Asset Management*, 16(7), 450–466. <https://doi.org/10.1057/jam.2015.31>

<sup>86</sup> Chatterji, A.K., Durand, R., Levine, D.I., Touboul, S. (2016). Do Ratings of Firms Converge? Implications for Managers, Investors and Strategy Researchers. *Strategic Management Journal*, 37(8), 1597–1614.

<sup>87</sup> Berg, F., Kölbel, J.F., Rigobon, R. (2022). Aggregate Confusion: The Divergence of ESG Ratings. *Review of Finance*, 26(6), 1315–1344.

<sup>88</sup> Christensen, D.M., Serafeim, G., Sikochi, A. (2022). Why is Corporate Virtue in the Eye of the Beholder? The Case of ESG Ratings. *The Accounting Review*, 97(1), 147–175.

harmonizing the way in which that data is used, then the divergence could persist or even increase.

Finally, Billio et al. (2021)<sup>89</sup> examined the consequences of divergence for asset pricing, demonstrating precisely that heterogeneity in ESG ratings leads to the identification of alternative benchmarks and that disagreement amongst providers disperses the effect of ESG investor preferences on asset prices, to the extent that even in cases where there is agreement amongst agencies, this does not produce significant impacts on financial performance. Therefore, in the context of this thesis, this result can be interpreted as a consequence of measurement divergence rather than as evidence of the absence of an underlying ESG signal.

---

<sup>89</sup> Billio, M., Costola, M., Hristova, I., Latino, C., Pelizzon, L. (2021). Inside the ESG ratings: (Dis)agreement and performance. *Corporate Social Responsibility and Environmental Management*, 28(5), 1426–1445. <https://doi.org/10.1002/csr.2177>

## 4.2 Literature on ESG and Financial Performance

In the present literature, the relationship between ESG performance and financial performance has been the subject of a growing number of studies, with results that are far from conclusive. In this regard, it is noted that Friede, Busch and Bassen (2015)<sup>90</sup>, in a meta-analysis based on over 2,000 empirical studies, found that around 90% of the studies reported a non-negative relationship between ESG and financial performance, with the majority documenting the presence of a positive link. However, the same authors caution that methodological heterogeneity makes it difficult to draw generalizable conclusions.

Subsequently, on a theoretical level, Eccles, Ioannou and Serafeim (2014)<sup>91</sup> demonstrated that companies with long-term sustainability policies tend to outperform their peers in the long run, but that these effects are moderated by sector and capital structure. Therefore, it is clear that this evidence is consistent with the findings of this thesis regarding the significant sectoral variability in ESG materiality.

Additionally, a more recent review conducted by Patekar and Mahajan (2025)<sup>92</sup>, which systematically analyses 287 articles published between 2015 and 2024, confirming that most studies report a positive relationship between ESG performance and firm value but also highlights negative or non-significant results in numerous cases, likely due to variations across sectors, regions and time horizons. Consequently, the review itself highlights three recurring methodological gaps in the literature, being: the lack of standardized ESG ratings; the need to separate the E, S and G components rather than using aggregate scores; and the importance of advanced causal methods to address endogeneity issues. Therefore, these gaps are directly relevant to this thesis, which adopts a controlled experimental approach precisely to isolate the sources of bias that plague studies based on real-world data.

---

<sup>90</sup> Friede, G., Busch, T., Bassen, A. (2015). ESG and Financial Performance: Aggregated Evidence from More than 2000 Empirical Studies. *Journal of Sustainable Finance & Investment*, 5(4), 210–233. <https://doi.org/10.1080/20430795.2015.1118917>

<sup>91</sup> Eccles, R.G., Ioannou, I., Serafeim, G. (2014). The Impact of Corporate Sustainability on Organizational Processes and Performance. *Management Science*, 60(11), 2835–2857. <https://doi.org/10.1287/mnsc.2014.1984>

<sup>92</sup> Patekar, A., Mahajan, S. (2025). Unraveling the ESG-firm value nexus: A systematic analysis of current literature and emerging themes. *Business Performance Review*, 3(2), 49–65. <https://doi.org/10.22495/bprv3i2p5>

### 4.3 Positioning of This Thesis in Literature

This thesis positions itself at the intersection of the two strands of research described in the previous sections; in particular, it makes a contribution that distinguishes itself from the existing literature on three levels.

1. **The identified gap:** the literature on divergence<sup>9394</sup> has focused on measuring divergence and identifying its causes, but has not systematically analysed how this divergence translates into instability in the empirical results of ESG-profitability analyses. Consequently, this implies that the literature has demonstrated that ratings diverge but has not, however, quantified the impact of this divergence on the conclusions drawn by researchers and investors.
2. **The methodological contribution:** unlike previous studies, this thesis does not merely document divergence, but instead provides systematic evidence of it in the empirical results, using a simulated dataset with known properties; in this way, it was therefore possible to systematically vary the measurement choices, the set of attributes and the aggregation weights, and thus isolate the systematic effects of each type of divergence; clearly, this approach is not feasible with real data, where, by contrast, the sources of distortion cannot be separated.
3. **The theoretical contribution:** this thesis proposes a reformulation of the central question in the ESG debate; that is, rather than asking “Does ESG improve profitability?”, the more productive question is “Using which metrics, on which drivers, in which sectors and under which controls does ESG emerge as relevant?”. Hence, this shift in perspective is empirically supported by the results and has direct implications for future research, investor practice and regulatory policy.

---

<sup>93</sup> Berg, F., Kölbel, J.F., Rigobon, R. (2022). Aggregate Confusion: The Divergence of ESG Ratings. *Review of Finance*, 26(6), 1315–1344.

<sup>94</sup> Chatterji, A.K., Durand, R., Levine, D.I., Touboul, S. (2016). Do Ratings of Firms Converge? Implications for Managers, Investors and Strategy Researchers. *Strategic Management Journal*, 37(8), 1597–1614. <https://doi.org/10.1002/smj.2407>

## 4.4 Discussion of Empirical Findings

### *4.4.1 Key Findings and Their Interpretation*

The empirical analysis reveals four key messages which, beyond the numerical details, define the substantive significance of the findings.

The first key finding concerns measurement divergence, which is not merely a theoretical issue, but a concrete source of empirical instability; in fact, by altering the measurement of a single driver (Labour), the same concept, that is labour practices, shifted from being statistically insignificant to significant and it is precisely this that demonstrates how it is possible for two researchers analyzing the same firms with the same hypotheses, but using slightly different metrics, to reach opposing conclusions. Therefore, this is direct evidence that the divergence in ratings is not merely a disagreement between agencies, but a problem that extends to empirical inference. Consequently, this result highlights that sensitivity to measurement choices is not a marginal specification issue but rather a structural characteristic of the ESG phenomenon, reflecting the sensitivity of ESG variables to the measurement choices adopted in the data construction process.

Furthermore, as noted previously, the sector to which a company belongs is a powerful confounding factor. Indeed, it has been shown that five out of twelve drivers lose statistical significance when sector-specific controls are introduced. Therefore, without such controls, there is a risk of attributing relationships to ESG factors that are in fact a reflection of the structural profitability of the sector in which the companies operate. In other words, the sector acts as a variable that implicitly incorporates competitive conditions, cost structure and regulatory intensity; that is, all characteristics that may be erroneously attributed to ESG drivers in the absence of adequate controls.

Additionally, it has been demonstrated that ESG contains economically relevant information but acts as a complementary rather than a dominant source; indeed, this is precisely what is inferred and demonstrated in the Random Forest model, where ESG drivers account for around 31% of the total predictive power, compared with approximately 38.5% attributable to sectoral variables. Therefore, ESG is not ‘noise’, but neither is it the main driver of profitability; consequently, it acts as a source of additional information that overlaps with structural factors, contributing to the model’s predictive power without representing a causal driver of corporate profitability. In addition, this estimate proves to be robust even when alternative methods are used to calculate the importance of features, although the exact magnitude of the ESG contribution varies between 16.9% and 31% depending on the metric used.

As regards weight divergence, this is not really a concern, since measurement divergence is the real issue. In fact, varying the weights of the drivers produces

minimal differences in the models' predictive power, whilst changing the measurement or excluding an entire pillar has significant effects. The debate over weights, which is often central to discussions on rating methodologies, is therefore secondary to the question of what is measured and how it is measured. Consequently, this finding suggests that the academic and regulatory debate focused on the weighting of indicators may be less relevant than the need to standardize the methods for collecting and defining the underlying metrics.

#### 4.4.2 Comparison with the Existing Literature

At this point, it can be confirmed that the results just described are largely consistent with the literature but extend it in some significant ways.

Indeed, as found by Berg et al. (2022)<sup>95</sup>, in this study too, measurement divergence is confirmed to be the most influential component, whilst weight divergence proves to be negligible. Consequently, the convergence between these findings and those of Berg et al. suggests that the importance of measurement divergence is not limited to a specific empirical setting.

Furthermore, the findings of Chatterji et al. (2016)<sup>96</sup>, which have documented the low commensurability between ratings without fully identifying the causes. Therefore, the experiments on false positives conducted in this thesis offer a possible explanation; that is, in the absence of adequate sectoral controls, spurious relationships are erroneously attributed to ESG and, consequently, the low commensurability is not necessarily accidental, but could be a predictable consequence of measurement divergence combined with sectoral confounding. Furthermore, the results obtained from the Labor experiment are consistent with the counterintuitive finding by Christensen et al. (2022)<sup>97</sup>; namely, that even when the underlying data are the same, the way in which they are measured and aggregated can lead to opposing conclusions. Thus, greater data availability does not automatically imply greater convergence in assessments. Finally, this thesis makes a contribution that complements the existing literature on the role of the sector in the ESG-performance relationship. Indeed, several studies have recognised that ESG materiality varies systematically across sectors; in particular, Khan, Serafeim e Yoon (2016)<sup>98</sup> demonstrated that

---

<sup>95</sup> Berg, F., Kölbel, J.F., Rigobon, R. (2022). Aggregate Confusion: The Divergence of ESG Ratings. *Review of Finance*, 26(6), 1315–1344.

<sup>96</sup> Chatterji, A.K., Durand, R., Levine, D.I., Touboul, S. (2016). Do Ratings of Firms Converge? Implications for Managers, Investors and Strategy Researchers. *Strategic Management Journal*, 37(8), 1597–1614. <https://doi.org/10.1002/smj.2407>

<sup>97</sup> Christensen, D.M., Serafeim, G., Sikochi, A. (2022). Why is Corporate Virtue in the Eye of the Beholder? The Case of ESG Ratings. *The Accounting Review*, 97(1), 147–175. <https://doi.org/10.2308/TAR-2019-0506>

<sup>98</sup> Khan, M., Serafeim, G., Yoon, A. (2016). Corporate Sustainability: First Evidence on Materiality. *The Accounting Review*, 91(6), 1697–1724. <https://doi.org/10.2308/accr-51383>

companies with good ratings on ESG dimensions that are material to their sector significantly outperform companies with good ratings on immaterial dimensions, which shows that not all ESG factors are equally relevant across all industrial contexts. Nevertheless, their work focuses on identifying material drivers rather than analyzing sectoral confounding as a source of false positives in empirical inference. Hence, the findings of this thesis show that this second effect is, in fact, far from marginal; this is evident from the fact that five out of twelve drivers lose statistical significance when sectoral controls are introduced, indicating that a significant proportion of the apparently significant relationships between ESG and profitability may reflect structural differences between sectors rather than a genuine ESG effect.

#### *4.4.3 Implications for Stakeholders*

At this stage, interpreting the results allows tangible conclusions to be drawn for the key stakeholders involved in the ESG debate, namely investors, regulators and rating agencies. In this regard, the variability of the results depending on the measurement approach suggests to investors that relying on a single aggregated ESG rating for portfolio decisions entails significant methodological risks. It is therefore advisable to compare assessments from multiple providers, analyze the disaggregated components of the ratings rather than the synthetic score, and take into account the sector to which a company belongs as an essential contextual factor for making companies comparable. As regards regulators, the results show that harmonizing definitions, that is, what to measure, as required by the EU Taxonomy and the CSRD, is not sufficient to reduce divergence; it is therefore also necessary to harmonize metrics, that is, how to measure. In this context, Regulation (EU) 2024/3005 of the European Parliament and of the Council of 27 November 2024 on the transparency and integrity of ESG rating activities<sup>99</sup> forms part of the broader European regulatory process aimed at strengthening the transparency and comparability of ESG ratings; therefore, requiring agencies to publish their methodologies and to separate ratings by pillar. However, as the experiments conducted in this thesis provides evidence, even within the same pillar, that different measurement choices can produce divergent results; one of which is that standardization must operate at the level of indicators, not just definitions. Furthermore, without a minimum level of methodological harmonization, there is a risk that ESG ratings will continue to produce inconsistent information signals, thereby reducing the effectiveness of market-based sustainability policies.

---

<sup>99</sup> Regulation (EU) 2024/3005 of the European Parliament and of the Council of 27 November 2024 on the transparency and integrity of ESG rating activities (OJ L, 2024).

Finally, for rating agencies, the findings on measurement divergence suggest that agencies should be more transparent regarding the measurement choices they adopt, and not merely regarding the weightings. Therefore, transparency regarding the specific indicators used for each driver would enable users to make an informed assessment of the comparability of scores across different providers.

## Chapter 5 – Conclusions

### 5.1 Overall Summary of Findings

At this point, the four initial research questions that guided the study can now be summarized.

1. In response to the question of which ESG dimensions are most relevant, it can be seen that environmental drivers (water, carbon, energy) and the governance driver relating to corruption are the most informative in predictive terms; while, by contrast, social drivers play a marginal role. This pattern therefore suggests that environmental drivers largely capture operational efficiency, whilst the governance driver reflects institutional risk.
2. Regarding the role of the sector, it is observed that this is a strong confounding factor. In fact, five out of twelve drivers lose significance when sector controls are applied, whilst ESG materiality varies significantly across sectors; indeed, no single non-environmental driver dominates in Technology, governance quality (Board) is most relevant in Healthcare, and anti-corruption remains central in Finance.
3. As regards the impact of measurement choices, it emerges that measurement divergence is the main source of empirical instability, whilst weight divergence has a negligible impact. Hence, by altering the measurement of a single driver (Labor), the same concept shifted from being non-significant to significant, demonstrating how measurement choices can alter empirical conclusions.
4. In terms of economic significance, ESG drivers account for around 31% of the model's predictive power, compared with approximately 38.5 per cent attributable to sectoral variables. Consequently, ESG acts as a complementary source of information, rather than as a dominant driver of profitability. Therefore, this estimate has proved to be robust, although it varies between 16.9% and 31% depending on the method used to calculate the ESG driver's share; in any case, it confirms the sector's predominant role.

Therefore, taken together, these four findings paint a coherent picture; being the fact that ESG is not devoid of informative content but this content is distributed unevenly across dimensions and sectors and is systematically sensitive to the methodological choices made in the construction of the metrics.

According to this, it implies that the empirical weakness characterizing part of the ESG literature; namely, studies finding positive, null or negative effects on the same research question, does not necessarily reflect a genuine instability in the underlying economic relationship, but can largely be attributed to differences in measurement choices, the selection of attributes and the handling of sectoral confounding. Therefore, in other words, the problem is not that ESG lacks a signal but that the signal is difficult to isolate consistently when metrics are not harmonized. With that being said, the key conclusion is that the empirically relevant question is not whether ESG is relevant to corporate profitability, but to understand under what conditions; that is, through which metrics, in which sectors, and with which controls, this relevance emerges in an economically interpretable way.

## 5.2 Contribution of the Thesis

The original contribution of this thesis is structured around three main areas:

1. **Theoretical contribution:** it does aim at reframing the debate on ESG, however the question “Does ESG improve profitability?” has proved to be poorly framed from an empirical perspective; in fact, the results show that the answer depends critically on the measurement choices and the sectoral context. Therefore, the main and most productive question for future research is: “Using which metrics, on which drivers, in which sectors, and with which controls does ESG emerge as relevant?”. It is precisely because of this shift in perspective that leads to concrete implications for investors, regulators and the design of future studies.
2. **Methodological contribution:** obtained with the introduction of a controlled computational experiment as a tool for analyzing the consequences of ESG rating divergence on empirical inference. Indeed, unlike previous studies, which merely documented the divergence, this thesis systematically isolates its consequences by varying the measurement choices, the set of attributes and the aggregation weights. Therefore, this approach is complementary to studies using real-world data; whereas the latter suffer from endogeneity and missing data, the simulated experiment offers control and causal transparency.
3. **Empirical contribution:** it also provides systematic evidence that measurement divergence is a concrete source of instability; thus, the same driver, measured slightly differently, can shift from being non-significant to significant. Furthermore, it shows that sector is a powerful confounder and that weight divergence is negligible, thereby confirming the findings of Berg et al. (2022)<sup>100</sup> in an alternative environment. Secondly, it quantifies the cancellation effect between drivers with opposite signs, which offers a mechanistic explanation of why aggregate ESG sometimes shows non-significant relationships with profitability despite the individual components being informative. Moreover, the thesis demonstrates that the choice of technique for measuring the importance of features, and not just the choice of the underlying ESG indicators, constitutes an additional source of methodological divergence, thereby extending the concept of ‘measurement divergence’ from the data level to that of model interpretation.

---

<sup>100</sup> Berg, F., Kölbel, J.F., Rigobon, R. (2022). Aggregate Confusion: The Divergence of ESG Ratings. *Review of Finance*, 26(6), 1315–1344.

Finally, the empirical findings also relate directly to the regulatory framework analyzed in Chapter 2. Indeed, the three European regulations examined, being the EU Taxonomy, the SFDR and the CSRD, share a common structural limitation, as while they define with increasing precision what needs to be measured in terms of sustainability, they leave the methods by which such measurements are to be carried out largely undefined. Therefore, the findings of this thesis show that this is precisely the critical issue: ‘measurement divergence’, which is differences in how things are measured, is responsible for the largest share of empirical instability, whilst ‘scope divergence’, instead is differences in what is measured, produces more limited effects. Hence, it follows that the European regulations, however ambitious and necessary they may be, are primarily addressing the less influential component of the divergence. It is precisely the Regulation (EU) 2024/3005 on the transparency of ESG ratings that represents a step in the right direction, requiring agencies to publish their methodologies; however, the publication of methodologies is not the same as their harmonization; indeed, as long as providers are able to continue using different indicators to measure the same concepts, structural divergence in ratings will persist, regardless of the level of formal transparency achieved.

### 5.3 Limitations and Directions for Future Research

The results of this thesis must be interpreted, however, considering certain methodological limitations that it is important to explicitly acknowledge; among these, the first and most significant concerns the use of a simulated dataset. Indeed, whilst offering a high degree of experimental control, which is precisely the advantage that motivated this methodological choice, the simulation implies that the results must be interpreted as structural evidence regarding the mechanisms of divergence rather than as point estimates applicable to real-world contexts. In particular, it is reasonable to assume that the simulated experiment tends to underestimate the actual instability present in real markets; that is, the simulated drivers were constructed with a known and controlled correlation structure, whereas in real data the relationships between ESG indicators are noisier, and as regards missing data, these are frequent and the providers' methodologies are opaque. Consequently, if volatility is already evident in a simulated and favorable context, it is plausible that it is even more pronounced in real-world data. Therefore, the quantitative results should not be extrapolated beyond the experimental context, but the direction of the observed effects is likely to be conservative compared with reality.

The second limitation concerns the absence of causal inference in the strict sense; indeed, whilst the temporal separation between the training and test sets reduces the risk of look-ahead bias, it does not eliminate the possibility of residual endogeneity; in fact, problems of reverse causality or omitted variables could simultaneously influence financial performance and sustainability practices. Therefore, it is not possible to assert that an improvement in ESG practices causes an increase in profitability; this is why the results suggest an association which, in the simulated context, is sensitive to measurement choices, yet the direction of causality remains an open question requiring more advanced methodologies.

The third limitation concerns the machine learning models used; since the Random Forest is the model with the best predictive performance, it can be sensitive to the choice of hyperparameters and prone to overfitting, as documented by the robustness tests which show that removing 1% of the outliers resulted in a 32% reduction in  $R^2$ , indicating that part of the predictive signal is concentrated in the tails of the distribution. Furthermore, the importance of variables in the Random Forest reflects their contribution to the model's predictive power and not necessarily direct economic relevance; thus, a variable may be predictively useful without being causally relevant.

Consequently, these limitations do not invalidate the main conclusions but define the scope of their applicability and point to the most promising directions for future research. In this regard, the following possibilities for future research can be highlighted:

1. One first thread concerns external validation using real-world data, which means replicating the analysis using databases provided by providers such as MSCI, Refinitiv or Sustainalytics would make it possible to verify whether the instability mechanisms documented in the simulated experiment also manifest themselves in the observed evidence. In particular, it would be interesting to test the measurement divergence experiment on the ‘Labor’ driver, which in this study showed the greatest sensitivity to measurement choices using real-world data from providers that employ different indicators to measure labor practices; in this way, it is possible to verify whether the variation in significance observed in the simulation is also replicated in a comparison between real-world agencies.
2. A second thread, on the other hand, concerns the identification of more robust causal relationships; that is, using methodologies such as instrumental variables, difference-in-differences models or event studies; for example, based on exogenous regulatory changes such as the entry into force of the CSRD. Hence, these could isolate the causal effect of ESG practices on profitability, thus reducing the endogeneity issues that plague the existing literature and that this thesis was unable to solve due to its design.
3. Finally, a third thread concerns the development of sector-specific ESG metrics. The results show that ESG materiality varies significantly across sectors, with distinct patterns in Technology, Healthcare and Finance, making a uniform measurement approach likely to be inadequate. Indeed, disaggregated and contextualized frameworks, similar to those promoted by the SASB and the ISSB, could improve both the comparability between studies and the usefulness of ratings for investment decisions. Because of this reason, a natural extension of this work would be to conduct sector-specific divergence experiments, in which measurement choices are varied not at an aggregate level but within each sector, in order to verify whether sensitivity to measurement parameters itself varies depending on the industrial context.

## 5.4 Concluding Remarks

In conclusion, this thesis has shown that the relationship between ESG and corporate profitability is significant but not linear, and that its relevance is strongly influenced by the measurement choices made, the sector in question and the complexity of the interactions between the various ESG dimensions. Therefore, the main contribution does not lie in demonstrating that ‘ESG works’, even though the results only partially support this claim and, above all, with considerable reservations, but rather in showing how ESG must be analyzed to avoid misleading conclusions. Indeed, the sensitivity of the empirical results to measurement choices is not a flaw in the analysis but a characteristic of the phenomenon itself. Consequently, the divergence in ESG ratings is not merely a temporary disagreement but rather reflects the intrinsic complexity of measuring sustainability in a context characterized by incomplete information, sectoral heterogeneity and the absence of globally agreed standards.

According to this, in a landscape where ESG ratings are increasingly influencing investment decisions, corporate strategies and regulatory policies, understanding the nature of their divergence and its implications for empirical analysis is not merely an academic exercise but a genuine practical necessity; which is why it is now widely recognized that there is no single correct way to measure ESG criteria, as there are measurement choices whose consequences must be explicitly acknowledged, communicated and managed.

In this sense, the challenge is not to eliminate divergence, but to understand its sources and consciously manage its analytical and decision-making implications.



## References

### EU LEGISLATION AND OFFICIAL DOCUMENTS

Commission Delegated Regulation (EU) 2021/2139 of 4 June 2021 supplementing Regulation (EU) 2020/852 (Climate Delegated Act). *Official Journal of the European Union*, L 442, 9 December 2021.

Commission Delegated Regulation (EU) 2023/2772 of 31 July 2023 supplementing Directive (EU) 2022/2464 as regards European Sustainability Reporting Standards (ESRS). *Official Journal of the European Union*, L 2023/2772, 22 December 2023.

Directive (EU) 2022/2464 of the European Parliament and of the Council of 14 December 2022 amending Regulation (EU) No 537/2014, Directive 2004/109/EC, Directive 2006/43/EC and Directive 2013/34/EU, as regards corporate sustainability reporting (CSRD). *Official Journal of the European Union*, L 322, 16 December 2022, pp. 15–80.

European Commission. *Action Plan: Financing Sustainable Growth*. COM(2018) 97 final, 8 March 2018.

European Commission. *The European Green Deal*. COM(2019) 640 final, 11 December 2019.

European Commission. *Proposal for a Corporate Sustainability Reporting Directive*. COM(2021) 189 final, 21 April 2021.

European Commission. *Call for Evidence: Simplification and Improved Effectiveness of the Sustainable Finance Disclosure Regulation (SFDR)*. 2 May 2025.

European Commission. *Omnibus I Package*. COM(2025) 80 final, 26 February 2025.

European Commission. *Proposal for a Regulation of the European Parliament and of the Council amending Regulation (EU) 2019/2088 on sustainability-related disclosures in the financial services sector (SFDR 2.0)*. COM(2025) 841 final, 20 November 2025.

Regulation (EU) 2019/2088 of the European Parliament and of the Council of 27 November 2019 on sustainability-related disclosures in the financial

services sector (SFDR). *Official Journal of the European Union*, L 317, 9 December 2019, pp. 1–16.

Regulation (EU) 2020/852 of the European Parliament and of the Council of 18 June 2020 on the establishment of a framework to facilitate sustainable investment (Taxonomy Regulation). *Official Journal of the European Union*, L 198, 22 June 2020, pp. 13–43.

Regulation (EU) 2024/3005 of the European Parliament and of the Council of 27 November 2024 on the transparency and integrity of Environmental, Social and Governance (ESG) rating activities, and amending Regulations (EU) 2019/2088 and (EU) 2023/2859. *Official Journal of the European Union*, OJ L 2024/3005, 12 December 2024.

## REPORTS AND INSTITUTIONAL PUBLICATIONS

European Financial Reporting Advisory Group (EFRAG). *ESRS Explanatory Note: Double Materiality*. November 2023.

European Securities and Markets Authority (ESMA). *Final Report on Greenwashing*. ESMA36-287652198-2699, 4 June 2024.

European Securities and Markets Authority (ESMA). *Opinion on the functioning of the Sustainable Finance Framework*. ESMA36-1079078717-2587, 24 July 2024.

European Supervisory Authorities (ESAs). *Principal Adverse Impact Disclosures under the Sustainable Finance Disclosure Regulation — Annual Report to the Commission under Article 18 of Regulation (EU) 2019/2088*. JC 2025 26, September 2025.

International Sustainability Standards Board (ISSB). *IFRS S1: General Requirements for Disclosure of Sustainability-related Financial Information*. June 2023.

Platform on Sustainable Finance. *Final Report on Social Taxonomy*. February 2022.

Platform on Sustainable Finance. *Report on Advancing Sustainable Finance*. April 2025.

Platform on Sustainable Finance. *Technical Working Group – Part A:*

*Methodological Report*. March 2022.

Principles for Responsible Investment (PRI). *Annual Report 2024*.

## ACADEMIC LITERATURE

Berg, F., Kölbel, J.F., Rigobon, R. (2022). Aggregate Confusion: The Divergence of ESG Ratings. *Review of Finance*, 26(6), 1315–1344. <https://doi.org/10.1093/rof/rfac033>

Billio, M., Costola, M., Hristova, I., Latino, C., Pelizzon, L. (2021). Inside the ESG ratings: (Dis)agreement and performance. *Corporate Social Responsibility and Environmental Management*, 28(5), 1426–1445. <https://doi.org/10.1002/csr.2177>

Chatterji, A.K., Durand, R., Levine, D.I., Touboul, S. (2016). Do Ratings of Firms Converge? Implications for Managers, Investors and Strategy Researchers. *Strategic Management Journal*, 37(8), 1597–1614. <https://doi.org/10.1002/smj.2407>

Christensen, D.M., Serafeim, G., Sikochi, A. (2022). Why Is Corporate Virtue in the Eye of the Beholder? The Case of ESG Ratings. *The Accounting Review*, 97(1), 147–175. <https://doi.org/10.2308/TAR-2019-0506>

Cochran, I., Mackenzie, C., Brander, M. (2025). EU's sustainable finance disclosure regulation: Does the hybrid reporting regime undermine the goal to reorient capital to climate action? *Climate Policy*, 25(1), 76–88. <https://doi.org/10.1080/14693062.2024.2353115>

Dorfleitner, G., Halbritter, G., Nguyen, M. (2015). Measuring the level and risk of corporate responsibility – An empirical comparison of different ESG rating approaches. *Journal of Asset Management*, 16(7), 450–466. <https://doi.org/10.1057/jam.2015.31>

Eccles, R.G., Ioannou, I., Serafeim, G. (2014). The Impact of Corporate Sustainability on Organizational Processes and Performance. *Management Science*, 60(11), 2835–2857. <https://doi.org/10.1287/mnsc.2014.1984>

Friede, G., Busch, T., Bassen, A. (2015). ESG and Financial Performance: Aggregated Evidence from More than 2000 Empirical Studies. *Journal of Sustainable Finance & Investment*, 5(4), 210–233. <https://doi.org/10.1080/20430795.2015.1118917>

Khan, M., Serafeim, G., Yoon, A. (2016). Corporate Sustainability: First Evidence on Materiality. *The Accounting Review*, 91(6), 1697–1724. <https://doi.org/10.2308/accr-51383>

Patekar, A., Mahajan, S. (2025). Unraveling the ESG-firm value nexus: A systematic analysis of current literature and emerging themes. *Business Performance Review*, 3(2), 49–65. <https://doi.org/10.22495/bprv3i2p5>

## **OTHER SOURCES (POLICY / INDUSTRY / NGO)**

Advokatfirmaet BAHR. *SFDR 2.0: From Paper Exercise to Impact?* 5 June 2025.

Finance Watch. *Response to the Consultation on Taxonomy Criteria and Disclosures*. 26 March 2025.

Frank Bold. *Analysis of the SFDR 2.0 Proposal: Progress and Remaining Challenges*. February 2026.

International Capital Markets Association (ICMA). *Paper on the Simplification of the EU Sustainable Finance Legislation*. 5 February 2025.

Sustainability Accounting Standards Board (SASB). *Materiality Finder*.