Annals of Operations Research
https://doi.org/10.1007/s10479-026-07378-5

ORIGINAL RESEARCH

From hundreds to dozens: stochastic hill climbing for parsimonious environmental, social, and governance scoring

Matteo Benuzzi[1] · **Özge Şahin**[2] · **Sandra Paterlini**[1]

Received: 13 October 2025 / Accepted: 3 August 2026

Abstract
Environmental, social, and governance (ESG) metrics are widely adopted tools for assessing corporate sustainability. Despite their importance for investors, regulators, and other stakeholders, ESG scores suffer from major drawbacks, including score divergence across providers, opaque and biased methodologies, and data quality issues. This paper focuses on Eikon Refinitiv, a prominent ESG data provider, to explore two interrelated challenges: the distortion of association measures due to missing data imputation and the potential redundancy among ESG key performance indicators (KPIs). First, we show that Refinitiv's percentile ranking scheme preserves rank correlations (Kendall's tau, Spearman's rho) but not Pearson's correlation when data are complete, but Refinitiv's imputation for missing values with zeros inflates all association measures. Second, we develop an optimization-based tool to identify informational redundancy among KPIs. We formulate it as a cardinality-constrained rank correlation maximization problem and solve it via stochastic hill climbing with random restarts. Across three sectors and 10 years, we find that only 20–25% of KPIs are needed to approximate pillar scores, suggesting that ESG scoring models can be substantially streamlined without sacrificing accuracy. Our theoretical and empirical findings have important implications for regulatory efforts to standardize ESG ratings and offer practical guidance for constructing interpretable, reliable, and efficient ESG scores.

Keywords OR in environment and climate change · ESG · Imputation · KPI · Stochastic hill climbing

M. Benuzzi and Ö. Şahin contributed equally to this work.

✉ Matteo Benuzzi
matteo.benuzzi@unitn.it

Özge Şahin
O.Sahin@tudelft.nl

Sandra Paterlini
sandra.paterlini@unitn.it

[1] Department of Economics and Management, University of Trento, Via Inama, 5, 38122 Trento, Italy

[2] Delft Institute of Applied Mathematics, Delft University of Technology, Mekelweg 4, 2628 CD Delft, The Netherlands

Published online: 13 August 2026

1 Introduction

Environmental, Social, and Governance (ESG) metrics are widely used for evaluating companies' sustainability and ethical performance. These metrics serve as an essential tool for investors, regulators, stakeholders and the public, offering insights into a company's environmental impact, social responsibility, and governance structures. However, ESG ratings suffer from numerous drawbacks, including significant divergence among providers (Berg et al., 2022; Billio et al., 2021), non-transparent and biased construction methodologies, and data quality issues (Sahin et al., 2023), such as missingness (Caprioli et al., 2025; Downing, 2025). Additionally, there is a notable bias toward larger companies (Dobrick et al., 2023; Drempetic et al., 2020). As the global financial system should shift toward more sustainable products, regulators have increasingly recognized the need for reform to enhance ESG score reliability and methodology transparency. For instance, on 13 June 2023, the EU proposed a regulation aimed at improving the transparency, impartiality, reliability, and comparability of ESG rating methodologies (Malecki, 2023). It has been approved as Regulation (EU) 2024/3005 by the European Commission in 2024. This regulation establishes a framework requiring ESG rating providers to be authorized by the European Securities and Markets Authority (ESMA) to prevent conflicts of interest. Besides, providers must disclose their methodologies, models, and key rating assumptions to reduce complexity and improve transparency and representativeness. We expect the new regulation to have a large impact not only on ESG providers but also on ESG score users.

A significant challenge in aggregating ESG data stems from the diverse types of key performance indicators (KPIs) used in these evaluations. These KPIs range from binary variables, such as whether a company has a diversity policy in place, to continuous variables, like the amount of carbon emissions a company produces (scaled for its revenues). Further, discrepancies in reporting standards across regions and industries lead to missing or inconsistent data, complicating the aggregation process (Sahin et al., 2022). Effectively handling these challenges is crucial to ensure the ESG scores offer a reliable reflection of a company's sustainability practices.

Another source of complexity lies in the percentile ranking methodologies frequently employed in ESG scoring (Benuzzi et al., 2025). Due to this data aggregation process, the scores of a company depend on the performance of its peers. A firm improving emissions might see its score drop if peers improve more than the firm itself. While useful for comparison, this methodology can complicate the interpretation of scores over time, especially as more companies adopt ESG reporting practices or improve their performance in specific areas (Sahin et al., 2023).

Moreover, while ESG scores aim to provide comprehensive coverage, including numerous KPIs can lead to redundancy, and overlapping KPIs offer little new information. The way in which missing data is handled (i.e., typically with an imputation of a zero) also represents a source of redundancy in the data. This redundancy can ultimately undermine the clarity and interpretability of ESG scores. Identifying redundancy is an important step, as it may suggest that the ESG scoring process could be streamlined by reducing the number of KPIs. Then, it enhances the efficiency and interpretability of the scores without losing meaningful information. Therefore, more efficient scoring models that maintain informational depth without unnecessary overlap may be needed.

This study contributes to applied operations research by addressing two key research questions related to the aggregation of ESG data of Eikon Refinitiv,[1] one of the largest and most used ESG data providers. While our analysis focuses on Eikon Refinitiv, we believe similar challenges may exist with other providers. Refinitiv's full transparency regarding the construction of its ESG scores makes it an ideal case for investigation.

First, we examine the impact of Refinitiv's scoring methodology on key association measures, distinguishing between binary and continuous KPIs while addressing the issue of missing data. Specifically, we go beyond the Pearson correlation coefficient by considering other association measures like Kendall's tau, Spearman's rho, polyserial and tetrachoric correlations. We demonstrate, both theoretically and empirically, that replacing missing data with zeros leads to inflated correlations between KPIs. This practice, particularly when transitioning from raw KPIs to mapped KPIs used in score computation, can alter the dependence relationships significantly.[2]

Second, we investigate whether all KPIs contribute distinct information or if there is redundancy among them. To address this, we propose an optimization tool specifically designed to detect redundancy in the data. We formulate KPI subset selection as a cardinality-constrained rank correlation maximization problem and address it using a tailored stochastic hill-climbing with random restarts. Accordingly, it enables a systematic evaluation of each KPI's informational contribution. Our analysis shows that only about 20–25% of KPIs are needed to build pillar scores that closely replicate those of Refinitiv, which typically aggregate over 40 KPIs per pillar, selected from more than 650 KPIs. However, we do not find a consistent set of KPIs always selected to replicate the pillar scores.

These two research questions are deeply connected because the existence of real and methodology-induced correlation among KPIs introduces redundancy, making it possible to replicate the scores with a fraction of the KPIs. Our theoretical framework allows us to better understand how the scoring process may distort or preserve the natural dependence between KPIs, both with and without missing values. By identifying potential redundancies, we aim to enhance the efficiency and interpretability of ESG scores while preserving their relevance and accuracy.

To the best of our knowledge, Billio et al. (2024) represents the most closely related work to ours. However, while Billio et al. (2024) compares ESG ratings across four providers and identifies the KPIs that have the most influence on these ratings, considering only Pearson's correlation, our study takes a different approach by focusing on a single data provider. Instead of comparing multiple providers, we examine the internal structure of Refinitiv's ESG scoring system. Our analysis provides a theoretical and empirical framework for evaluating the impact of imputing missing data with zeros on the dependence structure between KPIs. We show that this imputation inflates correlations (i.e., Pearson, Kendall's tau, Spearman's rho, polyserial and tetrachoric correlation coefficients) and alters the relationship between KPIs.

The novelty is not that rank correlations are invariant under strictly increasing transformations. Our innovation is to show that within Refinitiv's scoring framework, combining percentile ranking with handling of missing data (notably, imputing missing raw KPIs with

[1] Eikon Refinitiv, along with the broader Refinitiv business, was acquired by the London Stock Exchange Group (LSEG) in January 2021. In the paper, we will still refer to Refinitiv.

[2] We define "raw" KPIs the actual scoring variables data available in Refinitiv with missing values [i.e., the x_i in Eqs. (1) and (2) of Sects. 2.1 and 2.2], while we define mapped KPIs the scores obtained after applying Refinitiv's percentile ranking to the "raw" KPIs (i.e., the $f(x_i)$ and $g(x_i)$ in Eqs. (1) and (2)). Consider for example the KPI Policy Emissions Score: the raw value is a TRUE if the company has a policy on emissions and a FALSE otherwise; the mapped value is a number in the 0–100 range determined applying percentile ranking. See the example in Appendix A.

zeros), violates the conditions under which rank invariance would carry over from raw to mapped data. In the empirically relevant case with missingness, the methodology can destroy rank invariance on the full sample and may even reverse the sign of dependence, as shown by our theoretical contribution.

A second contribution is our unified treatment of dependence in mixed-type ESG data. We analyze continuous-continuous, binary-binary, and mixed KPI pairs, distinguishing measures preserved by mapping under complete data (Kendall/Spearman; tetrachoric for binary pairs) from measures that can change even without missingness (Pearson and polyserial). To the best of our knowledge, these preservation-versus-distortion results for Refinitiv's scoring framework have not been discussed from this angle before.

Our contribution is also not to propose a new ESG scoring model. Rather, we use Refinitiv's own transparent construction to show that its pillar scores can be closely replicated with a small subset of KPIs. In other words, the mapping from hundreds of inputs to published pillar scores is highly compressible: only about 20–25% of KPIs are sufficient to match the provider's scores with high accuracy. This provides direct evidence of substantial redundancy in the information set used by the provider and highlights that score outcomes are driven by a limited number of underlying measurement dimensions, even when the original methodology aggregates dozens of KPIs per pillar.

Our findings have significant implications, as we demonstrate that using Pearson's correlation, the most widely used dependence measure, might result in biased assessments of relationships. Despite this, many ESG studies rely heavily on Pearson's correlation to capture dependence, particularly in regression models for financial performance, risk, and related areas. Further, we offer practical guidance on constructing ESG scores that are both transparent and representative, built considering a small number of KPIs, making them more suitable for managerial decision-making.

The rest of the paper is structured as follows. In Sect. 2, we review the methodology Refinitiv uses to produce the scores starting from the raw values of the KPIs. In Sect. 3, we describe our datasets; in Sect. 4, we discuss the impact of Refinitiv's methodological choices on association measures from a theoretical perspective, while in Sect. 5, we discuss this from an empirical perspective. In Sect. 6, we discuss the process of selection of KPIs, and finally, we report our conclusions in Sect. 7.

2 Refinitiv's ESG scoring methodology overview

A company's ESG score is calculated as a weighted average of its Environmental, Social and Governance scores, with weights different for each sector and disclosed by Refinitiv. Each pillar score is calculated as a weighted average of its category scores (three for Environmental and Governance, and four for Social). The calculation of category scores is more complicated and based on relative performance against sector peers (for Environmental and Social) or country peers (for Governance).

Refinitiv collects (potentially) over 630 data points for each company. From these data points, it calculates and reports up to 186 raw KPIs, or scoring variables, depending on the sector in which the company operates. For instance, the KPI "Agrochemical products" is reported only for two sectors - Chemicals and Diversified industrial wholesale goods - while the KPI "Policy emissions" is reported for all sectors. Each KPI influences only one category score.[3] Then, raw values of KPIs are mapped into a score in the [0–100] range using a

[3] We provide the list of 186 KPIs in our online supplementary.

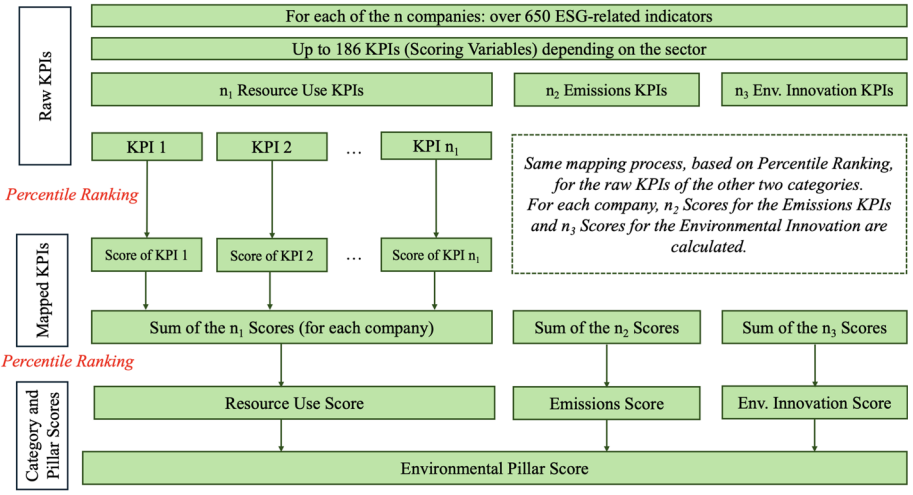


Fig. 1 Example of the flowchart for calculating Refinitiv's Environmental pillar score (from top to bottom)

methodology called percentile ranking, which we describe in detail in the next sections. This process is repeated for all KPIs classified under a given category. Then, for each company, the sum of all these scores is calculated, and a percentile ranking is applied to this sum. The score obtained in this way will be the category score. The pillar-specific category scores are combined as a weighted average, where weights are sector-specific and based on Refinitiv's materiality matrix. This process is shown for the Environmental pillar in Fig. 1, and we provide an example for score calculations in Appendix A.

2.1 Continuous KPIs

The following methodology is used to determine the scores of continuous KPIs (starting from their raw value) and to calculate category scores (starting from the sum of their mapped KPIs). Assuming positive polarity for the KPI under analysis, that is, the higher the raw value, the better the ESG performance, let $\boldsymbol{x} = (x_1, \ldots, x_n)^\top$ be a vector of raw values of the KPI for n companies with $x_i \in \mathbb{R} \cup \{\text{"NA"}\}$ for $i = 1, \ldots, n$. The Refinitiv's percentile ranking to calculate the corresponding mapped KPI is given by the function f, as follows: for any x_i and x_j,

$$f(x_i) = \begin{cases} \text{"NA"}, & \text{if } x_i = \text{"NA"}, \\ \left(\dfrac{\sum_{x_j \in \mathbb{R}} I(x_j < x_i) + \sum_{x_j \in \mathbb{R}} I(x_j = x_i)/2}{\sum_{x_j} I(x_j \in \mathbb{R})} \right) \times 100, & \text{otherwise}, \end{cases} \tag{1}$$

where "NA" denotes non-real-valued observations, i.e., missing values, $I(\cdot)$ is the indicator function, $\sum_{x_j \in \mathbb{R}} I(x_j < x_i)$ counts the number of real-valued observations of the vector $\boldsymbol{x}$ that are smaller than x_i, and $\sum_{x_j \in \mathbb{R}} I(x_j = x_i)$ counts the number of real-valued observations of the vector $\boldsymbol{x}$ that are equal to x_i, including itself.

The first case of Eq. (1) specifies that if a continuous raw KPI contains missing values denoted as "NA", they are encoded as "NA" in the corresponding mapped KPI. If this formula is used to calculate a category score, "NA" also includes the companies whose sum of the scores is equal to 0 (i.e., those companies that received a score of 0 for all the category-

specific scoring variables). Real-valued continuous raw values are transformed into their corresponding mapped values, as outlined in the second case of Eq. (1). We remark that only the companies with nonzero real-valued raw KPIs are considered in the denominator in the second case of Eq. (1).

Although missing values are encoded as "NA", in practice, they are treated as zeros. In fact, these "NA" are then excluded from the sum of the mapped KPIs' scores to determine the category score. The exclusion is equivalent to adding a zero. Since we are interested in using mapped values to determine the correlation between KPIs and to replicate the scores, we will treat these "NA" as zeros.

In case the KPI under analysis has negative polarity, that is, higher raw values imply worse ESG performance, the term $\sum_{x_j \in \mathbb{R}} I(x_j < x_i)$ will become $\sum_{x_j \in \mathbb{R}} I(x_j > x_i)$. Both for positive and negative polarity KPIs, this term represents the number of peer companies with performance worse than company i.

This scoring procedure ensures that scores of companies with non-missing observations are evenly spread in the [0-100] depending on their relative position against their peers, while scores of companies with missing observations are 0.

2.2 Binary (boolean) KPIs

Assuming positive polarity for the binary, i.e., boolean, KPI under analysis, let $\boldsymbol{x} = (x_1, \ldots, x_n)^\top$ be a vector of raw values of the KPI for n companies with $x_i \in \{1(\text{TRUE}), 0(\text{FALSE}), \text{"NA"}\}$ for $i = 1, \ldots, n$. The Refinitiv's percentile ranking method to calculate the corresponding mapped KPI is given by the function g, as follows: for any x_i and x_j,

$$g(x_i) = \begin{cases} 0, & \text{if } x_i = 0(FALSE) \text{ or } x_i = \text{"NA"}, \\ \left(\frac{\sum_{x_j \in \mathbb{R}} I(x_j < x_i) + \sum_{x_j \in \mathbb{R}} I(x_j = x_i)/2}{\sum_{x_j} I(x_j \in \mathbb{R} \cup \{\text{"NA"}\})} \right) \times 100, & \text{otherwise}, \end{cases} \tag{2}$$

where the representation of each term is given under Eq. (1).

In contrast to the continuous KPIs' transformation from the raw to mapped values in Eq. (1), missing and zero values of binary raw KPIs are encoded as zeros in the mapped KPI, as seen in the first case of Eq. (2). We will show theoretically in Sect. 4 and empirically in Sect. 5 that such zero encoding impacts association measures among KPIs. Otherwise, a percentile ranking scheme is applied, described in the second case of Eq. (2). Still, all companies are considered in the denominator in the second case in Eq. (2) in contrast to Eq. (1).

In the case of a negative polarity variable, the way in which true and false are coded into 0, 1 values is reversed, so $x_i \in \{0(\text{TRUE}), 1(\text{FALSE}), \text{"NA"}\}$.

3 Data

We focus our analyses on three sectors: (i) Machinery, Tools, and Heavy Vehicles, Trains and Ships (from now on Machinery), (ii) Oil & Gas (from now on Oil) and (iii) Chemicals. These sectors were selected because of their large sample size and active engagement in E and S pillar scores. We have ten separate datasets for each of these sectors, one for each

Table 1 Number of companies in the sample by sector and year

Sector\year	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021
Machinery	169	172	176	211	243	281	330	386	470	517
Oil	125	133	137	150	165	186	196	215	240	246
Chemicals	122	126	129	148	156	172	200	232	276	316

year in 2012–2021.[4] A company is included in a dataset if, in the year of reference, Refinitiv reported an ESG score for the company. Our datasets include the raw values of the KPIs, with the corresponding mapped KPIs, the category scores, the pillar scores, and the ESG scores. Table 1 shows the number of companies included in the datasets each year and sector.

Most of the results we will present in this paper will refer to the Oil sector in 2012, which we take as a benchmark. Yet, the presented results hold for the other years and sectors, and a summary of the results for other years and sectors is available on request.

The scatter plot in Fig. 2 shows the estimated correlation values[5] between the 150 KPIs and both the overall ESG score (x-axis) and their respective pillar scores (y-axis) for the 125 companies in the Oil sector in 2012. We observe that the correlation between individual KPIs and the pillar score they belong to is generally in line with their correlation with the overall ESG score. The consistency between these estimated correlations might imply that, at least for most KPIs, their impact on pillar scores is well-reflected in the aggregate ESG score.

The KPIs that have the highest estimated correlation with the pillar scores and the overall ESG score are predominantly binary and belong to the E and S pillars. Therefore, binary indicators within the pillars are more strongly associated with ESG performance than other variables. Such a result suggests that companies are more likely to score highly on ESG data based on whether they have certain policies in place rather than continuous performance measures. On the other hand, several KPIs within the G pillar exhibit low and/or negative correlations with the G pillar and overall ESG scores. It implies that Governance KPIs may not reflect a company's overall sustainability performance as clearly as E and S pillar KPIs. However, given that the G pillar and its category scores are assigned based on the companies' country, such a result might be expected.

4 Theoretical analysis of association measures

We now establish how Refinitiv's percentile ranking transformation and handling of missing values change five association measures, thereby grounding the KPI selection optimization in Sect. 6.

The bivariate association measures we consider include Pearson's correlation (ρ), tetrachoric correlation (ρ_t) (Castellan, 1966), polyserial correlation (ρ_p), Kendall's tau (τ) (Kendall, 1938), and Spearman's rank correlation or rho (ρ_s) (Spearman, 1904), The last two measures capture monotone association, which is a dependence of how much one variable tends to increase (or decrease) if the other increases. Thus, the association considered does not have to be linear. Nevertheless, Pearson's correlation is not a monotone association

[4] The datasets were all downloaded at a single point in time in December 2023. Thus, the scores of all years are calculated with the same methodology.

[5] If the KPI is continuous, the correlation reported is the Pearson correlation coefficient. If the KPI is binary, the correlation reported is the Polyserial correlation coefficient (Olsson et al., 1982).

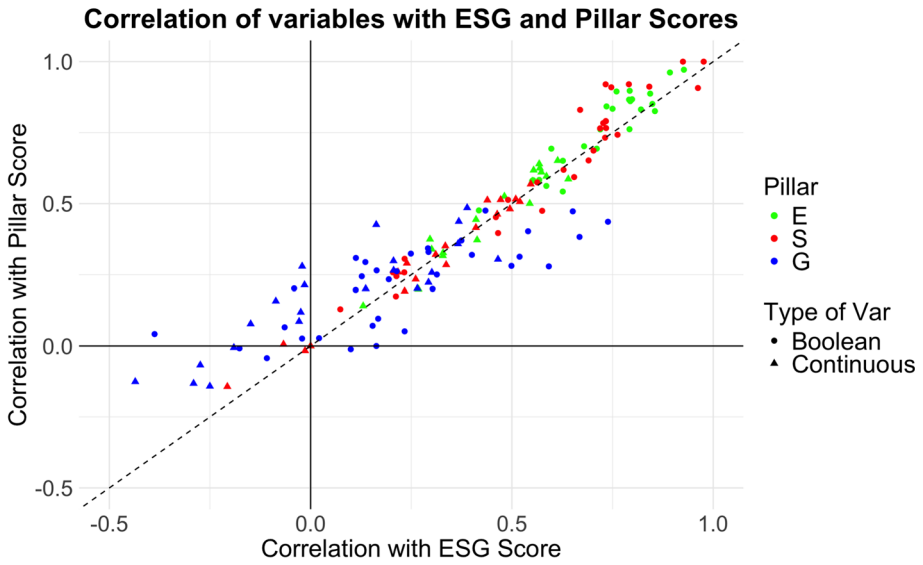


Fig. 2 The scatter plot of the estimated correlation (Pearson for continuous-continuous or polyserial for continuous-binary) of all mapped KPIs with the ESG score (x-axis) and their corresponding pillar (y-axis) in the Oil sector in 2012. The color of the points represents the pillar to which the variables belong, where green, red, and blue points correspond to the E, S and G pillars, respectively. The shape represents the variable type, with a circle representing boolean KPIs and a triangle representing continuous KPIs. Similar results apply to other years and sectors

measure but a linear association measure. In the next sections, we will discuss the disadvantages of using Pearson's correlation as an association measure on ESG data. Still, Pearson's correlation is the most popular measure and is widely used in the literature.

Further, in our empirical analyses, we estimate the polyserial correlation for the association between a continuous and binary variable and the tetrachoric correlation between two binary variables. Therefore, whenever a correlation is estimated in our empirical analyses, it refers to Pearson's correlation (ρ) for a pair of continuous variables, polyserial correlation (ρ_p) for a continuous and binary variable, and tetrachoric correlation (ρ_t) for two binary variables. We refer to Chapter 2 of Joe (2014) for more details on association measures and define them in Appendix B.

We present the impact of Refinitiv's methodology on the association measures between an ESG score of interest and raw and mapped KPIs with and without missing data, such as between the mapped KPI "Policy Emissions" and E pillar score, in Table 5 in Appendix B and prove them next. We observe that Pearson and Polyserial correlations are not the same in both cases, and all of them are different when imputing zero for missing data. Thus, the dependence structure among KPIs before and after the mapping process already changes for Pearson and Polyserial correlation when missing data are not present, and all the dependence measures change when missing values are imputed as zero.

In the following, assuming positive polarity of the given ESG KPI, let $\boldsymbol{x} = (x_1, \ldots, x_n)^\top$ be a vector of raw KPI of n companies.

Proposition 1 *The functions of Refinitiv's percentile ranking given in Eqs.* (1) *and* (2) *are strictly increasing, i.e., a monotone transformation, for real-valued inputs.*

Proof Let $\boldsymbol{x} = (x_1, \ldots, x_n)^\top$ be a vector of real-valued ESG raw KPI observations of n companies. We show that if $x_i < x_j$, then $f(x_i) < f(x_j)$, and if $x_i = x_j$, then $f(x_i) = f(x_j)$. Suppose $x_i < x_j$. Let m be the number of observations in $\boldsymbol{x}$ that are strictly smaller than x_i, let k be the number of observations equal to x_i, let q be the number of observations strictly between x_i and x_j, and let l be the number of observations equal to x_j. Then $f(x_i) = \left(\frac{m+0.5k}{n}\right) \times 100$, $f(x_j) = \left(\frac{m+k+q+0.5l}{n}\right) \times 100$. Hence, $f(x_j) - f(x_i) = \left(\frac{0.5k+q+0.5l}{n}\right) \times 100 > 0$, since $k \geq 1$, $l \geq 1$, and $q \geq 0$. Therefore $f(x_i) < f(x_j)$. If $x_i = x_j$, then both $f(x_i)$ and $f(x_j)$ are computed using the same number of observations below and equal to that common value, so $f(x_i) = f(x_j)$. The same argument applies to the binary case for the function g in Eq. (2), which proves the proposition. □

4.1 Impact of the ranking methodology: no missing data

We start by analyzing the impact of Refinitiv's percentile ranking scheme on the association measures in the absence of missing data.

Proposition 2 *For any two real-valued continuous ESG raw KPIs, $\boldsymbol{x}$ and $\boldsymbol{y}$, the sample Kendall's tau $\hat{\tau}$ between them $\hat{\tau}(\boldsymbol{x}, \boldsymbol{y})$ is the same as the sample Kendall's tau between their associated mapped KPIs $\hat{\tau}(f(\boldsymbol{x}), f(\boldsymbol{y}))$, i.e., $\hat{\tau}(\boldsymbol{x}, \boldsymbol{y}) = \hat{\tau}(f(\boldsymbol{x}), f(\boldsymbol{y}))$, where f is given by Eq. (1). The same result applies to sample Spearman's rho between their associated mapped KPIs, i.e., $\hat{\rho}_s(\boldsymbol{x}, \boldsymbol{y}) = \hat{\rho}_s(f(\boldsymbol{x}), f(\boldsymbol{y}))$.*

Proof Kendall's tau and Spearman's rho measure bivariate monotone association and are invariant concerning strictly increasing transformations on the variables; the result follows from Proposition 1. □

Proposition 2 confirms that rank correlations are preserved by Refinitiv's percentile ranking transformation when data are complete, justifying our later emphasis on rank-based association, and in particular Spearman's rho, in the objective function in Sect. 6.1.

Proposition 3 *For any two binary ESG raw KPIs, $\boldsymbol{x}$ and $\boldsymbol{y}$, the sample tetrachoric correlation $\hat{\rho}_t$ between them is the same as the sample tetrachoric correlation between their associated mapped KPIs, i.e., $\hat{\rho}_t(\boldsymbol{x}, \boldsymbol{y}) = \hat{\rho}_t(g(\boldsymbol{x}), g(\boldsymbol{y}))$, where g is given by Eq. (2).*

Proof Under Refinitiv's percentile ranking in Eq. (2), the 2×2 contingency table of two binary variables does not change. Since the tetrachoric correlation is fully determined by that table, it remains identical. □

Proposition 4 *There are a binary ESG raw KPI $\boldsymbol{x}$ and a continuous raw KPI $\boldsymbol{y}$ for some n such that the sample polyserial correlation coefficient between them, $\hat{\rho}_s(\boldsymbol{x}, \boldsymbol{y})$, is not the same as the sample polyserial correlation coefficient between their corresponding mapped KPIs $\hat{\rho}_s(g(\boldsymbol{x}), f(\boldsymbol{y}))$, i.e., $\hat{\rho}_s(g(\boldsymbol{x}), f(\boldsymbol{y})) \neq \hat{\rho}_p(\boldsymbol{x}, \boldsymbol{y})$, where g and f are given by Eqs. (2) and (1), respectively.*

Proof There are many counter-examples; for one, see Appendix B. □

Proposition 5 *There are two continuous ESG raw KPIs $\boldsymbol{x}$ and $\boldsymbol{y}$ for some n such that the sample Pearson's correlation coefficient between them, $\hat{\rho}(\boldsymbol{x}, \boldsymbol{y})$, is not the same as the sample Pearson's correlation coefficient between their corresponding mapped KPIs $\hat{\rho}(f(\boldsymbol{x}), f(\boldsymbol{y}))$, i.e., $\hat{\rho}(f(\boldsymbol{x}), f(\boldsymbol{y})) \neq \hat{\rho}(\boldsymbol{x}, \boldsymbol{y})$, where f is given by Eq. (1).*

Proof There are many counterexamples; for one, see Appendix B. □

Proposition 5 emphasizes the limitations of using Pearson's correlation coefficient in ESG data analysis, which could have actual impact. For example, biased Pearson correlations could mislead ESG-based portfolio optimization. The transformation of raw KPIs into mapped KPIs while embedding a mean of 50 (Sahin et al., 2022) might significantly change the linearity of their relationship. Still, further investigation is necessary to understand this phenomenon comprehensively, especially because most studies focus exclusively on the Pearson correlation.

4.2 Impact of the ranking methodology: missing data

In this section, we extend our analysis of the impact of Refinitiv's percentile ranking scheme on the association measures to the case of missing data.

Proposition 6 *There exist two continuous ESG raw KPIs* $\boldsymbol{x}$ *and* $\boldsymbol{y}$ *with at least one NA value for some n such that the sample Kendall's tau, Pearson's correlation, and Spearman's rho estimated on the subset of observations where both* $\boldsymbol{x}$ *and* $\boldsymbol{y}$ *have non-missing values, i.e., their joint complete part, are not the same as the sample Kendall's tau, Pearson's correlation, and Spearman's rho estimated on their corresponding mapped KPIs, respectively.*

Proof There are many counterexamples; for one, see Appendix B. □

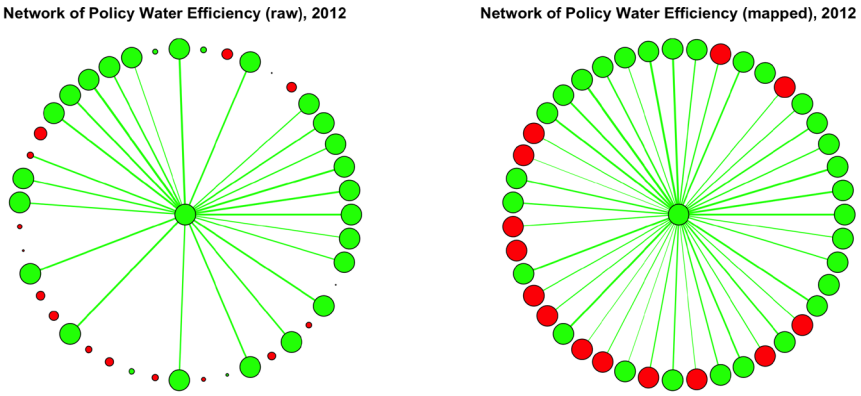
Proposition 7 *There exist two continuous ESG raw KPIs* $\boldsymbol{x}$ *and* $\boldsymbol{y}$ *with at least one NA value for some n such that the sample Kendall's tau, Pearson's correlation, and Spearman's rho between their joint complete observations have a different sign compared to the sample Kendall's tau, Pearson's correlation, and Spearman's rho between their corresponding mapped KPIs, respectively.*

Proof The proof of Proposition 6 applies. □

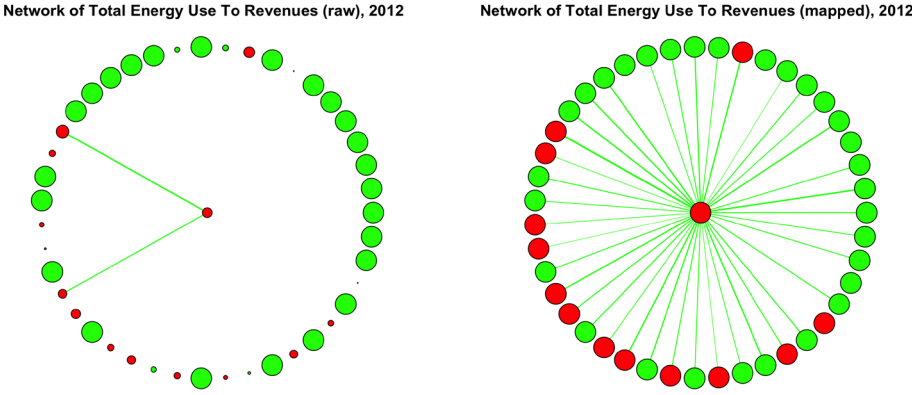
Similar statements can be extended to estimate polyserial and tetrachoric correlations in the presence of binary variables. Hence, the missing data encoding changes the strength of the bivariate monotone association. Still, we remark that if zero values are dropped from the analyses, considering them as missing, the raw and mapped KPIs have the same sample Kendall's tau, Spearman's rho, and tetrachoric correlations with other KPIs as shown in Propositions 2 and 3 contrary to Propositions 6 and 7. However, since zero is numeric, scholars likely include it in estimating association measures of ESG data. To sum up, the impact of Refinitiv's methodology on the association measures between KPIs and ESG data can be seen in Table 5.

5 Empirical analysis of association measures

We now focus on examining the association between raw KPIs and mapped KPIs through network visualization in Fig. 3. The Pearson, tetrachoric and polyserial correlation coefficients are reported for pairs of continuous variables, binary variables and pairs of continuous-binary variables, respectively. Figure 3 shows the network of KPIs under the Environmental pillar. In the left column, we show the network of correlations between the raw central KPI and the other raw KPIs. Conversely, the right column illustrates the network of correlations between



(a) The network of the KPI: Policy Water Efficiency.
Left: raw KPIs, right: mapped KPIs.



(b) The network of the KPI: Total Energy Use To Revenues.
Left: raw KPIs, right: mapped KPIs.

Fig. 3 The networks include the E pillar KPIs, focusing on their correlation with the KPI in the middle of the network. There is a boolean KPI (Policy Water Efficiency) in the upper networks and a continuous KPI (Total Energy Use / Revenues) in the lower networks. The networks on the left are based on the correlation in raw data, while the networks on the right are based on mapped data. The size of each node represents the number of observations; the node's color represents the variable's polarity (green = positive, red = negative). The edge indicates a statistically significant correlation between the nodes it connects. The data refers to the 125 companies in the Oil sector in 2012

the same central mapped KPI and the other mapped KPIs. If the central KPI exhibits a significant correlation (at a 5% level) with another KPI, then the central node, representing the central KPI, and the node of the other KPI appear connected with an edge. Otherwise, the two nodes are not connected. This analysis reveals several important observations.

First, the raw KPIs' networks feature several small nodes, indicating more missing data. In contrast, the mapped KPI networks show nodes of uniformly large size. Indeed, it reflects the filling of missing values with zeros in the mapped KPIs as shown in Eqs. (1) and (2). Several small nodes are not connected in the raw networks, but they become connected in the mapped network due to how missing data is handled. Second, the networks based on raw KPIs exhibit fewer edges (i.e., non-zero coefficients) than the mapped networks. This difference suggests that handling missing data in the mapping process introduces artificial correlations, potentially skewing the true relationships between KPIs as theoretically analyzed in Sects. 4.1 and 4.2. Third, numeric KPIs, which often have a higher rate of missing observations, tend to result in fewer statistically significant edges, particularly in the raw data.

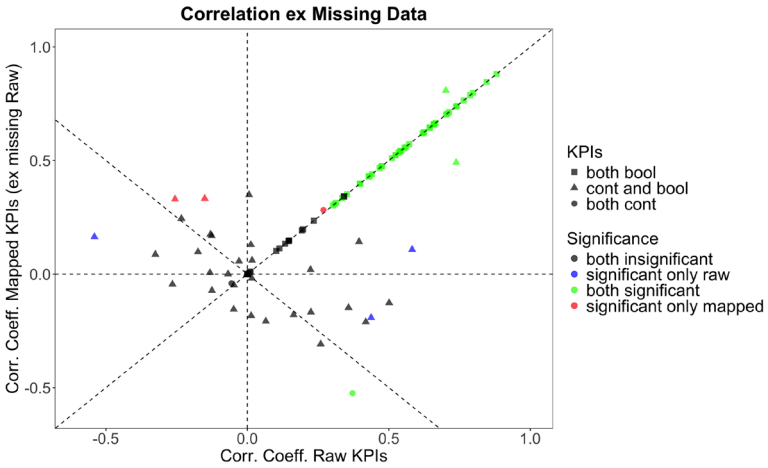
Figure 4 shows the impact of Refinitiv's percentile ranking to compute mapped KPIs from raw KPIs and highlights the impact of missing data on the sample correlation between KPIs. Each point in the scatter plots of Fig. 4 represents a pair of KPIs, and its coordinates represent the estimated correlation coefficient between the two KPIs before percentile ranking (raw data) and after percentile ranking (mapped data).[6] The coordinates of each point in Fig. 4a are determined using only the subset of companies for which the raw values of the two associated KPIs are not missing, while in Fig. 4b the value on the y axis is determined using all companies.[7]

The color of the points represents the results of the statistical tests on whether the correlation coefficient is equal to zero. The two main colors in Panel (a) are black and green. Black points represent pairs of KPIs for which the correlation coefficient is not statistically different from zero, neither for raw data nor for mapped data (excluding the firms for which there are missing values in at least one of the two KPIs). Green points, on the contrary, represent those pairs of KPIs for which both correlation coefficients are statistically different from zero. Red (blue) points represent those pairs of KPIs for which the correlation coefficient is significantly different from zero only in the mapped (raw) data, indicating that the mapping process alters the correlation structure.

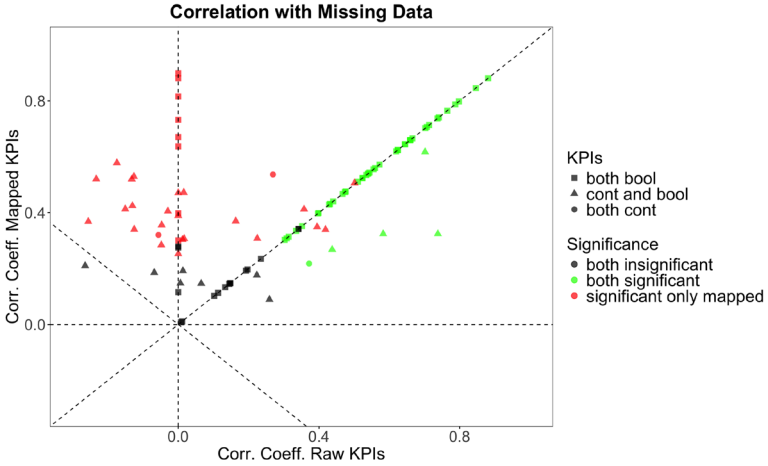
Panel (a), which is based solely on the observations that have no missing values in the plotted pair of KPIs, shows that the mapping process with percentile ranking usually does not alter the correlation structure between pairs of variables, a result in line with our theoretical analyses in Sect. 4.1. Indeed, most points lie on the $y = x$ line, indicating that the mapping does not change the value of the correlation coefficient. For binary versus binary KPI pairs, it is due to the measure chosen as shown in Proposition (3). Moreover, continuous versus binary KPI pairs have similar values before and after mapping. In any case, significance of correlations is rarely affected, with most correlations being not statistically significant both before and after. Many of these points also lay around the $y = -x$ line. In this case, the mapping process does not affect the absolute value of the correlation coefficient, but only its sign. This is due to the opposite polarity of the variables in the pair. Pairs of continuous KPIs also can lay off the two benchmark lines variables $y = x$ and $y = -x$, since the mapping process affects Pearson correlation. In general, the fact that most points of Panel (a) are black and green indicates that the significance of the correlation is unaffected by the mapping process.

[6] The Pearson, tetrachoric and polyserial correlation coefficients are reported for pairs of continuous KPIs, binary KPIs and mixed pairs, respectively.

[7] Thus, the correlation reported on the x axis is calculated using only the raw data of companies without missing data in the two variables, while the correlation on the y axis is calculated using the mapped data of all the companies. The mapped values of the missing observations are assigned the value 0 by Refinitiv's methodology.



(a) Scatter plots of the correlation coefficient between pairs of raw KPIs (x-axis) versus pairs of mapped KPIs (y-axis), where, for each pair, we exclude the observations with missing data in at least one of the two KPIs in the raw data.



(b) Scatter plots of the correlation coefficient between pairs of raw KPIs (x-axis) versus pairs of mapped KPIs (y-axis). No observation is excluded from calculating the correlation between mapped KPIs.

Fig. 4 Scatter plots of the correlation coefficient between pairs of raw KPIs (x-axis) versus pairs of mapped KPIs (y-axis). The shape of the points represents the type of KPIs in the pair. The color of the dots represents the significance of the correlation between the KPIs in the pair of raw and mapped data. All pairs of scoring KPIs being part of the Resource Use Category are plotted

However, how missing data is handled introduces an artificial correlation between pairs of KPIs. Unlike Panel (a), Panel (b) considers the whole set of companies, including those with missing data in at least one of the KPIs in the pair. The presence of several red dots in Panel (b) indicates that the correlation in raw data was not statistically significant for those pairs of KPIs, whereas it is significant in mapped data, as previously shown for two KPIs with the networks of Fig. 3. An explanation for this phenomenon is offered below.

Assume we are considering two uncorrelated raw KPI vectors, $\boldsymbol{M}_1$ and $\boldsymbol{M}_2$. However, suppose several companies have missing data for both KPIs. In that case, the sample correla-

tion in raw data will reflect the true structure, that is, the absence of correlation, but the sample correlation in mapped data will show some positive value since several companies with a score of 0 in $\boldsymbol{M}_1$ also have a score of 0 in $\boldsymbol{M}_2$. Similarly, several companies with a non-zero score in $\boldsymbol{M}_1$ will also have a non-zero score in $\boldsymbol{M}_2$. Figure 11 in the Appendix C shows that, in general, for any pair of KPIs $\boldsymbol{M}_1$ and $\boldsymbol{M}_2$ belonging to the same pillar, the unconditional probability that the value of $\boldsymbol{M}_2$ is missing is smaller than the probability that the value of $\boldsymbol{M}_2$ is missing conditional on the value of $\boldsymbol{M}_1$ being missing too. In other words, missing values are not random, but there tend to be clusters of companies with low missingness and clusters of companies with high missingness, which in turn produces an artificial correlation in mapped KPIs. We provide further illustration of missing data patterns in the Appendix C.

6 Optimal KPIs selection

As shown in Sect. 5, several raw KPIs exhibit a high correlation with each other. Moreover, most mapped KPIs exhibit a stronger correlation due to the *de facto* imputation of zero scores for missing values. Thus, some of the KPIs may be redundant in the creation of an aggregated pillar score and introduce unnecessary complexity. In this section, we propose an optimization tool designed to facilitate the optimal selection of KPIs for the replication of the Environmental and Social pillar scores.[8] This tool addresses the potential redundancy among KPIs and their respective pillar scores. Our formulation is a cardinality-constrained subset-selection problem: select k KPIs that maximize a rank correlation objective, an NP-hard task for which we use stochastic hill-climbing with random restarts.

6.1 Optimization problem: KPIs selection tool

Let x_{ij} denote the entry in the ith row and jth column of $\boldsymbol{x}$, which practically refers to the jth KPI for ith company. The corresponding pillar score for that company is denoted by y_i. Using a set of selected KPIs $\mathcal{I}$, i.e., $\mathcal{I} \subset \{1, \ldots, p\}$, we can construct a new pillar score z_i for a given company i. The construction is similar to Refinitiv's methodology[9] described in Sect. 2. Specifically, let z_i be the sum of selected KPIs in $\mathcal{I}$ for the ith observation and given by

$$z_i = \sum_{j \in \mathcal{I}} x_{ij}, \quad \text{for} \quad i = 1, \ldots, n. \tag{3}$$

We define the objective function, constraints, and initialization for selecting KPIs to maximize Spearman's rank correlation between the newly constructed scores $\boldsymbol{z} = (z_1, \ldots, z_n)^\top$, and Refinitiv's pillar scores $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$. Our optimization framework is specifically designed to maximize Spearman's correlation coefficient, which is a measure of rank correlation and better captures the strength and direction of monotonic relationships between variables. As shown in Proposition 2, the estimated Spearman's rho, $\hat{\rho}_s$, between the con-

[8] Since we work at a sector level, we also do not replicate the Governance pillar scores as they are calculated by grouping firms based in the same country. For the same reason, we do not replicate the ESG scores.

[9] For each company, Refinitiv calculates the sum of all category-specific mapped KPIs and applies percentile ranking to this sum to determine the category score. The pillar scores are then calculated as the weighted average of the three/four category scores, where the weights are based on the materiality matrix. With our approach, we directly aggregate pillar-specific KPIs because we aim to drastically reduce the redundancy.

structed and original pillar score, $\hat{\rho}_s(\boldsymbol{z}, \boldsymbol{y})$, is the same as $\hat{\rho}_s(f(\boldsymbol{z}), \boldsymbol{y})$,[10] where the function f corresponds to the percentile ranking transformation used by the Refinitiv and specified in Eq. (1). We remark that we can formulate our optimization problem for minimization by changing the sign of the objective function or using other association measures as well.

Next, we aim to select the variables that maximize (or minimize) the sample Spearman's rho between their sum and the original pillar score while constraining the number of selected KPIs to a predefined value k (e.g., $k = 5$). Our optimization problem is formulated as:

$$\max_{\mathcal{I}} \hat{\rho}_s \left(\sum_{j \in \mathcal{I}} \boldsymbol{x}_j, \boldsymbol{y} \right), \quad \text{subject to} \quad |\mathcal{I}| = k, \tag{4}$$

where $\hat{\rho}_s(\cdot, \cdot)$ represents the sample Spearman's rank correlation coefficient, $\boldsymbol{x}_j$ denotes the vector of the j-th KPI values across all observations (companies), and $\boldsymbol{y}$ is the vector of original pillar scores. The set $\mathcal{I}$ represents the subset of k selected KPIs that will be used to construct the new score for each company. Here, $|\mathcal{I}|$ denotes the cardinality (number of elements) in the set $\mathcal{I}$, ensuring that exactly k KPIs are selected.

Our optimization problem in Eq. (4) to select the optimal set of k KPIs out of p has a discrete search space with $\binom{p}{k}$ elements. Thus, the solution space is large, making finding the global optimum hard. To compute approximate solutions, we rely on a stochastic optimization method (stochastic hill-climbing with random restarts) outlined below.

Algorithm 1 Stochastic hill-climbing with random restarts for KPI selection in ESG data.

Require: $\boldsymbol{x}$, $\boldsymbol{y}$, k, max_iter
$\mathcal{I} \leftarrow$ initialize_with_random_KPIs
current_solution $\leftarrow \mathcal{I}$
current_score $\leftarrow$ objective_function($\mathcal{I}$, $\boldsymbol{x}$, $\boldsymbol{y}$)
best_solution $\leftarrow \mathcal{I}$
best_score $\leftarrow$ current_score
for iter = 1 to max_iter **do**
  neighbor_indx $\leftarrow$ randomly select an index from $[1, k]$
  candidate_pool $\leftarrow \{1, \ldots, p\} \setminus$ current_solution
  neighbor_KPI $\leftarrow$ randomly select an index from candidate_pool
  new_solution $\leftarrow$ current_solution
  new_solution[neighbor_indx] $\leftarrow$ neighbor_KPI
  new_score $\leftarrow$ objective_function(new_solution, $\boldsymbol{x}$, $\boldsymbol{y}$)
  if new_score > current_score ***OR*** exp(100 · (new_score − current_score)/iter) > runif(1) **then**
    current_solution $\leftarrow$ new_solution
    current_score $\leftarrow$ new_score
    if current_score > best_score **then**
      best_solution $\leftarrow$ current_solution
      best_score $\leftarrow$ current_score
    end if
  end if
end for
return best_solution

Mainly, stochastic hill-climbing with random starts begins with an initial solution (i.e., a random combination of k KPIs), and at each iteration, it replaces a KPI in the current

[10] This would not happen if the Pearson's correlation coefficient had been chosen. With Refinitiv's percentile ranking, the linear correlation between two measures is affected, but their rank correlation is not. In this way, we can pass from $\boldsymbol{z}$ to f($\boldsymbol{z}$) to account for the differences in the range of $\boldsymbol{z}$ due to the different number of KPIs selected k.

solution with a KPI not in the solution. Both KPIs are uniformly randomly selected from the set of KPIs in the current solution and not in the current solution, respectively. Then, the new candidate solution's objective function is assessed. The new solution is accepted either if its objective function is better than the current one or, in case it is worse, with some acceptance probability given by $\exp(100 \cdot (\text{new\_score} - \text{current\_score})/\text{iter})$. Hence, the closer the candidate is to the current solution in terms of objective value, the more likely it is to be accepted. This allows us to approximate the global optimum. We remark that we update the best solution only if the new solution has a higher objective function. Otherwise, while the new suboptimal solution will be the starting point of the subsequent iterations, the algorithm will keep track of the (old) best solution. For example, assume the following values: iter = 10, current_score $= 0.90$, new_score $= 0.89$. Then, the acceptance probability is given by $exp(100 * (0.89 - 0.9)/10) = 0.37$. Thus, with probability 0.37, the new solution becomes the current solution explored in the next iteration, but the incumbent best solution is unchanged because $0.89 < 0.90$.

Our algorithm continues until the predefined number of iterations, max_iter, is reached. This number is given by $k \cdot p$. The process just described represents one run of the stochastic hill-climbing algorithm. For a predefined number of selected KPIs k, we repeat this for 300 runs, each of which will identify one best solution. Even though a different number of runs could be analyzed, 300 restarts stabilized the empirical objective function values as reflected with low variance in the green boxplots in Fig. 7. From a computational perspective, one restart has $1 + kp$ objective function evaluations. Therefore, the full procedure uses $300(1 + kp)$ evaluations for a given k. Since each evaluation involves computing the candidate score and estimating Spearman's rho across n firms, the runtime scales linearly with the number of restarts and neighborhood evaluations, making it practical for our sector and year analyses.

In Fig. 5, we plot the simulated E pillar scores of the best combination of KPIs (i.e., the combination with the highest Spearman's rho with the original scores) against the actual E pillar for different numbers of KPIs selected. As the number of KPIs increases, the simulated pillar scores should converge towards the actual ones. We see this in Fig. 5, indicating that the ability of the simulated scores to mimic the actual scores improves when the number of considered KPIs, k, increases. Further, even with relatively few variables, the simulated scores are very close to the actual ones. However, Refinitiv's methodology relies on a materiality matrix to assign different weights to the sets of KPIs. Since such weights are not considered in our optimization framework, i.e., we consider equal weights, we would not expect to produce the actual scores with perfect accuracy.[11] Still, our results show that the combination of a few KPIs can already produce a remarkable similarity with the original score. Thus, the choice of the number of KPIs and their choices seem to play a more significant role in determining the final score than the weighting system applied. Also, it highlights the effectiveness of our framework in replicating the actual E pillar scores through an equal-weighted aggregation process with just k KPIs, for k small.

6.2 Empirical analysis: finding the optimal KPIs set

We run the optimization formulated in Eq. (4) on the data sets described in Sect. 3. We run it for every year in 2012–2021 for all three sectors and a different number of KPIs. Here, we

[11] It should be noted that in practice in the computations of the category scores all KPIs have the same weight. Then, when different category scores are aggregated into a single pillar score, these 3 or 4 category scores receive different weights. However, these weights can sometimes be close to each other. For instance in the oil sector the weights of the three environmental categories are about 30%, 30% and 40%.

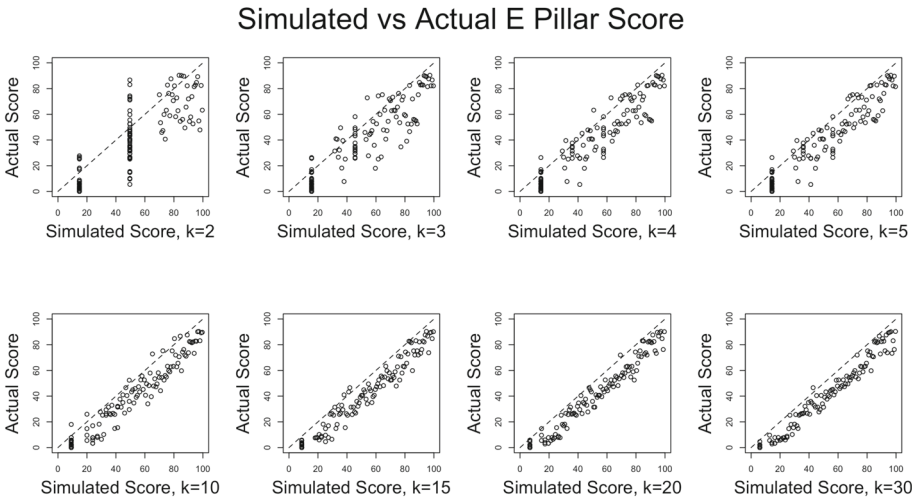


Fig. 5 The comparison of simulated E pillar score, which consists of the KPIs whose combination gives the highest estimated Spearman's rho with the original scores, with the actual E pillar score for a given number of KPIs entering the combination. As k increases, the simulated scores exhibit a pattern ever closer to the real scores. For comparability, the simulated scores were rescaled using the percentile ranking given in Eq. (1)

report the results of the Environmental pillar for the Oil sector in 2012, which includes 125 companies. The results reported for a single year and sector also hold for the other years and sectors; and a summary of the results for other years and sectors is available on request.

As an objective function, we use both maximization and minimization of the estimated Spearman's rho to analyze the maximum and minimum association one can have using k KPIs with the original pillar scores.

6.2.1 Maximizing correlation

First, we focus on maximizing the estimated Spearman's correlation of the constructed scores with the Environmental pillar score. The green lines in Fig. 6 provide a visualization of two different estimated association measures (i.e., Kendall and Spearman[12]) with the number of KPIs used for replication. The results refer to the best combination found in the 300 runs of the stochastic hill-climbing (each of which had hundreds of iterations). Kendall's tau correlation coefficient is significantly lower than Spearman's. This is consistent with the fact that our objective is with respect to Spearman's correlation coefficient. As we increase the number of KPIs used, the estimated correlation tends to plateau when using k =5 to k = 7 KPIs, though there is still a slight improvement with additional KPIs. Further, with just 15 KPIs, we achieve an almost perfect estimated correlation for both Pearson and Spearman coefficients. Thus, using more than 15 KPIs might not be necessary to construct the aggregated E pillar scores, which currently, with Refinitiv's approach, rely on 44 KPIs for the sector under analysis.

In Appendix D, we extend the analysis to all the 300 solutions found, not just the best, showing that with only 4 KPIs, the estimated Spearman's correlation coefficient of more than 75% of the solutions is above 90%, and with 9 KPIs, it is above 95% for over 75% of the

[12] The Pearson correlation coefficient is practically equal to the Spearman's in this context, indicating a strong alignment between these two measures of association. Since Spearman's rho is a rank-based counterpart to Pearson's rho, such a result is expected.

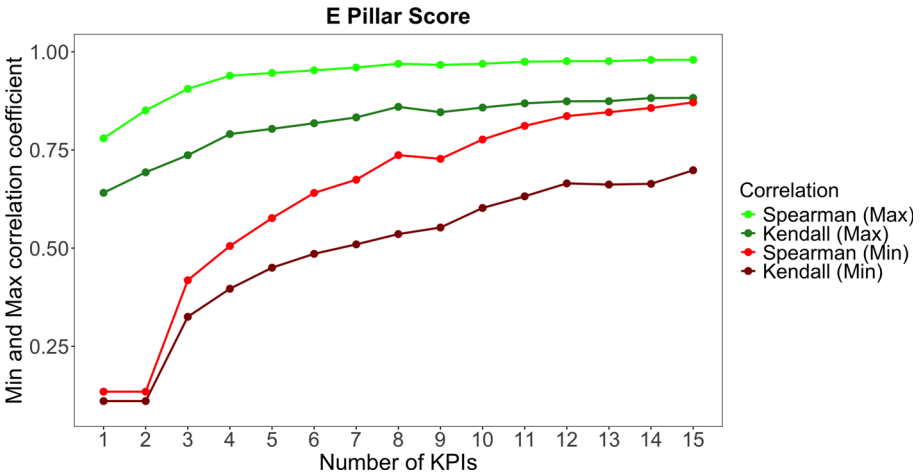


Fig. 6 The plot of the number of KPIs, k, on the x-axis and the highest (in green) and lowest (in red) correlation across the 300 runs of the stochastic hill-climbing algorithm (to maximize/minimize the estimated Spearman's rho as an objective in Eq. (4)) on the y-axis. The number of iterations equals the total number of selected KPIs times the total number of KPIs ($k \times p$) in each run. The results refer to the Oil sector in 2012 for the E pillar

solutions. Therefore, both the analysis of the best combination and the analysis of the overall results imply a significant redundancy of information: out of the 44 Environmental KPIs, less than 10 could be enough to replicate the score. In addition, using fewer KPIs could simplify the interpretation of the results. Further, fewer KPIs might make the scoring process more manageable, as the new ESMA regulation envisioned.

6.2.2 Minimizing correlation

Next, we minimize the estimated Spearman's rho of the constructed scores with the Environmental pillar score. The results are shown in Fig. 6 (red lines). Since all KPIs positively correlate with the Environmental pillar score regarding Spearman's rho, the lowest achievable estimation is positive. Further, although the purpose is to minimize the correlation between the original and constructed scores, combining just four KPIs already leads to a correlation higher than 0.50 in Fig. 6. This result matches the previous findings that there is significant redundancy in the KPIs and that even considering the worst combinations (in terms of the estimated Spearman's rho of the constructed scores with the Environmental pillar score) results in a high estimated correlation.

In Fig. 7, we show the distribution of the estimated Spearman's rho between the constructed scores obtained using a given set of combinations with k KPIs and the original pillar scores by Refinitiv. We see that for each k, we have three boxplots showing the distributions of the estimated correlation when the KPIs were selected to minimize it (in red), maximize it (in green), and when they were randomly selected (in yellow). For a given value of k, the distributions are in line with the order we would expect. The estimated correlations are higher when our objective was their maximization and lower when their objective was minimization. Random combinations fall in between. These results indicate that although there is redundancy in the information conveyed by the different KPIs because they are all correlated with each other, the choice of the KPIs is critical. As the number of KPIs considered

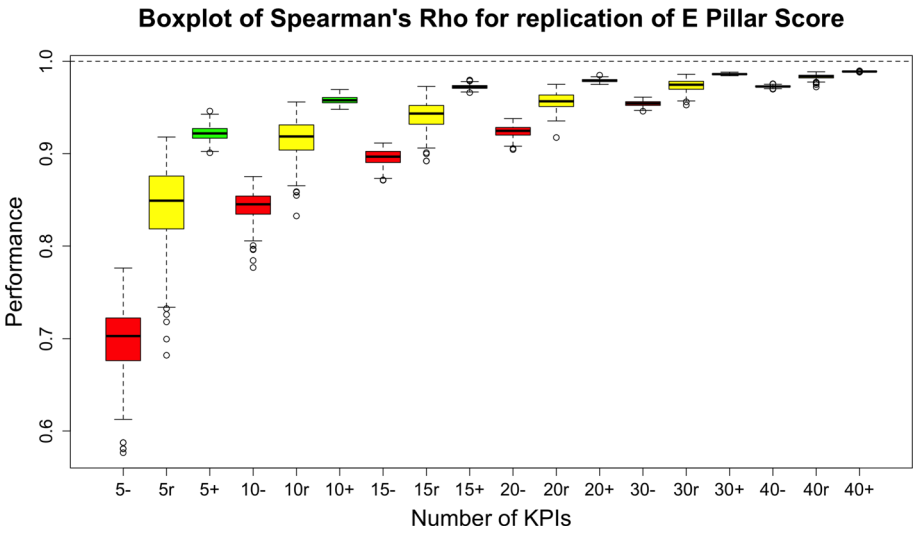


Fig. 7 The results of the 300 runs of stochastic hill-climbing with random restarts to maximize (in green, x-axis label "$k+$") and minimize (in red, x-axis label "$k-$") the estimated Spearman's rho between a constructed score and original score by Refinitiv, as well as the estimated Spearman's rho between the randomly selected k KPIs (in yellow, x-axis label "kr"), and given score. It refers to replicating the E-pillar score in the Oil sector in 2012

increases, the order is still respected, but the gap between maximization, randomization, and minimization reduces significantly.

Figure 7 also provides an empirical convergence analysis across random restarts. For each k, the runs with the objective of maximization have smaller empirical variance in their final performance, i.e., different initial solutions lead to very similar objective values. We can therefore interpret the algorithm as converging to good local optima for our analyses, while still not claiming guaranteed convergence to the global optimum.

In Appendix E, we investigate the presence of multimodality in the KPI selection process, which occurs when different combinations of KPIs yield the same level of estimated correlation with the actual scores.

In Appendix F, we perform a robustness check on the solutions found by the algorithm by performing an out-of-sample analysis. In particular, we use the optimal set of KPIs found using the dataset of a given year to reproduce a score using the actual mapped scores of those KPIs in a different year. The results indicate that the solutions found in a given year still produce correlations with the actual scores above 90% several years later.

6.3 Empirical analysis: computational efficiency

In this section, we analyze the computational efficiency of the algorithm and of some variants, comparing efficiency and performance.

First, we consider the procedure described in the previous sections as a baseline. Second, we consider the same procedure applied to a small dataset that includes only those KPIs with a missingness lower than 5% (Ex missing). Since the number of iterations is proportional to the number of KPIs, reducing the number of eligible KPIs would reduce the iterations, making the results not directly comparable. So, our third procedure is to use the smaller

Table 2 Average completion time (expressed in seconds) of one run of the algorithm given in Algorithm 1 by number of KPIs selected k and method

Method	5 KPIs	10 KPIs	15 KPIs	20 KPIs
Baseline	0.035	0.084	0.138	0.221
Ex missing	0.018	0.047	0.088	0.173
Ex missing robust	0.034	0.086	0.162	0.324
Weighted	0.119	0.348	0.689	1.153
Weighted ex post	0.036	0.084	0.139	0.222

dataset, but keep the same number of iterations of the baseline (Ex missing robust). Our fourth procedure tests different weights for KPIs. In particular, instead of applying equal weights, we apply to each KPI the weight that Refinitiv applies to the category to which that KPI belongs (Weighted).[13] The weighting occurs at each iteration when the z_i is calculated as a (now weighted) sum of k KPIs.

Finally, we consider another alternative where we take the KPIs from the baseline solutions, and, instead of combining them with equal weight, we apply Refinitiv's weights (Weighted ex post). In this way, the optimization process occurs using equal weights, but the KPIs obtained are weighted ex post based on their relative importance in Refinitiv's scoring methodology.

Table 2 shows the average time required to complete one run of the algorithm for each method, considering combinations of 5, 10, 15, and 20 KPIs. Increasing the number of KPIs increases proportionally the number of iterations per run, requiring more time to complete one complete run. Further, the process of selecting the KPIs and adding them requires more terms to be added together, making the increase in time required more than proportional.

Compared with the baseline, as expected, using a smaller dataset reduces the number of iterations, which implies a shorter completion time. Increasing the number of iterations to match those in the baseline results in an average completion time slightly higher than that in the baseline. The method of applying different weights appears much slower due to the need to identify the weight to apply to each selected KPI and then calculate the weighted sum of the KPIs. On the contrary, the ex post weighting only marginally increases computational with respect to the baseline because weights are applied only once at the end, instead of at each iteration.

In terms of performance, Fig. 8 shows the Spearman's correlation coefficient between the true scores and the scores produced through the 300 runs per method, sorted from the worst to the best. For $k = 5$, it seems that the reduction in the number of iterations resulting from the smaller datasets produces a small but noticeable reduction in the final correlation (green line in the upper left subplot). The difference in terms of performance for the other four methods (which all have the same number of iterations) is negligible. For a large enough number of selected KPIs, the lines appear mostly overlapped and flat. Indeed, in Fig. 8 the range of the y axis reduces from 5% when $k = 5$ (top left subplot) to just 1% when $k = 20$ (bottom right subplot). This means that for $k = 20$, the difference between the best and worst solution

[13] Since Refinitiv uses equal weights for all KPIs in the calculation of the Category Score, and then determines the Pillar Scores using different weights for the different Category Scores, there is not an explicit weight associated to each KPI in the determination of the Environmental Pillar Score. With our method we use the relative importance of the Category to which a KPI belongs to proxy for the weight of that KPI.

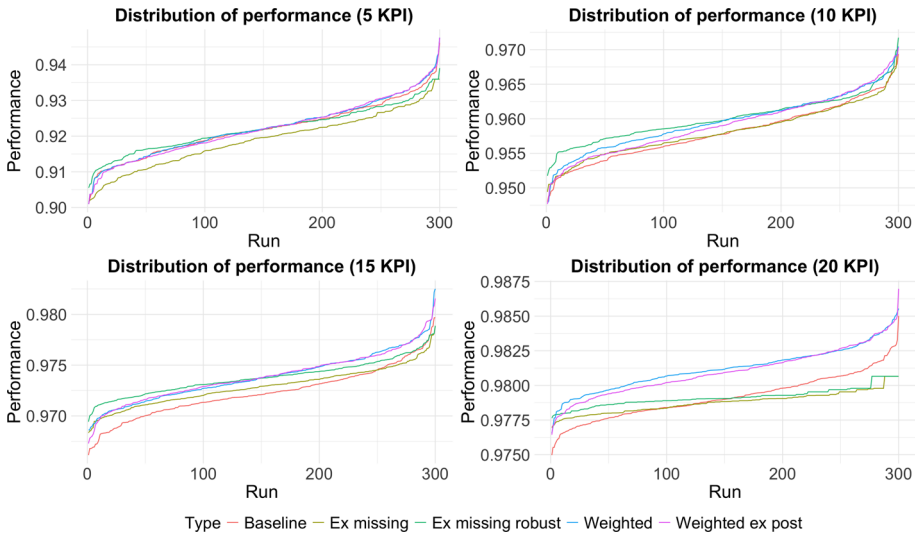


Fig. 8 The results of the 300 runs of stochastic hill-climbing given in Algorithm 1 sorted by performance. Each subplot shows a different number of KPI selected: 5, 10, 15, and 20. For each subplot, the performance of the 300 runs, sorted from worst to best, for the five methods is shown

for all 5 methods is barely above 1%. Despite the similarity in performance, it appears that the highest performance is obtained with the weighted method, but at a high cost in terms of completion time. The ex post weighting produces a performance level in line with the weighting, but it is much faster.

These results pertain to the Environmental pillar for the Oil sector in 2012; variations occur across other years, pillars, and sectors. In the Chemicals sector, Refinitiv assigns equal weights to the Environmental KPIs, resulting in no differences among the Baseline and the two Weighted methods. These equivalent approaches generally outperform the two methods based on the exclusion of high-missingness KPIs, particularly for sufficiently large k. However, performance in the different methods is very close, with the average distance between curves within 1%. In contrast, within the Machinery sector, the Weighted and Weighted ex post methods underperform the Baseline by about 3.5%, while the two Exclusion methods are generally superior by about 1%.

For the Social pillar, across all three sectors, both Weighted approaches yield inferior performance compared to the Baseline: average underperformance is 1.8% and 2.1% for the Weighted and the Weighted ex post methods, respectively. Conversely, excluding the KPIs with higher levels of missingness tends to improve performance relative to the Baseline by about 0.8% (Ex missing) and 1% (Ex missing robust).

Finally, as the sample size increases over time, the number of firms in the dataset grows, leading to longer computation times in line with the increased sample size.

Overall, when jointly considering performance and computational efficiency, the Baseline method emerges as the most effective method for the Environmental pillar. Applying weights ex post yields only marginal performance gains in two of the three sectors analyzed, while leading to a deterioration in the third. For the Social pillar, the Ex missing method appears to be the most consistent across all sectors. Although the Ex missing robust variant, characterized by an increased number of iterations, provides some performance improvements, these gains are marginal and do not appear to be worth the increased cost in computational time.

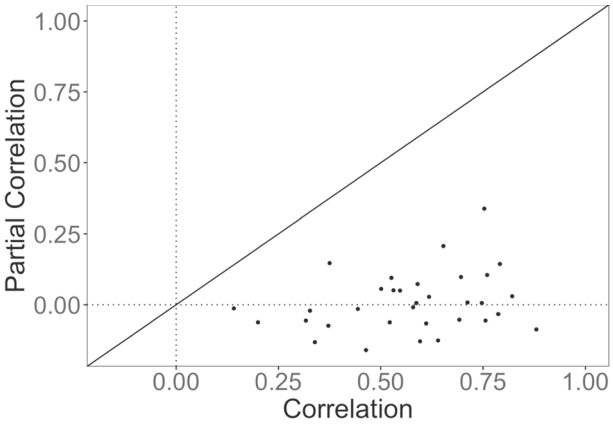


Fig. 9 For $k = 10$, we identify the set of 10 KPIs out of the 300 runs with the highest estimated Spearman's rho with the original pillar scores. We construct a score using the sum of the 10 KPIs and rescale this sum using percentile ranking. Then, we subtract the rescaled constructed score from the original score, labeling this as residual. We select the p-k KPIs that are not part of the best combination and calculate their correlation with the original score (x-axis) and with the residual (y-axis). Each point in the scatter plot represents a KPI, and its coordinates represent the two correlations. The data refers to the Environmental pillar of the Oil sector in 2012

6.4 Empirical analysis: partial correlation

In this section, we show that the set of KPIs identified with the best combination(s) is enough to represent all the information provided by all the KPIs, including those not included. We consider the best combination for any fixed number of KPIs. For each company, we calculate the Environmental score associated with the combination of selected KPIs,[14] and subtract this constructed score from the given Environmental pillar score. In this way, we find what we may call a residual. Then, we calculate Pearson's correlation between the residual and each KPI not included in the best combination, and we label this as a *Partial Correlation* of a KPI with the pillar score given the selected KPIs. Our calculations are formulated in Appendix G.

Figure 9 shows the estimated correlation between the original score and a given KPI on the x-axis and between the residual and the same KPI on the y-axis.[15] This result is based on the constructed scores obtained using the best combination of 10 KPIs. We see in Fig. 9 that most KPIs with a non-zero correlation with the original pillar score have a near-zero correlation with the residual, i.e., a near-zero partial correlation with the original pillar score one the effect of the selected KPIs is accounted. Indeed, only 1 KPI among 32 KPIs still has a correlation significantly different from 0 with the residuals.[16]

More generally, for other values of k and other years and sectors, these results hold with both a reduction in the number of (residual) KPIs correlated with the original score, and a reduction in the absolute correlation values for those (residual) KPIs that still remain significantly correlated.

[14] We do so by applying the percentile ranking to the sum of the scores of the selected variables as given in Eq. (1) and call it *constructed Environmental score*.

[15] Pearson correlation is used for score/residual versus continuous KPIs, whereas polyserial correlation is used for score/residual versus binary KPIs.

[16] For 2 KPIs the correlation with the original score was not significant, and neither was it with the residuals.

6.5 Empirical analysis: selected KPIs for three sectors

In this section, we identify the set of KPIs that appear most commonly in the best combinations, and we discuss their characteristics. To do so, we consider all the runs of our optimization in Eq. (4) for three sectors and 10 years (2012–2021). For each year and sector we identify the best 10 KPIs for the Environmental Pillar. The identification of the list of the best 10 KPIs (for a given year and sector) follows this procedure: for each k between 1 and 15 we select the best combination of k KPIs out of the 300 runs. Then we select the 10 KPIs appearing most frequently in the 15 best combinations.

This procedure is repeated a total of 30 times (10 years and 3 sectors), obtaining 30 lists of 10 KPIs. Then we count how many times a single KPI is included in these lists. This absolute frequency will range between 0 (if the KPI is never in the list of the best 10 KPIs) and 30 (if the KPI is always in the list of the best 10 KPIs). Further, we compute the relative frequency, which is the ratio between the absolute frequency and the maximum potential absolute frequency (which is usually 30, but could be 20 if the KPI is scoring for only two sectors, or 10 if the KPI is scoring for only one sector).

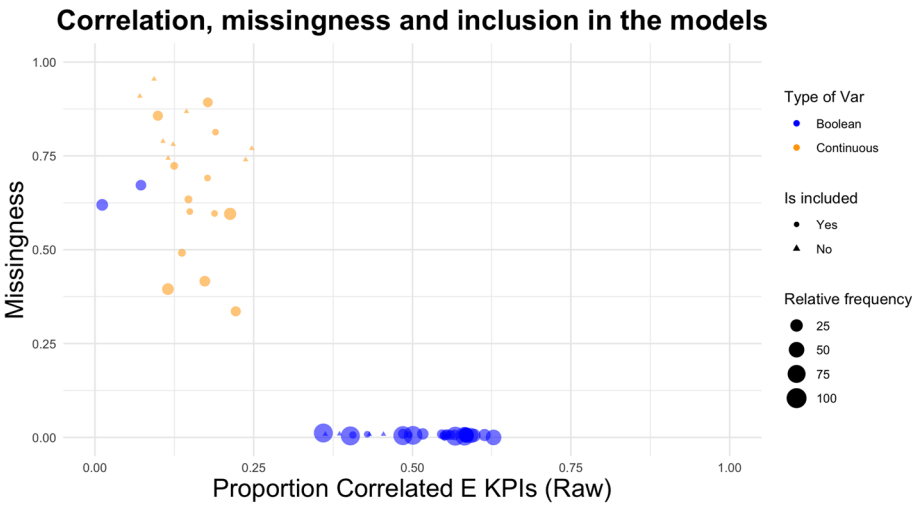
Tables 7 and 8 in Appendix H present the selected KPIs alongside their frequency, type, average correlation, and missingness across the sectors for E and S pillars, respectively.

Considering that the Environmental pillar has a number of KPIs in the 44-48 range, depending on the sector, if we had selected 10 KPIs randomly each of the 30 times, then a single KPI would have had a relative frequency of inclusion around 20-25%. Thus, taking some margin around this range, the KPIs appearing more frequently than 30% of the time in the list of best KPIs will be those that carry more information to replicate the pillar scores, while those appearing less than 15% of the time will be those that carry less information. We identify 15 Environmental KPIs appearing at least 30% of the time in the best combinations, and 14 of these are boolean (93.3%). The baseline frequency of boolean KPIs for the Environmental pillar is 61.8%; a proportion test comparing these two proportions indicates that boolean KPIs are significantly more likely to be in the best combinations ($p < 5\%$). Similarly, we identify 28 KPIs appearing at most 15% of the times, with 12 of these being boolean (42.9%). The equal proportion test against the base rate of 61.8% indicates that boolean KPIs are significantly less likely to be among the least influential KPIs ($p < 5\%$).

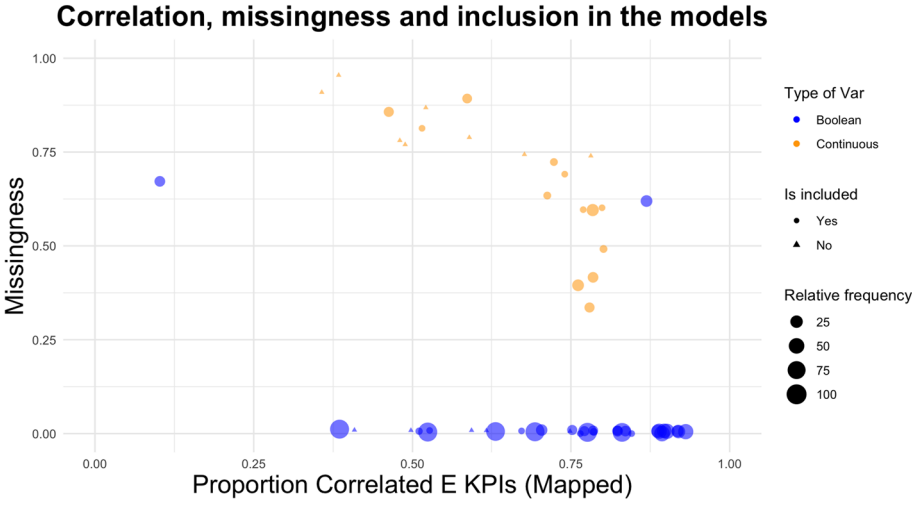
For each of these KPIs, we also compute the average missingness in our datasets and the average proportion of KPIs with which the KPI has a correlation significantly different from 0.[17] In this section, for the sake of simplicity, we will refer to this average proportion of correlated KPIs as correlation. We show the plot of raw correlation against missingness in Fig. 10a, and the plot of mapped correlation against missingness in Fig. 10b. Additional dimensions are shown with color (boolean vs. continuous), shape (included vs. non included in the best solutions) and size of the dots (relative frequency of inclusion).

Figure 10a shows the relationship between raw correlation and missingness. Boolean KPIs (shown in blue) exhibit near-zero missingness and tend to be more likely to be correlated with each other. Continuous KPIs (shown in orange), on the contrary, exhibit higher missingness and lower correlations. The size of the dots represents the relative frequency of inclusion in the list of best KPIs, which is much higher for boolean KPIs. On average, the 15 most influential KPIs are correlated with slightly over 50% of the other Environmental KPIs,

[17] We use this measure instead of simply taking the average correlation for two reasons. First, because some correlations are negative due to opposite polarities, so the average may be misleading. Second, because we would be averaging different correlation measures (Pearson, polyserial and tetrachoric coefficients) depending on the type of variables involved. The process is first executed at a single dataset level. Then we average for all 10 years of data for the same sector, and finally we average across sectors.



(a) Scatter plot produced using correlations in raw data.



(b) Scatter plot produced using correlations in mapped data.

Fig. 10 Scatter plots of the proportion of average correlated Environmental KPIs between pairs of KPIs (x-axis) versus missingness (y-axis). Color indicates variable type, size indicates relative frequency of inclusion among the best combinations of KPIs, and shape indicates whether the KPI was even included in the list of best KPIs

while the 28 least influential are correlated with slightly less than 30% of the other KPIs. This difference of over 20 percentage points is statistically significant ($p < 0.1\%$). However, this gap is mostly due to the variable type: almost all influential KPIs are boolean, which are most likely to be correlated with other KPIs. Once we control for KPI type, the difference in correlations between the most and influential KPIs shrinks to less than 10% for both types of KPIs (the differences are still significant at the 5% level). 13 KPIs are never included in the list of best KPIs and they are shown with a triangular shape. 8 of these KPIs are continuous, and even if their correlation level is in line with the correlation of the other continuous KPIs

($p > 5\%$), they tend to have higher missingness ($p < 0.1\%$). This can be seen in the cluster of small red triangles on the top left. The 5 boolean KPIs that never enter the lists of best KPIs have features comparable to the KPIs that enter, both in terms of missingness ($p > 5\%$) and correlation ($p > 5\%$).

Figure 10b mirrors Fig. 10a, with only one difference: instead of reporting the correlation among raw KPIs on the x axis, it reports the correlation among mapped KPIs. This figure complements the previous one. First, since most dots moved to the right, it confirms our previous finding that the mapping process tends to artificially boost correlations. Second, while the correlation in raw data was not statistically different for the subgroups of entering versus non-entering KPIs, this pattern does not hold for mapped correlations: entering KPI are more likely to be correlated with other KPIs than non-entering KPIs ($p < 1\%$ for continuous KPIs and $p < 5\%$ for boolean KPIs).

The average proportion of correlated KPIs for the most influential KPIs (i.e., those entering with a relative frequency of at least 30%) is 79%, while it is 63% for the least influential KPIs (i.e., those entering with a relative frequency of at most 15%). This difference is statistically significant ($p < 1\%$) and, unlike for raw correlations, it does not change after accounting for variable type.

From the joint analysis of the two figures, we can conclude that the most influential KPIs are boolean KPIs that tend to be more correlated with the other KPIs when using mapped data. The main feature of such KPIs is having near-zero missingness, which reduces the noise of zero imputations. Continuous KPIs tend to be less likely to be included in the best combinations. However, when they are included, they tend to exhibit relatively lower missingness and higher correlation in mapped data. The analysis of the best KPIs for the Social pillar confirms these findings (see Appendix I for more details).

7 Conclusion

In this study, we examined how missing data imputation and KPI aggregation influence Environmental, Social, and Governance (ESG) scores and proposed a parsimonious KPI selection framework with stochastic hill-climbing with random restarts.

Our theoretical and empirical results demonstrate that imputing missing data with zero can introduce artificial correlations when mapped KPIs are considered, which complicates the interpretation of ESG scores. Specifically, rank correlations such as Kendall's tau and Spearman's rho remain consistent between raw and mapped data when there is no missingness. In contrast, the Pearson correlation deviates even without missing data. Therefore, using rank correlations provides more consistent results among ESG data than Pearson's correlation when no missing data is present. However, when missing data is introduced, the dependency structure artificially changes.

We employed stochastic hill-climbing (cardinality-constrained) with random restarts to identify the optimal set of KPIs to construct the actual E- and S-pillar scores as closely as possible, thereby identifying redundancy among KPIs. We show that constructed scores using a relatively small number of KPIs (around 10) can achieve near-perfect association measures with the actual scores. Thus, additional KPIs beyond this point may introduce redundancy without significantly improving performance. In our baseline approach, we use equal weights for KPIs. However, we also consider an alternative weighting scheme based on Refinitiv's category weights. Our results show that including Refinitiv's weights directly in the optimization does not consistently lead to an improvement in performance, especially

for the Social pillar. On the other hand, it always entails a higher computational time. Further, applying Refinitiv's weights ex post to the KPIs sets selected under the Baseline approach yields a performance very close to that of the Weighted approach, with no increase in computational burden. Therefore, our findings imply that the choice of KPIs matters more than the specific weighting scheme in replicating the pillar scores, even though the role of alternative weighting systems remains open to further research, as suggested by Muck and Schmidl (2024).

We assessed robustness through repeated random initializations and out-of-sample evaluation. Appendix E shows that the solution space is multimodal: different KPI subsets can attain very similar objective values while sharing relatively few indicators. This is both a substantive finding and a limitation. The selected subsets should therefore not be interpreted as a universal or definitive minimum set of ESG indicators, but rather as evidence that the scoring system is highly compressible and that several combinations of KPIs contain similar information for reproducing the final pillar scores. Appendix F further shows that KPI subsets selected in one year retain high replication performance in subsequent years. Together, these results support the robustness of our main conclusion to random initialization and temporal variation, while also showing that the relevance of individual KPIs is context-dependent and may vary across sectors. Future research could therefore investigate whether more stable KPI subsets emerge within sectors or alternative ESG frameworks.

The results of our study also highlight the potential for simplification and improvement in ESG scoring processes: fewer KPIs, if selected optimally, can provide similar insights as many. As ESG reporting continues to evolve, our findings emphasize the need for rigorous methodologies that balance data completeness with practical scoring accuracy. This parsimonious replicability has direct implications for ESG-based decision-making, regulatory standardization, and the transparency of ESG reporting frameworks.

From an investor and managerial perspective, KPI redundancy implies that ESG scores may give the impression of being based on a broad and diversified information set, while in practice the resulting rankings can be largely driven by a much smaller number of informational dimensions. This has two consequences. First, users of ESG scores should not interpret a higher number of aggregated KPIs as necessarily implying a more informative or more robust score. Second, portfolio managers, corporate sustainability officers, and other decision-makers may benefit from identifying the subset of KPIs that materially drives score variation, since this can improve monitoring, engagement, benchmarking, and internal reporting. In this context, a promising direction for future research would be to combine statistical redundancy analysis with decision-making criteria, so that selected KPIs are not only able to replicate existing scores, but are also interpretable, relevant for decision-makers, and aligned with substantive ESG outcomes.

For regulators, our results suggest that the standardization of ESG scores should not focus only on increasing the number of disclosed indicators, but also on assessing their marginal informational contribution. Highly redundant KPIs may increase reporting costs and methodological complexity without improving the discriminatory power of ESG scores. A more effective standardization process would therefore require ESG rating providers to disclose not only the set of indicators used, but also the relative contribution of each KPI group to the final score, the treatment of missing values, and the extent to which some indicators may duplicate information already captured by others. In this context, redundancy analysis can support the development of ESG reporting frameworks that are more transparent, interpretable, and auditable. Our findings, for instance, show that the selected KPIs in the E- and S-pillars across various sectors rely heavily on boolean KPIs. While these boolean KPIs offer more straightforward insights, the continuous KPIs, such as greenhouse

gas emission, would provide richer information and allow for more granular evaluation of ESG data-driven performance. Nevertheless, these KPIs exhibit substantially higher rates of missingness across ESG pillars, which makes them noisier measures. This widespread missingness likely contributes to their reduced influence in replicated pillar ratings, despite their relevance. As a result, continuous KPIs play a less prominent role than might be desirable, provided their ability to capture underlying sustainability performance more directly than many binary KPIs. Standardization and focus on few critical quantitative KPIs could therefore improve the interpretability of ESG evaluations.

In addition to the aforementioned search for the optimal subset of KPIs, future research should focus on some methodological aspects. From a data perspective, a promising direction is to study the impact of different missing imputation techniques, as also done by Downing (2025) and Caprioli et al. (2025). In particular, we could study how missing data imputation affects measures of association when dimensionality-reduction methods (e.g., principal component analysis) are applied. This is a valuable extension because PCA-based scores and factor models are increasingly used to summarize high dimensional ESG information.

From a more practical side, comparing ESG scoring methodologies across different providers, including examining variations in KPI aggregation and weighting strategies, could help identify best practices and highlight discrepancies in ESG assessments. In particular, investigating how different ESG score methodologies impact stakeholder decision-making, such as investors' choices, regulatory responses, and consumer perceptions, could provide a clearer understanding of the practical implications and significance of accurate and representative ESG scoring.

Appendix A Example of percentile ranking

Table 3 shows an example of the application of percentile ranking on a continuous KPI. The KPI has negative polarity, so the companies rank higher the lower the value of the KPI. Since missing values are excluded from the computations of the sum of scores for the calculation of the category score, in practice, they receive a score of 0.

Table 3 Example of the application of percentile ranking for a negative polarity continuous KPI

Company	Raw value	Rank	Score (mapped value)
A	9.5	2	62.5
B	NA	Missing	0
C	12.7	3	37.5
D	NA	Missing	0
E	8	1	87.5
F	16.2	4	12.5

Table 4 shows an example of the application of percentile ranking on a binary KPI. The KPI has positive polarity, so it receives a non-zero score if the raw value is True. Missing values receive a score of 0.

Table 4 Example of the application of percentile ranking for a positive polarity binary KPI

Company	Raw value	Rank	Score (mapped value)
A	True	1	75
B	False	4	0
C	NA	Missing	0
D	False	4	0
E	True	1	75
F	True	1	75

Appendix B Association measures

Pearson's correlation coefficient (ρ) between two random variables X and Y is defined as:

$$\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}, \tag{B.1}$$

where $\text{Cov}(X, Y)$ is the covariance of X and Y, and σ_X and σ_Y are the standard deviations of X and Y, respectively.

Kendall's tau measures the rank correlation between two random variables X and Y. Let (X_1, Y_2) and (X_1', Y_2') be independent random pairs of X and Y. Kendall's tau is then defined as:

$$\tau(X, Y) = P[(X_1 - X_1')(Y_2 - Y_2') > 0] - P[(X_1 - X_1')(Y_2 - Y_2') < 0]. \tag{B.2}$$

Spearman's rank correlation measures the rank correlation between two random variables X and Y with marginal distributions F_X and F_Y. It is the same as the Pearson correlation of the two uniform random variables $F_X(X)$ and $F_Y(Y)$.

$$\rho_s(X, Y) = \rho(F_X(X), F_Y(Y)). \tag{B.3}$$

Polyserial correlation measures the association between a continuous variable X and a binary variable Y. Assuming the binary variable Y is derived from an underlying latent continuous variable Y^* with thresholds $\{\theta_k\}$, the polyserial correlation coefficient is defined as:

$$\rho_s(X, Y) = \frac{\text{Cov}(X, Y^*)}{\sigma_X \sigma_{Y^*}}, \tag{B.4}$$

where $\text{Cov}(X, Y^*)$ is the covariance between X and the latent variable Y^*, and σ_X and σ_{Y^*} are their respective standard deviations.

Tetrachoric correlation measures the association between two binary variables, X and Y, assuming they are thresholds of latent continuous variables X^* and Y^*. It is defined implicitly through:

$$\rho_t(X, Y) = \frac{\Phi^{-1}(P_{11}) - \Phi^{-1}(P_{10})\Phi^{-1}(P_{01})}{\sqrt{1 - \rho_t^2}}, \tag{B.5}$$

where P_{ij} are the joint probabilities of the binary outcomes, and Φ^{-1} is the inverse of the cumulative standard normal distribution function. Tables 5 reports a summary of the comparison of the association measures.

Table 5 Comparison of association measures between a score of interest, e.g., pillar scores, and raw and mapped ESG KPIs under scenarios with and without missing data, considering an imputation of zero in case of missingness

Variable types	Measure	No missingness	With missingness
Continuous versus continuous	Pearson	Not same	Not same (even different sign)
Continuous versus continuous	Kendall	Same	Not same (even different sign)
Continuous versus continuous	Spearman	Same	Not same (even different sign)
Continuous versus binary	Polyserial	Not same	Not same (even different sign)
Binary versus binary	Tetrachoric	Same	Not same (even different sign)

Proof (Proposition 4) Assume $n = 20$. Let $\boldsymbol{x} = (1, 0, 1, 1, 1, 1, 0, 0, 0, 0, 1, 1, 0, 0, 0, 1, 1, 0, 1, 0)^\top$ and $\boldsymbol{y} = (-1.02, -0.31, 0.84, -0.30, -0.20, 0.64, 0.82, -1.02, -1.16, 0.24, 1.27, 0.16, 0.79, 0.53, -0.07, 1.89, 0.71, -1.73, 0.45, -0.60)^\top$ be a binary and continuous ESG raw KPI, respectively. After applying Eqs. (2) and (1), we have $g(\boldsymbol{x}) = (75, 0, 75, 75, 75, 75, 0, 0, 0, 0, 75, 75, 0, 0, 0, 75, 75, 0, 75, 0)^\top$ and $f(\boldsymbol{y}) = (15.0, 27.5, 87.5, 32.5, 37.5, 67.5, 82.5, 15.0, 7.5, 52.5, 92.5, 47.5, 77.5, 62.5, 42.5, 97.5, 72.5, 2.5, 57.5, 22.5)^\top$. Then, we have $\hat{\rho}_p(g(\boldsymbol{x}), f(\boldsymbol{y})) = 0.47$ and $\hat{\rho}_p(\boldsymbol{x}, \boldsymbol{y}) = 0.49$. Thus, $\hat{\rho}_p(g(\boldsymbol{x}), f(\boldsymbol{y})) \neq \hat{\rho}_p(\boldsymbol{x}, \boldsymbol{y})$, which proves the statement.

□

Proof (Proposition 5) Assume $n = 20$. Let $\boldsymbol{x} = (0.89, 3.31, 0.14, 1.66, 4.34, 2.12, 0.35, 0.13, 3.37, 1.97, 2.03, 1.64, 0.90, 3.50, 2.68, 1.43, 0.98, 1.25, 0.51, 1.40)^\top$ and $\boldsymbol{y} = (0.68, 2.40, 0.52, 2.13, 3.76, 2.62, 2.89, 2.54, 4.28, 2.61, 3.09, 1.36, 1.79, 3.41, 4.95, 1.30, 2.62, 1.36, 1.70, 1.86)^\top$ be two continuous ESG raw KPIs. After applying Eq. (1), we obtain $f(\boldsymbol{x}) = (22.5, 82.5, 7.5, 57.5, 97.5, 72.5, 12.5, 2.5, 87.5, 62.5, 67.5, 52.5, 27.5, 92.5, 77.5, 47.5, 32.5, 37.5, 17.5, 42.5)^\top$ and $f(\boldsymbol{y}) = (7.5, 47.5, 2.5, 42.5, 87.5, 62.5, 72.5, 52.5, 92.5, 57.5, 77.5, 22.5, 32.5, 82.5, 97.5, 12.5, 67.5, 17.5, 27.5, 37.5)^\top$. Then, it holds that $\hat{\rho}(f(\boldsymbol{x}), f(\boldsymbol{y})) = 0.62$ and $\hat{\rho}(\boldsymbol{x}, \boldsymbol{y}) = 0.66$. Thus, $\hat{\rho}(f(\boldsymbol{x}), f(\boldsymbol{y})) \neq \hat{\rho}(\boldsymbol{x}, \boldsymbol{y})$, proving the statement.

□

Proof (Proposition 6) Assume $n = 20$. Let $\boldsymbol{x} = (0.55, 0, 1.02, 0.03, 0.13, 1.91, 0.17, 0.58, 1.76, 0.25, 0.35, \text{NA}, \text{NA}, \text{NA}, 1.47, \text{NA}, \text{NA}, 0.66, 0.32, 0.96)^\top$ and $\boldsymbol{y} = (\text{NA}, 0.98, 1.69, 0.68, 0.49, -1.48, \text{NA}, 0.90, -1.27, 0.94, 1.16, \text{NA}, -1.57, -1.28, -0.63, 2.09, \text{NA}, \text{NA}, 0.87, 1.32)^\top$ be two continuous ESG raw KPIs.

Let $\boldsymbol{x}^c$ and $\boldsymbol{y}^c$ correspond to the joint complete part of $\boldsymbol{x}$ and $\boldsymbol{y}$ together, respectively. Then we have $\hat{\tau}(\boldsymbol{x}^c, \boldsymbol{y}^c) = -0.18$, $\hat{\rho}(\boldsymbol{x}^c, \boldsymbol{y}^c) = -0.73$, and $\hat{\rho}_s(\boldsymbol{x}^c, \boldsymbol{y}^c) = -0.31$. After applying Eq. (1), we obtain $f(\boldsymbol{x}) = (50.00, 3.33, 76.67, 10.00, 16.67, 96.67, 23.33, 56.67, 90.00, 30.00, 43.33, 0.00, 0.00, 0.00, 83.33, 0.00, 0.00, 63.33, 36.67, 70.00)^\top$ and $f(\boldsymbol{y}) = 0.00, 70.00, 90.00, 43.33, 36.67, 10.00, 0.00, 56.67, 23.33, 63.33, 76.67, 0.00, 3.33, 16.67, 30.00, 96.67, 0.00, 0.00, 50.00, 83.33)^\top$. Then, it holds that $\hat{\tau}(f(\boldsymbol{x}), f(\boldsymbol{y})) = 0.09$, $\hat{\rho}(f(\boldsymbol{x}), f(\boldsymbol{y})) = 0.08$, and $\hat{\rho}_s(f(\boldsymbol{x}), f(\boldsymbol{y})) = 0.13$, proving the statement.

□

Proposition 8 *For any two nonzero, real, and distinct-valued ESG mapped KPIs, $f(\boldsymbol{x})$ and $f(\boldsymbol{y})$, where f is given by Eq. 1, the sample Kendall's tau $\hat{\tau}$ between them $\hat{\tau}(f(\boldsymbol{x}), f(\boldsymbol{y}))$ increases or stays the same with the inclusion of a company, x_{new}, whose KPI is encoded as zero, i.e., $f(x_{new}) = 0$, $f(y_{new}) = 0$.*

Proof Let n be the total number of companies with nonzero, real, and distinct-valued ESG-mapped KPIs, $f(\boldsymbol{x})$ and $f(\boldsymbol{y})$. The Kendall's tau estimate between $f(\boldsymbol{x})$ and $f(\boldsymbol{y})$ is given by

$$\hat{\tau}(f(\boldsymbol{x}), f(\boldsymbol{y})) = \frac{N_c - N_d}{\frac{n\cdot(n-1)}{2}},$$

where $N_c(N_d)$ are the number of concordant (discordant) pairs.

Let $N_c = k$ and $N_d = \frac{n\cdot(n-1)}{2} - k$ be for $k \geq 0$ and $\frac{n\cdot(n-1)}{2} \geq k$. Then

$$\hat{\tau}(f(\boldsymbol{x}), f(\boldsymbol{y})) = \frac{k - \left[\frac{n\cdot(n-1)}{2} - k\right]}{\frac{n\cdot(n-1)}{2}} = \frac{4\cdot k - n\cdot(n-1)}{n\cdot(n-1)}.$$

The inclusion of the $(n+1)^{\text{th}}$ company whose KPI is encoded as zero due to the missingness implies that $f(x_{new}) = 0$, $f(y_{new}) = 0$. Then, the new Kendall's tau estimate between the new KPIs, $f(\boldsymbol{x}*)$ and $f(\boldsymbol{y}*)$, has n more concordant pairs than the pairs of $f(\boldsymbol{x})$ and $f(\boldsymbol{y})$. Hence,

$$\hat{\tau}(f(\boldsymbol{x}*), f(\boldsymbol{y}*)) = \frac{k + n - \left[\frac{n\cdot(n-1)}{2} - k\right]}{\frac{(n+1)\cdot n}{2}} = \frac{4\cdot k + 2n - n\cdot(n-1)}{n\cdot(n+1)}.$$

Since it holds that $\frac{n\cdot(n-1)}{2} \geq k$, we see that $\hat{\tau}(f(\boldsymbol{x}*), f(\boldsymbol{y}*)) \geq \hat{\tau}(f(\boldsymbol{x}), f(\boldsymbol{y}))$, proving the statement.

□

Appendix C Empirical results: missing KPIs

Consider an Environmental KPI M_1 with at least one missing observation (i.e., a company with no reported value for that KPI) for the Oil sector in 2012. The unconditional probability of missingness for KPI M_1, denoted as $P(\text{Missing in } M_1)$, is the proportion of companies with missing observations for that KPI relative to the total number of companies. Now, consider a second Environmental KPI M_2, also with at least one missing observation. The conditional probability of missingness for KPI M_1, given that KPI M_2 is missing, is defined as:

$$P(\text{Missing in } M_1 \mid \text{Missing in } M_2) = \frac{P(\text{Missing in } M_1 \cap \text{Missing in } M_2)}{P(\text{Missing in } M_2)}.$$

This represents the likelihood that a firm has a missing observation for KPI M_1, conditioned on the same firm having a missing observation for KPI M_2.

We calculate the conditional probabilities $P(\text{Missing in } M_1 \mid \text{Missing in } M_2)$ for all pairs of Environmental KPIs where both M_1 and M_2 have at least one missing observation. These conditional probabilities, alongside the unconditional probabilities of missingness $P(\text{Missing in } M_1)$, are plotted in Fig. 11. The fact that all points lie above the $y = x$ line in Fig. 11 demonstrates that if a firm has a missing observation for one KPI, it is more likely also to have a missing observation for another KPI. This indicates a pattern of correlated missingness among the Environmental KPIs. Therefore, there seem to be systematic factors influencing the reporting of these metrics.

Figure 12 is a heatmap reporting the cluster of missing data. Each column represents one KPI with at least one missing observation. Each row represents one company. If a cell is

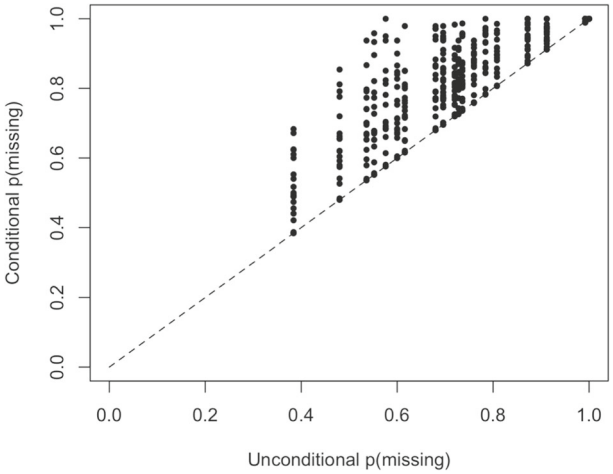


Fig. 11 The probability for a given company to have missing values in a given KPI M_1, conditional on having a missing value in KPI M_2, is higher than the relative frequency of missingness in KPI M_1 in the Oil sector in 2012

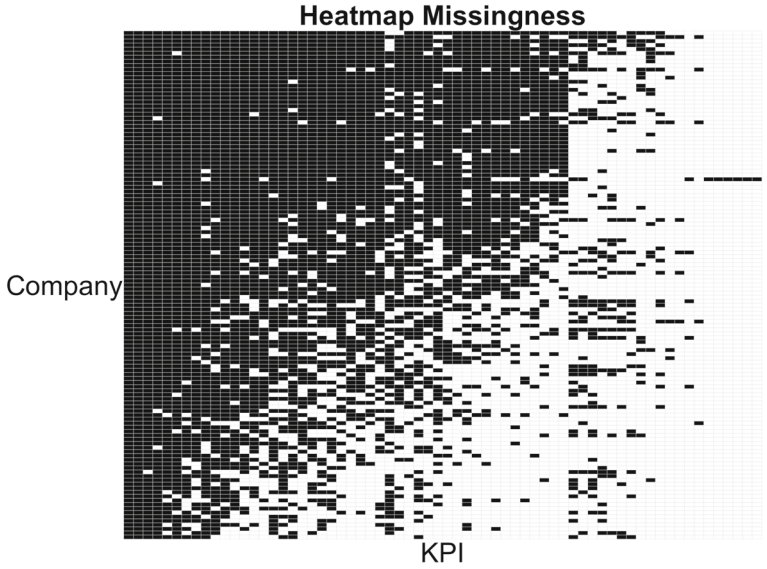


Fig. 12 Heatmap showing the missingness clusters. The dataset is the Oil sector in 2012

black, it means that the company in the row has a missing value for the KPI in the column. Both rows and columns are sorted so that high missingness KPIs are on the left and high missingness companies are at the top. The figure shows that missingness is not a random process. It tends to cluster both at a KPI level and a company level.

Appendix D Empirical results: maximizing correlation

Figure 13 shows the distribution of correlation between the original score and the reproduced scores for all the 300 solutions found in the runs of the algorithm, for multiple values of k (from 2 to 15 at unitary increments, and then from 20 to 40 at 10 KPI increments). The results show that even with a relatively low number of selected KPIs, most solution have a correlation over 90%. Moreover, as k increases the variability in performance of the solutions reduces, suggesting that local optima are found. In the next Section we discuss the level of similarity of these different solutions.

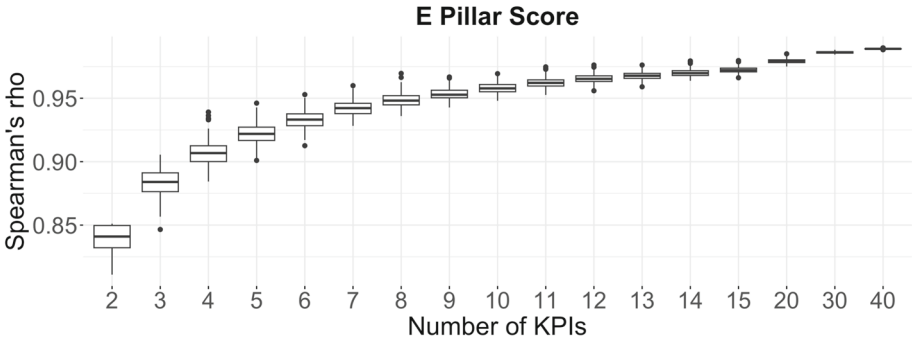


Fig. 13 Distribution of the estimated Spearman's correlation coefficient between the constructed and given scores by Refinitiv for a different number of KPIs selected. Each boxplot includes the results of the 300 runs for each number of KPIs. The dataset is the Oil sector in 2012 for the E pillar

Appendix E Empirical analysis: multimodality

We have already shown in Fig. 7 that for sufficiently high values of k (but still much lower than p), the performance of the 300 runs is similar compared to the random combinations or the combinations aimed at minimizing correlation. However, we have not considered the similarity of the selected KPI combinations obtained across the 300 runs.

In this section, we address this by taking the best combination (for a given k) as a benchmark and determining the distance of the remaining 299 combinations from this benchmark. We calculate two types of distances. First, the distance, based on the similarity of the selected KPIs, is defined as $d_{\text{similarity}} = \frac{k-\#\text{commonKPIs}}{k}$, where #commonKPIs denotes the number of KPIs shared between a combination and the benchmark. Second, the distance, based on the difference in performance, is defined as $d_{\text{performance}} = \left|\frac{\hat{\rho}_s(z_{\text{bench}}, \boldsymbol{y}) - \hat{\rho}_s(z_{\text{comb}}, \boldsymbol{y})}{\hat{\rho}_s(z_{\text{bench}}, \boldsymbol{y})}\right|$, where $\hat{\rho}_s(z_{\text{bench}}, \boldsymbol{y})$ is estimated the Spearman's rho coefficient between the constructed scores using the benchmark combination and the actual scores $\boldsymbol{y}$, and $\hat{\rho}_s(z_{\text{comb}}, \boldsymbol{y})$ is the coefficient for a given combination. These distances allow us to evaluate the KPI combinations' structural similarity and performance consistency relative to the benchmark. Moreover, for a given k, we will denote the average distances of 299 combinations with the benchmark by $\hat{d}_{\text{similarity}}$ and $\hat{d}_{\text{performance}}$.

Table 6 lists the similarity between the optimal KPI combination and the remaining 299 KPI combinations, and it shows that, on average, as k increases, the average distance of the sub-optimal combinations from the best combination increases in absolute terms

Table 6 The table shows for each k (first column) the average distance of the 299 combinations from the best combination calculated as $k - \widehat{\#\text{common}}\text{KPIs}$ (second column), the average proportion of KPIs different from the best combinations given by $\hat{d}_{\text{similarity}}$ (third column), and the average relative distance in terms of estimated Spearman's rho given by $\hat{d}_{\text{performance}}$ (fourth column). The data refers to the Environmental pillar of the Oil sector in 2012

# KPIs	Avg dist KPI	Avg % different KPI	Avg dist Corr
2	1.65	82.44%	1.31%
3	2.31	77.03%	2.41%
4	2.94	73.58%	3.42%
5	3.77	75.32%	2.56%
6	3.99	66.50%	2.07%
7	4.87	69.52%	1.85%
8	5.13	64.17%	2.18%
9	6.23	69.23%	1.38%
10	6.58	65.79%	1.75%
11	6.31	57.40%	1.30%
12	7.41	61.76%	1.11%
13	7.19	55.34%	0.87%
14	7.86	56.16%	0.97%
15	8.38	55.90%	0.76%
20	8.98	44.90%	0.60%
30	6.85	22.84%	0.21%
40	1.89	4.72%	0.08%

($k - \widehat{\#\text{common}}\text{KPIs}$) since there are more KPIs selected. However, it decreases in relative terms ($\hat{d}_{\text{similarity}}$). The sub-optimal combination, on average, has only 25% of the KPIs in common ($\hat{d}_{\text{similarity}} = 25\%$) with the best combination when only 5 KPIs are selected, 35% when 10 KPIs are selected, and 45% when 15 KPIs are selected. When k gets higher, the distance tends to reduce in relative and absolute terms. However, despite the relatively high distance in terms of similarity ($\hat{d}_{\text{similarity}}$), performance is quite close, with average performance distance ($\hat{d}_{\text{performance}}$) as small as 1%-2%. This confirms once again multimodality: different combinations produce similarly high performance.

Appendix F Empirical analysis: out of sample performance

In this section, we analyze the performance of the best KPI set identified by our optimization in a given year for the other years, i.e., we perform an out-of-sample analysis. Figure 14 shows the in-sample and out-of-sample performance of the best combination identified in 2012, 2017, and 2021 for 2012–2021 for the Environmental pillar score in the Oil sector. Therefore, we have three different combinations of k selected KPIs. The red line represents the estimated Spearman's rho coefficient between the Environmental pillar score (for the year reported on the x-axis) and the constructed score using the ten KPIs in the optimal combination for 2012. The red dot represents the in-sample performance, whereas the line represents the out-of-sample performance in years other than 2012. The same logic applies to

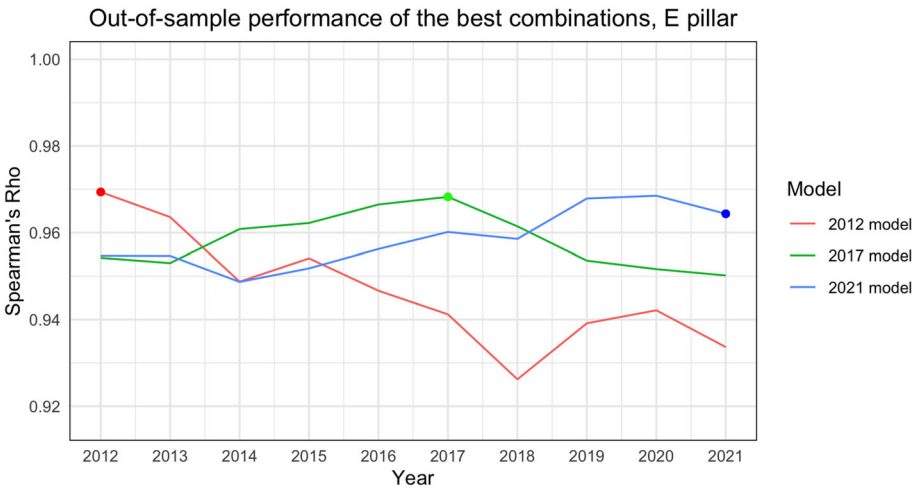


Fig. 14 In-sample (dots) and out-of-sample (lines) performance of the best combination (of 10 KPIs) to reproduce the Environmental pillar score for the Oil sector regarding the estimated Spearman's rho between the constructed scores with the best KPI combination and original score at the given year

green and blue dots and curves for 2017 and 2021 when taken as the in-sample. Even though the estimated correlation is lower in the out-of-sample, as seen in red, green, and blue curves in Fig. 14, it is still very high: the optimal KPI combination of 2012 can replicate the pillar score a decade later with a loss of 4% in the estimated correlation, and a correlation still well above 90%.

Unsurprisingly, the out-of-sample performance gets worse almost monotonically since a variable that works well in replicating the score in year t still works well in year $t + 1$ (or would have worked well in year $t - 1$), but may not work as well as we move away from year t. This is likely since many of the KPIs, especially the boolean ones, stay the same from one year to the next, but are more likely to change over longer time frames.

We confirm that not only does the best KPI combination have stable performance regarding the estimated Spearman's rho over time, but also the other combinations. We perform an out-of-sample analysis by taking all the 300 combinations resulting from the stochastic hill-climbing on the 2012 dataset (for the Oil sector), and show the boxplots of the estimated Spearman's rho in out-of-sample in Fig. 15. The black triangles show the out-of-sample performance of the best KPI combination identified in 2012, while the green triangles show the in-sample performance of the best combination identified for that year by our optimization in Eq. (4). We can infer two relevant points from Fig. 15. First, most of the combinations of KPIs identified based on the 2012 data exhibit a correlation above 0.92 in the subsequent years, suggesting that combinations that worked well to replicate the score in 2012 can also do so in later years. Second, the combination of KPIs that had the highest Spearman's correlation with the original score of 2012 (i.e., the best combination) has an out-of-sample performance in the subsequent years, which is in line with the other suboptimal combinations identified in 2012. This reinforces the message that there is nothing special about the optimal combination of a given year: since there is multimodality, we can reconstruct pillar scores using a random set of 10 KPIs from a subset of "good KPIs" that appear often in the optimal combinations, which is much smaller than the set of the original KPIs. Based on our out-of-sample analysis, this replicated score will most likely have an estimated Spearman's rho with the original scores above 0.90.

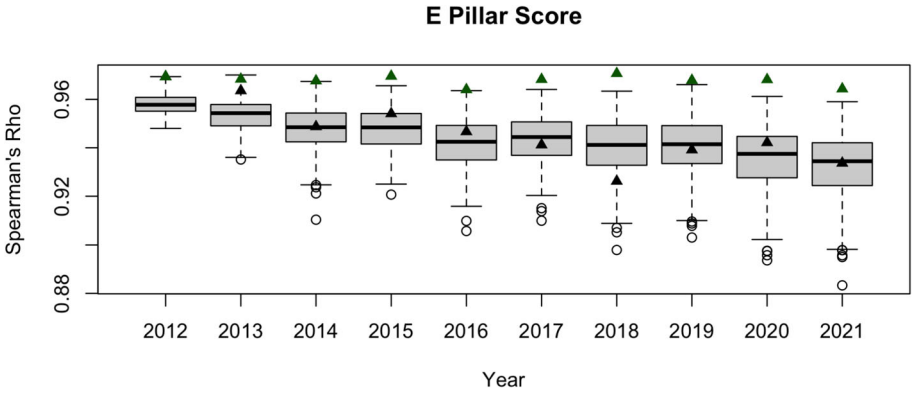


Fig. 15 Distribution of the out-of-sample performance of the 300 combinations identified with the 2012 stochastic hill-climbing model in the years indicated on the x-axis to reproduce the Environmental pillar score for the Oil sector. The black triangles show the out-of-sample performance of the best combination of 2012. The green triangles show the in-sample performance for that year's best combination of the stochastic hill-climbing. The in-sample performance for 2012 is shown as a benchmark

Appendix G Empirical results: partial correlation

Let k denote the number of KPIs in the best combination. For each firm, we compute the *constructed Environmental score* based on the selected KPIs. This score is obtained by applying the percentile ranking to the sum of the scores of the selected variables, as given in Eq. (1).

Denote by y_i the actual Environmental pillar score for firm i, and by $f(z_i)$ the constructed score derived from the best combination of KPIs. The *residual* for firm i is then defined as:

$$r_i = y_i - f(z_i).$$

For each KPI that is not included in the optimal combination, we compute Pearson's correlation coefficient defined in Appendix B between the residual vector $\mathbf{r} = (r_1, r_2, \ldots, r_n)^\top$ and the vector of scores for that KPI across all firms. We denote this correlation as the *Partial Correlation* of the KPI with the Environmental pillar score, conditional on the selected KPIs.

A high partial correlation suggests that the excluded KPI contains additional, non-redundant information relative to the selected KPIs. Conversely, a low partial correlation indicates that the optimal combination already captures the information from the excluded KPI.

Appendix H Selected KPIs for three sectors for E Pillar

In Table 7, we report for each KPI, its type (B for Boolean, C for Continuous), the average missingness in our 30 datasets (3 sectors, 10 years), the correlations with the Environmental Score for the raw and mapped values of the KPI, and the absolute and relative frequency of inclusion in the best KPI list. Absolute frequency indicates how many times the KPI is in the list of the best 10 KPIs. That column must sum up to 300 since the number of KPIs the algorithm selects is $3 \cdot 10 \cdot 10$ corresponding to three sectors, 10 years, and 10 KPIs. Relative frequency divides the relative frequency by 30 (or 20 if the KPI is a scoring KPI in only two

sectors, or 10 if the KPI is a scoring KPI in only one sector). The table is arranged by relative frequency in descending order.

Appendix I Selected KPIs for three sectors for S Pillar

We replicate the analysis of Sect. 6.5 for the Social Pillar. Table 8 replicates Table 7, but for the Social Pillar.

We identify 12 Social KPIs appearing at least 30% of the time in the best combinations, all of which are boolean (100%). The baseline frequency of boolean KPIs for the Social pillar is 61.5%; a proportion test comparing these two proportions indicates that boolean KPIs are significantly more likely to be in the best combinations ($p < 1\%$). Similarly, we identify 30 KPIs appearing at most 15% of the times, with 13 of these being boolean (43.3%). The equal proportion test against the base rate of 61.5% indicates that boolean KPIs are significantly less likely to be among the least influential KPIs ($p < 5\%$).

Figure 16a shows the relationship between raw correlation and missingness. Boolean KPIs (shown in blue) exhibit near-zero missingness and tend to be more likely to be correlated with each other. Continuous KPIs, on the contrary, exhibit higher missingness and lower correlations. The size of the dots represents the relative frequency of inclusion in the list of best KPIs, which is much higher for boolean KPIs. On average, the 12 most influential KPIs are correlated with slightly over 50% of the other Social KPIs, while the 30 least influential are correlated with slightly less than 25% of the other KPIs. This difference of almost 30 percentage points is statistically significant ($p < 0.1\%$). It partially reduces once we control for KPI type, but it is still larger than 15 percentage points for boolean KPIs and significant at a 1% level. 19 KPIs are never included in the list of best KPIs, and they are shown in the Figure with a triangular shape. 11 of these KPIs are continuous and 8 are boolean. For both types of KPIs, entering and non-entering KPIs have the same missingness levels ($p > 5\%$), while their correlation is lower ($p < 5\%$). The non-entering versus entering correlation gap is about 5.5% for continuous KPIs and 13% for boolean KPIs.

Figure 16b mirrors Fig. 16a, with only one difference: instead of reporting the correlation among raw KPIs on the x axis, it reports the correlation among mapped KPIs. This Figure complements the previous one. First, since most dots moved to the right, it confirms our previous finding that the mapping process tends to artificially boost correlations. Second, it shows that even if correlation increased, the continuous KPIs that never enter the lists of best KPIs tend to have lower correlation in mapped data than the other continuous KPIs that enter ($p < 0.1\%$). The same is true for boolean KPIs ($p < 5\%$), with entering KPIs being more likely to be correlated with other KPIs than non-entering KPIs.

The average proportion of correlated KPIs for the most influential KPIs (i.e., those entering with a relative frequency of at least 30%) is 65%, while it is 41% for the least influential KPIs (i.e., those entering with a relative frequency of at most 15%). This difference is statistically significant ($p < 0.1\%$).

This analysis confirms that the most influential KPIs are boolean KPIs that tend to be more correlated with the other KPIs when using mapped data and have low levels of missing data. Continuous KPIs tend to be less likely to be included in the best combinations. However, when they are included, they tend to exhibit higher correlation in mapped data.

Table 7 The selected KPIs for the Environmental pillar score across three sectors and 10 years

KPI	Type	Missingness	Corr raw	Corr mapped	AbsFreq	RelFreq
Hybrid Vehicles	B	0.48	48.48	69.35	10	100.00
Noise Reduction	B	0.33	56.74	77.61	10	100.00
Sustainable Building Products	B	1.18	35.96	38.51	10	100.00
Water Technologies	B	0.41	40.22	52.39	10	100.00
Environmental Products	B	0.32	58.24	83.06	29	96.67
Renewable/Clean Energy Products	B	0.55	50.07	63.14	29	96.67
Product Impact Minimization	B	0.00	62.83	89.35	6	60.00
Environmental Supply Chain Management	B	0.63	58.41	88.90	17	56.67
Policy Emissions	B	0.51	59.22	93.08	17	56.67
Policy Environmental Supply Chain	B	0.60	58.53	90.12	17	56.67
Targets Emissions	B	0.72	58.40	90.12	15	50.00
Policy Energy Efficiency	B	0.53	59.73	91.94	12	40.00
Policy Water Efficiency	B	0.64	58.69	88.68	11	36.67
Environmental Partnerships	B	0.64	61.42	91.80	9	30.00
NOx Emissions To Revenues	C	59.55	21.28	78.43	9	30.00
Environmental Supply Chain Monitoring	B	61.95	1.13	86.91	8	26.67
Water Use To Revenues	C	39.51	11.49	76.13	8	26.67

Table 7 continued

KPI	Type	Missingness	Corr raw	Corr mapped	AbsFreq	RelFreq
Emissions Trading	B	0.94	51.61	70.40	5	25.00
Climate Change Commercial Risks Opportunities	B	0.63	55.18	82.35	6	20.00
Environmental Materials Sourcing	B	0.66	56.68	83.60	6	20.00
Total Energy Use To Revenues	C	41.62	17.30	78.49	6	20.00
Agrochemical Products	B	67.20	7.23	10.21	2	20.00
Biodiversity Impact Reduction	B	0.65	55.98	77.12	5	16.67
Targets Energy Efficiency	B	0.82	55.44	82.34	5	16.67
Targets Water Efficiency	B	0.96	48.50	75.21	5	16.67
Total CO2 Equivalent Emissions To Revenues	C	33.59	22.17	77.93	5	16.67
Total Renewable Energy To Energy Use	C	85.71	9.90	46.26	5	16.67
Env R&D Expenditures To Revenues	C	89.25	17.77	58.67	3	15.00
Toxic Chemicals Reduction	B	0.85	54.62	78.51	3	15.00
CO2 Equivalent Emissions Indirect, Scope 3 To Rev.	C	72.34	12.47	72.33	2	6.67
Environment Management Team	B	0.44	54.84	78.62	2	6.67
Total Hazardous Waste To Revenues	C	63.42	14.72	71.28	2	6.67
Total Waste To Revenues	C	49.18	13.69	80.14	2	6.67
e-Waste Reduction	B	0.64	40.61	51.03	1	5.00
Environmental Expenditures Investments	B	0.00	58.04	84.57	1	3.33

Table 7 continued

KPI	Type	Missingness	Corr raw	Corr mapped	AbsFreq	RelFreq
Green Buildings	B	0.83	42.87	52.69	1	3.33
NOx and SOx Emissions Reduction	B	0.70	49.45	67.24	1	3.33
SOx Emissions To Revenues	C	59.66	18.83	76.94	1	3.33
VOC Emissions To Revenues	C	69.11	17.72	74.03	1	3.33
VOC or Particulate Matter Emissions Reduction	B	0.00	55.03	76.49	1	3.33
Waste Recycled To Total Waste	C	60.16	14.93	79.90	1	3.33
Water Recycled	C	81.33	18.98	51.49	1	3.33
Env Supply Chain Partnership Termination	B	0.80	43.21	61.70	0	0.00
Land Environmental Impact Reduction	B	0.93	38.51	40.85	0	0.00
Flaring Gases To Revenues	C	73.98	23.72	78.14	0	0.00
Ozone-Depleting Substances To Revenues	C	90.92	7.06	35.71	0	0.00
Water Pollutant Emissions To Revenues	C	78.90	10.70	59.01	0	0.00
EMS Certified Percent	C	74.36	11.52	67.69	0	0.00
Environmental Restoration Initiatives	B	0.64	55.66	74.87	0	0.00
Accidental Spills To Revenues	C	77.00	24.69	48.84	0	0.00
Self-Reported Environmental Fines To Revenues	C	78.08	12.31	48.00	0	0.00
Internal Carbon Price per Tonne	C	95.48	9.30	38.37	0	0.00
Policy Sustainable Packaging	B	0.86	45.43	59.36	0	0.00
Renewable Energy Use Ratio	C	86.83	14.41	52.09	0	0.00
Staff Transportation Impact Reduction	B	0.86	36.27	49.75	0	0.00

The table details their Type (B for Boolean, C for Continuous), the extent of missing data (Missingness), the average proportion of other Environmental KPIs that are significantly correlated with the KPI indicated in the first column in raw data (Corr raw) and in mapped data (Corr mapped), the absolute frequency of inclusion in the best KPIs list (AbsFreq) and the relative frequency of inclusion (RelFreq)

Table 8 The selected KPIs for the Social pillar score across three sectors and 10 years

KPI	Type	Missingness	Corr raw	Corr mapped	AbsFreq	RelFreq
Fundamental Human Rights ILO UN	B	0.70	58.57	74.971	29	96.67
Human Rights Contractor	B	0.71	60.68	74.69	28	93.33
Human Rights Breaches Contractor	B	0.80	49.52	58.37	23	76.67
Policy Human Rights	B	0.61	58.57	75.85	23	76.67
Policy Freedom of Association	B	0.78	51.16	63.06	22	73.33
Policy Customer Health & Safety	B	0.62	53.40	65.24	21	70.00
Policy Child Labor	B	0.76	52.18	65.65	20	66.67
Corporate Responsibility Awards	B	0.68	46.94	61.97	13	43.33
Quality Mgt Systems	B	0.00	48.91	57.89	11	36.67
Training and Development Policy	B	0.00	52.59	71.22	11	36.67
Employees Health & Safety OHSAS 18001	B	0.60	53.06	68.30	9	30.00
Policy Fair Competition	B	0.40	40.61	44.22	9	30.00
Improvement Tools Business Ethics	B	0.38	39.46	41.02	8	26.67
Policy Business Ethics	B	0.35	42.86	45.65	8	26.67

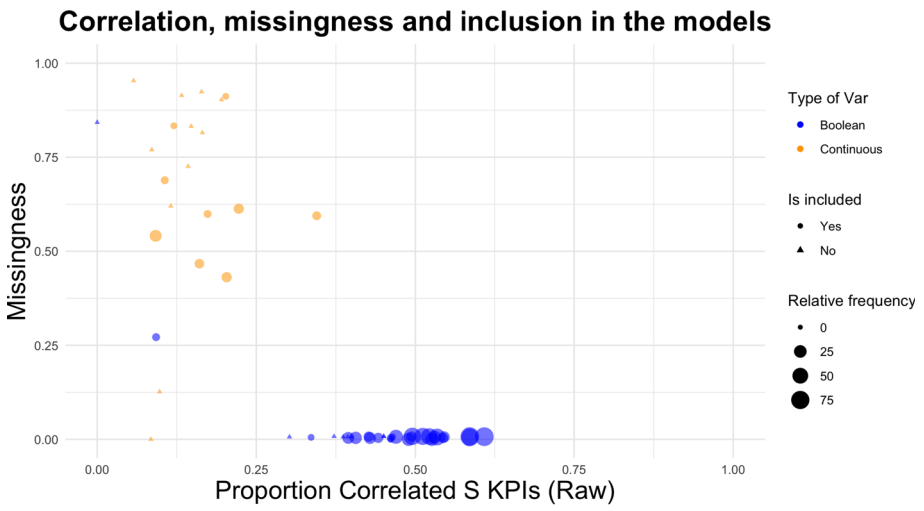
Table 8 continued

KPI	Type	Missingness	Corr raw	Corr mapped	AbsFreq	RelFreq
Total Donations To Revenues	C	54.11	9.18	52.86	8	26.67
Employees Health & Safety Team	B	0.62	54.49	72.18	7	23.33
Policy Bribery and Corruption	B	0.39	44.15	46.80	5	16.67
Policy Diversity and Opportunity	B	0.37	54.35	64.63	5	16.67
Policy Forced Labor	B	0.75	49.39	61.22	5	16.67
Product Responsibility Monitoring	B	0.72	42.65	47.28	5	16.67
Women Employees	C	43.11	20.34	62.93	5	16.67
Women Managers	C	61.32	22.24	66.73	5	16.67
Injuries To Million Hours	C	46.70	16.05	56.46	4	13.33
Trade Union Representation	C	59.44	34.49	62.86	3	10.00
Average Training Hours	C	68.88	10.61	58.44	2	6.67
Policy Community Involvement	B	0.34	46.12	61.02	2	6.67
Turnover of Employees	C	59.92	17.35	62.18	2	6.67
Whistleblower Protection	B	27.17	9.25	37.14	2	6.67
Flexible Working Hours	B	0.66	46.39	55.17	1	3.33
Health & Safety Policy	B	0.00	46.12	56.94	1	3.33
HRC Corporate Equality Index	C	91.19	20.20	38.71	1	3.33
Policy Data Privacy	B	0.53	33.61	38.71	1	3.33
Training Costs Per Employee	C	83.36	12.04	47.14	1	3.33

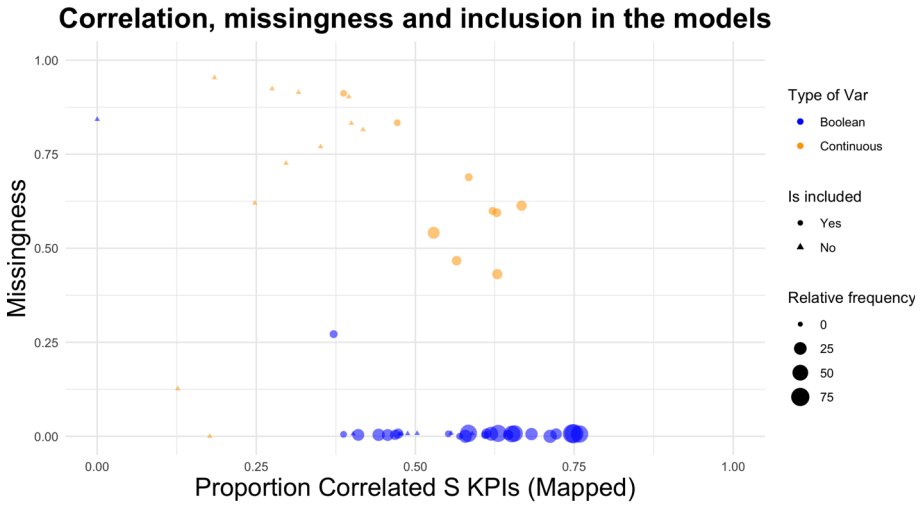
Table 8 continued

KPI	Type	Missingness	Corr raw	Corr mapped	AbsFreq	RelFreq
Targets Diversity and Opportunity	B	0.71	39.93	48.78	0	0.00
Employee Satisfaction	C	91.44	13.27	31.63	0	0.00
Salary Gap	C	62.03	11.56	24.76	0	0.00
Net Employment Creation	C	12.66	9.80	12.65	0	0.00
Announced Layoffs To Total Employees	C	0.00	8.44	17.69	0	0.00
Gender Pay Gap Percentage	C	95.36	5.71	18.44	0	0.00
Day Care Services	B	0.66	38.64	47.82	0	0.00
Employees With Disabilities	C	83.22	14.76	39.93	0	0.00
Employee Health & Safety Training Hours	C	90.29	19.52	39.52	0	0.00
Occupational Diseases	C	81.54	16.53	41.77	0	0.00
Lost Days To Total Days	C	72.57	14.29	29.66	0	0.00
HIV-AIDS Program	B	0.65	39.39	47.55	0	0.00
Internal Promotion	B	0.78	37.21	50.27	0	0.00
Supplier ESG training	B	0.74	45.03	58.98	0	0.00
Extractive Industries Transparency Initiative	B	0.65	30.20	40.20	0	0.00
Critical Country 1	B	84.26	0.00	0.00	0	0.00
Customer Satisfaction	C	92.39	16.39	27.48	0	0.00
OECD Guidelines for Multinational Enterprises	B	0.81	45.00	55.61	0	0.00
QMS Certified Percent	C	76.99	8.57	35.10	0	0.00

The table details their Type (B for Boolean, C for Continuous), the extent of missing data (Missingness), the average proportion of other Social KPIs that are significantly correlated with the KPI indicated in the first column in raw data (Corr raw) and in mapped data (Corr mapped), the absolute frequency of inclusion in the best KPIs list (AbsFreq) and the relative frequency of inclusion (RelFreq)



(a) Scatter plot produced using correlations in raw data.



(b) Scatter plot produced using correlations in mapped data.

Fig. 16 Scatter plots of the proportion of average correlated KPIs between pairs of Social KPIs (x-axis) versus missingness (y-axis). Color indicates variable type, size indicates relative frequency of inclusion among the best combinations of KPIs, and shape indicates whether the KPI was even included in the list of best KPIs

Acknowledgements Sandra Paterlini would like to thank the EU COST Action CA21163 (HiTEc) "Text, functional and other high-dimensional data in econometrics: New models, methods, applications". We thank Sebastian Utz for his insightful comments and positive feedback, which improved this manuscript. We are also grateful to two anonymous referees for commenting and leading to a considerably improved version of the paper.

Author contributions Matteo Benuzzi: Conceptualization, Methodology, Software, Data Curation, Visualization, Validation, Writing—Original Draft, Writing—Review and Editing. Ozge Sahin: Conceptualization, Methodology, Formal analysis, Software, Validation, Writing—Original Draft, Writing—Review and Edit-

ing. Sandra Paterlini: Conceptualization, Methodology, Formal analysis, Resources, Writing—Review and Editing.

Funding Open access funding provided by Università degli Studi di Trento within the CRUI-CARE Agreement.

Declarations

Competing interests The authors have no competing interests to declare that are relevant to the content of this article.

Declaration of generative AI and AI-assisted technologies in the writing process During the preparation of this work, the authors used ChatGPT in order to improve the readability and language of the manuscript. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

References

Benuzzi, M., Bax, K., Paterlini, S., & Taufer, E. (2025). Chasing ESG performance: How methodologies shape outcomes. *International Review of Financial Analysis, 104*, Article 104239.

Berg, F., Koelbel, J. F., & Rigobon, R. (2022). Aggregate confusion: The divergence of ESG ratings. *Review of Finance, 26*, 1315–1344. https://doi.org/10.1093/rof/rfac014

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*, 1426–1445. https://doi.org/10.1002/csr.2112

Billio, M., Fitzpatrick, A. C., Latino, C., & Pelizzon, L. (2024). Unpacking the ESG ratings: Does one size fit all? Working Paper 415. SAFE. https://ssrn.com/abstract=4742445, https://doi.org/10.2139/ssrn.4742445. Available at SSRN.

Caprioli, S., Foschi, J., Crupi, R., & Sabatino, A. (2025). Denoising ESG: Uncertainty-aware scoring through probabilistic imputation of missing data. *Quantitative Finance,1–10,*. https://doi.org/10.1080/14697688.2025.2542507

Castellan, N. J., Jr. (1966). On the estimation of the tetrachoric correlation coefficient. *Psychometrika, 31*, 67–73.

Dobrick, J., Klein, C., & Zwergel, B. (2023). Size bias in refinitiv ESG data. *Finance Research Letters, 55*, Article 104014. https://doi.org/10.1016/j.frl.2023.104014

Downing, N. J. (2025). Missing value imputation in environmental, social, and governance data: an impact on emissions scores. *Finance Research Letters, 85*, Article 107818. https://doi.org/10.1016/j.frl.2025.107818

Drempetic, S., Klein, C., & Zwergel, B. (2020). The influence of firm size on the ESG score: Corporate sustainability ratings under review. *Journal of Business Ethics, 167*, 333–360. https://doi.org/10.1007/s10551-019-04174-9

Joe, H. (2014). *Dependence modeling with copulas*. New York: Chapman and Hall/CRC.

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika, 30*, 81–93.

Malecki, C. (2023). The EU proposal for a regulation on the transparency and integrity of ESG rating activities on 13 June 2023: The missing piece of sustainable finance regulation. *Law and Financial Markets Review, 17*(2), 156–170.

Muck, M., & Schmidl, T. (2024). Comparing ESG score weighting approaches and stock performance differentiation. *Finance Research Letters, 67*, Article 105924. https://doi.org/10.1016/j.frl.2024.105924

Olsson, U., Drasgow, F., & Dorans, N. J. (1982). The polyserial correlation coefficient. *Psychometrika, 47*, 337–347.

Sahin, Ö., Bax, K., Czado, C., & Paterlini, S. (2022). Environmental, social, governance scores and the missing pillar - why does missing information matter? *Corporate Social Responsibility and Environmental Management, 29*, 1782–1798. https://doi.org/10.1002/csr.2326

Sahin, Ö., Bax, K., Paterlini, S., & Czado, C. (2023). The pitfalls of (non-definitive) environmental, social, and governance scoring methodology. *Global Finance Journal, 56*, Article 100780. https://doi.org/10.1016/j.gfj.2022.100780

Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology, 100*, 441–471.