# Four Things No One Will Tell You About ESG Data

## Citation

## Published Version

## Permanent link

## Terms of Use

# Share Your Story

# Four Things No One Will Tell You About ESG Data

Sakis Kotsantonis and George Serafeim*

## Abstract

As the ESG finance field and the use of ESG data in investment decision-making continue to grow, we seek to shed light on several important aspects of ESG measurement and data. This article is intended to provide a useful guide for the rapidly rising number of people entering the field. We focus on the following:

- The sheer variety, and inconsistency, of the data and measures, and of how companies report them. Listing more than 20 different ways companies report their employee health and safety data, the authors show how such inconsistencies lead to significantly different results when looking at the same group of companies.
- 'Benchmarking,' or how data providers define companies' peer groups, can be crucial in determining the performance ranking of a company. The lack of transparency among data providers about peer group components and observed ranges for ESG metrics creates market-wide inconsistencies and undermines their reliability.
- The differences in the imputation methods used by ESG researchers and analysts to deal with vast 'data gaps' that span ranges of companies and time periods for different ESG metrics can cause large 'disagreements' among the providers, with different gap-filling approaches leading to big discrepancies.
- The disagreements among ESG data providers are not only large, but actually increase with the quantity of publicly available information. Citing a recent study showing that companies that provide more ESG disclosure tend to have more variation in their ESG ratings, the authors interpret this finding as clear evidence of the need for 'a clearer understanding of what different ESG metrics might tell us and how they might best be institutionalized for assessing corporate performance.'

What can be done to address these problems with ESG data? Companies should 'take control of the ESG data narrative' by proactively shaping disclosure instead of being overwhelmed by survey requests. To that end, companies should 'customize' their metrics to some extent, while at the same time seeking to self-regulate by reaching agreement with industry peers on a 'reasonable baseline' of standardized ESG metrics designed to achieve comparability. Investors are urged to push for more meaningful ESG disclosure by narrowing the demand for ESG data into somewhat more standardized, but still manageable metrics. Stock exchanges should consider issuing—and perhaps even mandating—guidelines for ESG disclosures designed in collaboration with companies, investors, and regulators. And data providers should come to agreement on best practices and become as transparent as possible about their methodologies and the reliability of their data.

---

* Sakis Kotsantonis is the managing partner of KKS Advisors, an advisory services firm helping clients to find innovative solutions that enable the creation of more sustainable business models and communities. George Serafeim is the Charles M. Williams Professor of Business Administration at Harvard Business School.

## Introduction

The demand for information that captures how companies use different forms of capital—natural, social, and intellectual, as well as financial—to provide their products and services, and how their activities affect society through positive and negative externalities, has led to the creation of environmental, social, and governance (ESG) metrics and related corporate reporting efforts. These efforts have resulted in a proliferation of ESG reports, associated ESG data and ratings, and organizations trying to develop a more rigorous and systematic reporting of ESG information.

As the ESG finance field and the number of people using ESG data in investment decisions continue to grow, we feel it is important to shed light on and express our concerns about several important aspects of ESG measurement and data. This is not meant to be a criticism of any efforts in the field to measure, analyze, and communicate ESG activities and outcomes. It is intended rather to provide a useful guide for the rapidly rising number of people entering the field.

## Back to Basics

The primary goal of ESG metrics is to capture as accurately as possible a firm's performance on a given ESG issue. Only when this goal is achieved will investors be able to use the data to hold companies accountable for their ESG performance as part of their engagement efforts, or to integrate the data into their business analysis and valuation tools. From a corporate perspective, only then will companies be able to tell when their efforts are effective in producing the intended outcomes, and how to systematically integrate these efforts into their operating processes, corporate strategy, and executive compensation plans. And the same is true for customers when using the data to guide their purchasing decisions, for employees when choosing where to work, for regulators when monitoring companies and creating incentives and sanctions, and for NGOs when designing their efforts to drive social progress.

The question then becomes whether ESG data accurately capture a firm's performance. We are sceptical that this is the case, and we are continually amazed that we find signals and meaningful relationships with economic outcomes given the poor quality of the data. At the same time, the success of ever more researchers in detecting such relationships continues to reinforce our belief in the underlying phenomenon we have been studying—namely, the effectiveness of management teams that improve their performance on *material* ESG issues in increasing the competitiveness, financial performance, and value of their companies. We believe that this is the new reality—and that if we had better data, we would find even stronger relationships.

## Our Working Model

Throughout this article, we use simple figures that show the distribution of the performance of a group of companies with regard to a given ESG metric as a normal distribution, even though the actual distributions are far from normal. Our aim in so doing is

to demonstrate some of the biggest challenges with the current ESG data. The same logic can be applied to any ESG metric, or aggregate ESG scores.

So, let's begin by assuming that the performance of a group of companies on an ESG metric looks like the normal distribution shown in Figure 1. A few companies will perform at the top, a similar number will perform at the bottom, and many of the companies will be clustered around a mean value of performance. In that case, it is straightforward to understand how well any company is performing given its relative position on the distribution.
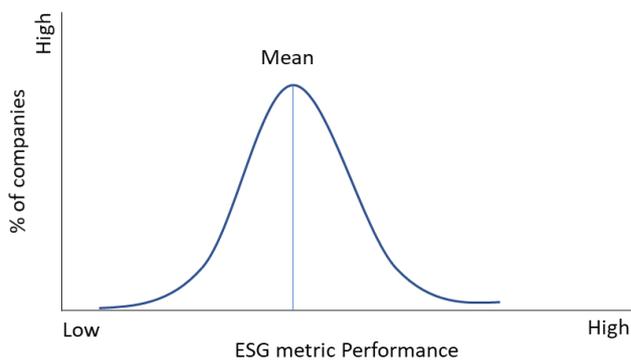


*Figure 1. Example of a normal distribution describing observed (real) data of performance on an ESG metric from a list of companies.*

Keeping this "normal" distribution in mind, we will focus on four limitations of the data that might not be obvious to consumers of it and, importantly, the steps that we can take to overcome those limitations.

1. Data inconsistency is worse than you think it is.
2. Distortions are introduced by the mystic art of "benchmarking" (or defining your peer group matters).
3. ESG data imputation can be a problem (or not all models are created equal).
4. ESG data providers disagree a lot (and even more, surprisingly, when there is publicly available information).

## 1. Data inconsistency is worse than you think it is

Inconsistency in the way different companies report ESG data is a commonly cited challenge when trying to analyse the effects of ESG investment and performance. But what forms does this inconsistency take, and what kinds of problems does it create?

To provide more context, we selected a random sample of 50 large (Fortune 500) publicly listed companies across a variety sectors. For these companies we hand-collected information on how they report on the issue of Employee Health and Safety in their latest sustainability reports. As can be seen in Tables 1 and 2, we found more than 20 different ways that companies report their Employee Health and Safety data, using different terminology and, most importantly, different units of measure.

| Various metrics used to describe Employee Health and Safety |
|---|
| Lost time Incident Rate (per 100, 5,000, 200,000) |
| Lost Time Frequency Rate |
| Injuries Resulting in Lost Time |
| Accident Rate |
| Accidents Requiring Time Off, |
| Days Lost Due to Injury |
| Financial Loss Due to Accidents, |
| Injuries resulting in more than 1 lost day |
| Injury Rate |
| Lost Day rate |
| Lost time of Fewer than 61 days due to occupational illness or injury |
| Lost time incident severity rate |
| Number of accidents |
| Number of accidents without lost time |
| Number of lost workdays |
| Number of severe accidents occurred |
| Rate of Injury per 200,000 hours worked |
| Reduced working days caused by work related injuries |
| Time loss Claims |
| Lost time |
| Incident numbers of occupational disease, |
| Occupational Illness rate |
| Occupational Illness Cases |
| Occupational Disease Rate |

*Table 1. Various metrics used to describe Employee Health and Safety for a random sample of 50 large publicly listed companies.*

| Company | Metric | Unit |
|---------|--------|------|
| 1 | Number of accidents with fatal consequences | Number |
| 2 | Rate of injury per 200,000 hours worked | Number (ratio) |
| 3 | Occupational injury rate-related fatalities | Number |
| 4 | Lost-time incident frequency rate | Percentage |
| 5 | Injury rate | Percentage |
| 6 | Total case incident rate | Percentage |

*Table 2. Various units of measure to describe Employee Health and Safety for a random sample of 50 large publicly listed companies.*

This inconsistency poses a significant challenge when making comparisons among companies. Since these metrics are not necessarily measuring the same thing, it is not obvious which company would be a top performer on Health and Safety. To illustrate this point, consider again our normal distribution of performance. Instead of having one distribution, we now have multiple distributions, and those distributions might not be comparable (as shown in Figure 2). Which of these metrics best captures what good performance means when it comes to employee health and safety? If all of them are relevant, how should they be aggregated to describe a company's performance (equal basis, more weight on injury rate vs. number of accidents, etc.)? Another key point is that we assume that data is available for all the companies we want to compare, an assumption that is far from the reality as will be discussed in more detail below.
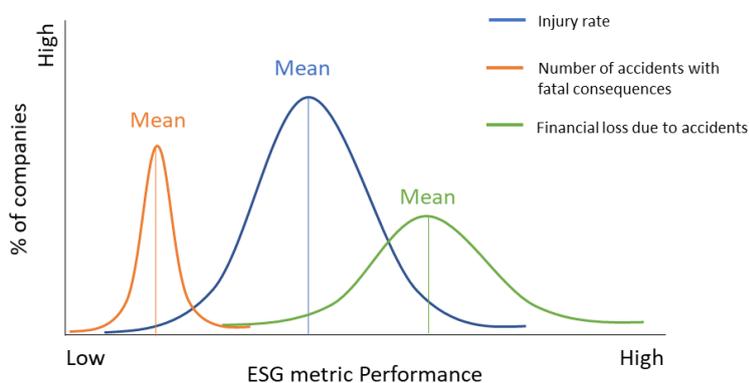


*Figure 2. Illustrative example of the challenge introduced when different ESG metrics are used to describe the same issue, in this case employee health and safety. These metrics might be*

*described by different distributions with different top and bottom performers, standard deviations and associated mean values.*

## 2. The mystic art of benchmarking (or defining your peer group matters)

So far, we have discussed how data inconsistencies can lead to different results when looking at the same group of companies. There is another key step in defining good or bad performance: the selection of the benchmark.

An ESG metric is a snapshot of performance that is assessed in relation to a range of values that defines best and worst possible performance for a given sample. Any ESG data provider will assign the best-performing companies on relevant ESG metrics the highest possible performance score, and the worst performing companies the lowest possible performance score, with the remaining companies all falling somewhere in between. Therefore, a critical decision point, as well as a potential root cause of discrepancies among ESG providers, is the definition of the range of best and worst performance that will determine the benchmark for the sample's scores.

Defining the range of performance can be done either by looking at a peer group, or by assessing absolute levels of performance based on a pre-defined "optimal" level of performance on ESG metrics. Both options and their implications will be discussed below.

### *Defining a peer group*

Although the definition of the peer group is entirely at the discretion of the data provider, there are a limited number of possible options:

- Universal peer group: A sample of companies across countries and sectors (e.g. MSCI ACWI). ESG metric performance calculated based on a universal peer group will unavoidably have an industry-level bias (for example, oil and gas companies will perform much lower than commercial banks on environmental issues).
- Industry peer groups: A sample of companies that belong to the same primary industry or subindustry. Benchmarking a company's ESG performance to an industry peer group allows providers to create more direct comparisons.

Regardless of how the peer group is defined, it is important to note that the range of performance observed in a peer group determines the final assessment of a company's performance on an ESG metric. For this reason, the definition of the peer group is crucial in determining the performance ranking of a company. This is especially so if different data providers use different peer groups either because they use different industrial classifications, such as the Global Industrial Classification System (GICS), MSCI IVA industries, or the Bloomberg Industrial Classification System (BICS), or because the sample they cover, and therefore the peers, are different. The following illustrative example can help visualize the extent of the problem.

The circle marks a company that we want to assess its performance on Diversity, based on the two peer groups. As shown in the figure, depending on the Peer Group selection and its associated distribution of performance, the same company would be assessed very differently. Compared to Peer Group A, the company will be closer to top performers and well above the mean value. Compared to Peer Group B, the company will be slightly above the mean.
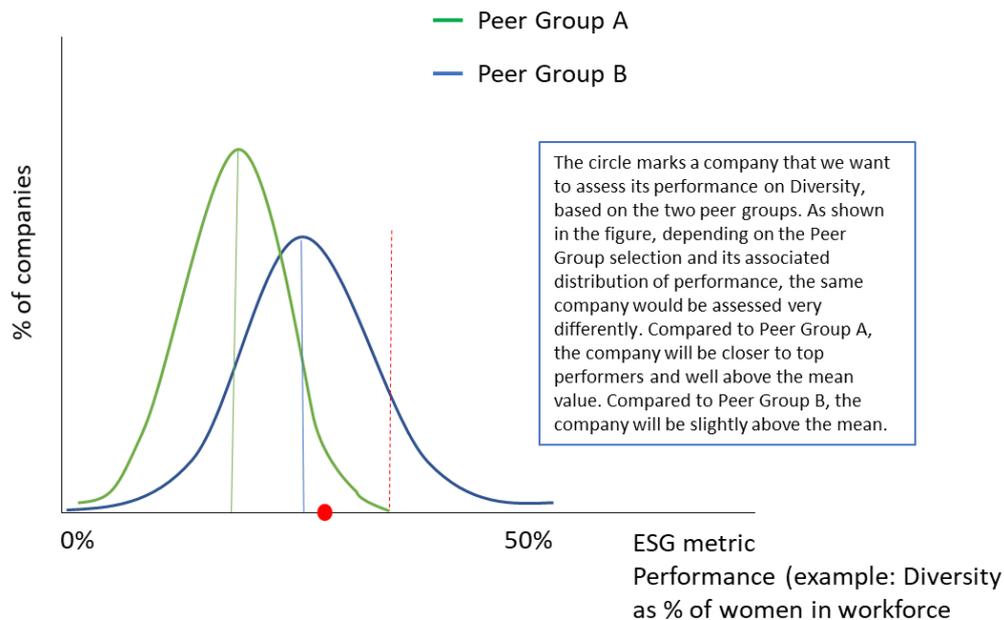
*Figure 3. Illustrative example that shows the importance of the Peer Group in assessing the relative performance of a company on any given ESG metric.*

Two different peer groups (A and B), as shown in Figure 3, are used to evaluate gender diversity, as measured by percentage of women in the workforce. The main difference is that Peer Group B includes more companies that perform at the top end of the distribution (top performers). Now let's assume that we want to understand a company's relative performance—represented by the red circle—on the given ESG metric, as compared to each of the two different peer groups. As shown in Figure 3, the same company would be assessed very differently in the two cases; it would be well above the mean value and close to the top performers within Peer Group A, whereas it would sit slightly over the mean within Peer Group B.

While changes in the tails of the distributions are straightforward to visualize, the implications for the companies falling within the range of performance are less clear and more heavily dependent on the specific formula applied to calculate ESG metric performance. Regardless of the proprietary scoring methodology applied, benchmarking sample performance to a peer group will, by construction, affect the final assessment of a company's performance.

### Defining performance

To overcome some of the issues posed by benchmarking, instead of defining a peer group to assess performance, another option is to assign scores based on pre-defined ranges of performance on a given ESG metric. And let's go back to the issue of gender diversity, measured again as the percentage of women in the workforce. Here the case for benchmarking arises from the reality that in certain industries (and countries), you find systematically higher levels of diversity than in others.

For the sake of argument let's assume that one designates 50% as the best possible performance on the metric, and 0% as the worst possible performance. Though straightforward

for metrics defined on an absolute scale, this approach makes sense to the extent that there is agreement about what best performance means, and whether this definition is truly "industry-agnostic."  However, for variables that by definition or construction are not bounded within a range of possible values, this approach implies a detachment from the reality of the data that could lead to distorted results.  Take the case of water usage and replacement.  How can one determine which companies deserve the best or worst performance on water replacement without a peer group for reference?

Setting pre-defined ranges of performance for ESG metrics provides a way to assess the real impact of a company to the external world.  For example, many companies around the world have made hundreds of commitments on deforestation and their environmental "performance" may well have increased as a result of their initiatives; but that hasn't stopped forests from disappearing. Similarly, a company might be scoring as a number two on diversity with 15% women in its workforce only because the number one performer happens to have 16%.  Again, this shouldn't be a case for celebration but rather a push to rethink and redesign assessment methodologies.

Initiatives such as science-based climate targets can provide some context for such ESG metrics, and possibly an answer to the question of how to define absolute benchmarks.  For example, data providers could use the level of decarbonization required to keep global temperature increase below 2 degrees Celsius and calculate for each industry what the absolute level of good performance means. That could provide companies with a clearly defined pathway to future-proof growth by specifying how much and how quickly they need to reduce their GHG emissions.

The reality is that the choice of the peer group significantly drives the meaning and interpretation of an assessment.  The lack of transparency about peer group components and observed ranges create market-wide inconsistencies in ESG metrics and undermines their reliability. Users need to be informed about the composition of the peer group and the corresponding range of values observed to appreciate the meaning of the assessment at hand.

---

**How about diversified businesses?**

Large companies might have revenue streams from various business units that sometimes do not fall under their primary industry classification. Take for example American Express, in 2017 the company had 82% of its revenues coming from the Financials Sector (Specialized Finance, Consumer Finance) and 18% of its revenues from the Information Technology Sector (Data Processing and Outsourced Services).  An assessment of American Express' ESG performance based only on metrics that fall under its primary industry is not reflective of all the risks and opportunities the company might face. There is currently no agreed method on how to handle diversified businesses, in terms of which ESG issues are material to them. Several data providers compare these companies' performance to peer groups from all industries where they create significant revenues (in this case comparing American Express to companies from the Technology Sector).

The question now becomes what if American Express displays leading performance on ESG metrics that are material within the Data Processing and Outsourced Services subindustry? Surely that should warrant its inclusion in a "reference" Peer Group that would be used to compare any other company within that sub industry.

---

### 3. ESG data imputation—not all models are fed equally

We have just established that even in the presence of disclosure it can be hard to judge what makes for a good or bad ESG performer. But what happens when there is no disclosure of a given metric? The lack of regulation and standardization around ESG disclosure implies that not every company will report on ESG issues, and those that will won't do it in a consistent way. Going back to our sample of 50 Fortune 500 companies, we manually collected raw data on the topic of employee health and safety (more specifically on the existence of a health and safety policy, the lost time incident rate, and the number of workplace fatalities). We found that roughly 50% of the companies report having a health and safety policy, and about 15% disclose their lost time incident rates and workplace fatalities. It is interesting to note that employee health and safety is a material ESG issue for 9 out of the 11 sectors, according to SASB's framework and that all these companies have large market capitalizations. And as one might infer from this number, disclosure of ESG issues tends to be much more limited in smaller companies.[1]

The implication, then, is that researchers and analysts face the problem of dealing with vast "data gaps" that span ranges of companies, time periods, and ESG metrics. So how can data providers claim to offer full coverage on ESG data and analytics for vast universes of companies and long time series?

This is where data imputation is likely to be playing an important role. Although rarely discussed by ESG data providers, their imputation approaches have the potential to be major differentiator among the providers because different gap-filling approaches can deliver drastically different results. Imputation directly affects the rankings of companies on ESG metrics; and depending on the assessment model applied, two different imputed figures can deliver significantly different ESG performance ratings. This is a significant part of the picture when looking at inconsistencies and disagreements across ESG ratings.[2] In the following paragraphs, we analyze some of the most widely used imputation approaches for ESG data gaps, from the least to the most computationally demanding.

**Rules-based**

The rules-based approach consists of an arbitrary assessment of a missing datapoint based on ad hoc rules constructed for a given ESG metric. For example, taking the metric "*Number of fatalities in the workplace*," one might decide to assume that all the missing values are 0, thereby taking an "innocent until proven guilty" approach. A different rule could be to apply regional or sector averages to avoid promoting non-disclosure incentives. Examples of such rules are endless, since they are metric-specific and can be combined to create different layers of complexity as needed. For example, one could observe that, on average, banks operating in heavily regulated countries will perform above industry average on the issue of "business ethics" and therefore assign a higher performance score for banks in those countries that do not disclose.

This approach is the simplest in terms of computational efforts, since no advanced statistical knowledge is needed. However, a joint understanding of business, industry

---

[1] https://www.sasb.org/standards-overview/materiality-map/

[2] ACCF, July 2018, Ratings that don't rate – The subjective world of ESG ratings agencies

dynamics, and specific ESG issues is required to create rules that can deliver meaningful imputations.

## Input-Output Model

The input-output approach to estimating overall performance of a company on a given ESG metric relies on relevant industry-specific and macroeconomic level data. Industry-level data is scaled for the company's size of operations to get an estimate of its direct impact. Additionally, to estimate impact deriving from production and handling of the company's main inputs, macroeconomic level data is used to estimate the flow of goods and services across sectors. An input-output model is then able to assign performance on a given ESG metric based on the estimate of a company's direct and indirect impact of business activities. These models are in fact best suited for scientific environmental metrics in cases where top-down methodologies and assumptions like the ones embedded in the input-output model hold and macroeconomic data is accessible.

Nevertheless, such applications are limited by the kinds of ESG metrics that the input-output model can be applied to. Scaling down industry averages to estimate company-level figures is likely to be much more challenging in the case of social metrics, whose correlation with firm size or scale of operations is much less clear than in the case of environmental metrics.

## Statistical

When discussing more advanced statistical imputation models that go beyond mean imputation,[3] the door opens onto a vast landscape of more or less computationally-demanding possibilities. Single imputation processes impute data for each missing ESG metric only once. Multiple processes impute data more than once, run the desired statistical analysis on each imputed dataset, and finally aggregate the different outputs into one set of statistical estimates.[4] Multiple imputation takes better account of variability in the unobserved data and thus is able to provide more accurate imputations.[5] (We have included an appendix with more information on these different statistical imputation techniques.)

*Regression methods:* Regression methods estimate an ESG metric's missing values based on the inferred relationship between its observed data and a set of time-varying and non-time varying predictors. To achieve this, a regression model is set up for each ESG metric with missing data (dependent variable, Y) to regress its information against a set of predictors (independent variables, X). Regression methods include several variations, such as the single imputation approach and multivariate imputation via chained equations (MICE).

*Predictive Mean Matching (PMM)*: Predictive mean matching represents an attractive machine learning alternative to regression imputations, both in the context of single and multiple

---

[3] Like a qualitative rules-based approach, mean imputations replace all missing data with the mean of observed values across a specific dimension (sector, industry, geography, size, etc.)

[4] Little, R.J.A and Rubin, D.B., 1987, 'Statistical Analysis with Missing Data', John Wiley & Sons, NY
Schafer JL, Graham JW, 'Missing data: our view of the state of the art', Psychol Methods, 2002 Jun, 7(2):147-77

[5] Multiple imputation requires optimization processes that are computationally heavier to support, and assumptions need to be made on how to pool results into one set of outputs. For a thorough overview of multiple imputation methods, see van Buuren (2012)

imputation. PMM still employs linear regressions in its algorithm; and for each missing entry, the method finds a set of observed values that are potentially similar to the missing entry and from this set it chooses a value to impute.[6]

To illustrate the different answers that different techniques are likely to give, consider the following example of the large airline company Lufthansa. The reported employee turnover number for Lufthansa in 2017 was 12.9%. This number was increasing over time as the company restructured its operations and labor force.

What if we used different models to estimate this number when we did not know it? Starting with the rules-based approaches, one could simply assign Lufthansa the worst performance within its GICS industry. If one assumes that the highest turnover is the worst performance, then the imputed value would be 20%, which was more than 7% higher (worse) than the actual value. But if one instead assumed the average value across companies in the industry, that would have been 8.9% which would understate Lufthansa's turnover by more than 4%. Using regression methods that account for country, industry, sales, profitability, market multiples, and total number of employees, and allow for five iterative imputations and averaging across them would give us only 4.0%, an understatement of almost 9%. And the PPM method would give us an answer of 4.6%, an understatement of 8.3%. Of course, such models can work well when a company is highly representative of the average firm for the variables modelled, but the point remains that the rules and models one chooses can provide very different answers.

| Lufthansa real turnover in 2017: **12.9%** | Imputation method | | |
|---|---|---|---|
| | Rules based | Regression method | PMM method |
| | Worst performance within industry: **20%** | Highest: **12.9%** Lowest: **-8.5%** Average of 5 iterative imputations: **4.0%** | Highest**: 5.97%** Lowest**: 1.38%** Average for various imputations**: 4.61%** |
| Difference from real | +7.1% | | |
| | Average value within industry: **8.9%** | | |
| Difference from real | -4% | -8.9% | -8.3% |

Another major problem with data imputation and gap filling is that because large capitalization companies are more likely to have the resources to report comprehensively on ESG issues, their values are likely to impart an upward bias to the observed averages, while the observed range fails to reflect the worst performances. And to the extent this is so, most imputation techniques will be inappropriate since they cannot estimate data with any degree of confidence outside the range of observed values.

---

[6] C Yuan, Yang. (2005). Multiple Imputation for Missing Data: Concepts and New Development.

### 4. ESG data providers disagree – mostly so when there is publicly available information

A reasonable question to ask is whether as disclosure has improved dramatically in the last ten years, we have made progress towards defining what we consider good or bad performance. Another way of thinking about this is to ask which effect seems to be more dominant—the general lack of agreement among the metrics we use, or the lack of disclosure and the resulting reliance on imputations. If the lack of agreement about metrics is the problem, we expect that when evaluating companies that disclose more information, ESG analysts might use different metrics and therefore disagree more on their assessments of those companies. But if the main problem is really lack of disclosure, then we would expect less disagreement among ESG analysts when evaluating companies that provide more comprehensive ESG disclosure.

A recent study found strong evidence that it is companies that provide more ESG disclosure that tend to have more disagreement, or variation, in their ESG ratings.[7] To us this constitutes clear evidence of the need for not only more effective disclosure, but also for a clearer understanding of what different ESG metrics might tell us and how they might best be institutionalized for assessing corporate performance.

**Implications for Companies**

What can companies do to address these problems with ESG data and how it is currently being used and, in many cases, misused? We have two main suggestions:

(1) Take control of the ESG data narrative.

(2) Accept a reasonable baseline of ESG metrics and self-regulate in ways that aim to provide comparability.

We understand the value for companies to being able to convey the uniqueness of their business models by "customizing" their reporting practices to some degree. At the same time, we believe that most companies should be able to accept and work within a reasonable baseline for reporting standards. Organizations like the Sustainability Accounting Standards Board (SASB) have made significant progress in providing such a baseline. Taking control of the ESG data narrative can also help with one of the major frustrations of sustainability departments, which is "survey fatigue."

Many companies have identified the need to come together as industries and take control of the narrative that gets communicated to their investors. An example of such efforts can be found within the electric utilities industry in the U.S. The Edison Electric Institute, the association that represents all U.S. investor-owned electric companies, assembled a working group of companies and investors to develop industry-focused and investor-driven ESG reporting practices.[8] The outcome of the working group was a simple reporting template that includes an excel-based tool for utilities to use when reporting qualitative as well as quantitative information.

Taking a similar approach, a project run by KKS Advisors and sponsored by the Rockefeller Foundation is creating Industry ESG working groups that aim to bring together

---

[7] Dane Christensen, George Serafeim and Anywhere Sikochi, 2019, "Why is corporate virtue in the eye of the beholder? The case of ESG ratings," Harvard Business School Working Paper.

[8] Edison Electric Institute: http://www.eei.org/issuesandpolicy/finance/Pages/ESG-Sustainability.aspx

leading companies within an industry and their major long-term investors to agree on certain ESG issues and metrics. The goal is to promote "pre-competitive" collaboration among companies to solve common sustainability issues by developing industry standards, generating data, and creating industry knowledge.[9] To achieve this goal requires effective communication with investors.

## Implications for Investors

For investors, our message is to push for meaningful disclosure of metrics. Like companies, investors need to come together and narrow the demand for ESG data into manageable and meaningful disclosure of metrics. In the past, investors have made vague and, in many ways, unfocused requests for data. As a recent study found, the most important barrier for the use of ESG data in investment decisions is the lack of comparability of metrics across companies and across time.[10] We feel that it is time for investors to reach agreement on a baseline level of indicators and metrics that would be informative on a core set of ESG issues that are of prime importance, such as climate change, labor conditions, and diversity.

## Implications for Stock Exchanges

Stock exchanges should give serious consideration to issuing guidelines for and even mandating ESG disclosure. The exchanges can be the coordinating mechanism, working with companies, investors, and regulators to design smart disclosure guidelines. The Sustainable Stock Exchanges initiative is an effort taking us in that direction, exploring how exchanges together with investors, regulators, and companies can enhance corporate transparency.[11] (As an example, consider the Greek stock exchange's plan to issue ESG guidelines for listed companies in the Athens Stock Exchange that is described in this issue.)[12]

## Implications for Data Providers

Data providers need to come to an agreement on best practices and become as transparent as possible about their methodologies and the reliability of their data. In discussing the methods they use to assess a company's performance, data providers should include not only a list of material issues and a description of their scoring methodology, but more detail on the peer groups used, and clearly distinguishing between real and imputed data. Data providers could also establish some best practices for assessing performance that could be followed by the whole industry—which could be especially helpful in the case of diversified businesses.

---

[9] Serafeim, George. "Can Index Funds Be a Force for Sustainable Capitalism?" *Harvard Business Review* (December 7, 2017).

[10] Amel-Zadeh, A. and Serafeim, G., 2018, "Why and how investors use ESG information: Evidence from a global survey," *Financial Analysts Journal*, *74*(3), pp.87-103.

[11] UN Sustainable Stock Exchange Initiative: http://www.sseinitiative.org

[12] Tania Bizoumi, Sokrates Lazarides, and Anatassa Stamou, "Innovation in Stock Exchanges: Driving ESG Disclosure and Performance," *Journal of Applied Corporate Finance*, Vol. 31 No. 1 (Spring 2019).

**APPENDIX**

Regression methods

Regression methods estimate a variable's missing values based on the inferred relationship between its observed data and a set of time-varying and non-time-varying predictors. To achieve this, a regression model is set up for each variable with missing data (dependent variable, Y), to regress its information against a set of predictors (independent variables, X). The regression output will provide an estimate of the marginal contribution of each predictor in explaining variability in the observed data, conditional on all other predictors remaining constant.

Multivariate imputation via chained equations (MICE) allows you to designate a model to every variable that needs to be imputed (whether they are continuous, categorical, nominal, etc.), and the number *m* of imputed datasets to create. By design, the method differentiates from single imputation by initializing the imputation via random sampling first (coefficients are randomly drawn from a distribution of possible coefficients), thus adding a stochastic feature to the imputation that better captures variability in the unobserved data and thus allows for a more valid statistical inference (Rubin, 1996; Buuren and Groothuis-Oudshoorn, 2011).
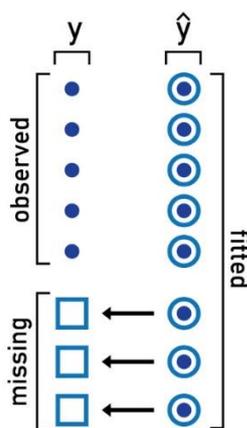


Regression-based imputation is straightforward but poses a number of challenges. Firstly, the choice of explanatory variables needs to be tailored to the imputed variable.[13] Model selection also introduces potential cases of endogeneity and model misspecification. Cases of model misspecification and other endogeneity problems can hover over model selection and lead to biased estimators, which, in turn, leads to biased data imputations. Another issue with regression-based imputation on a more practical level is the breadth of missing values in the dataset. There is no test or rule to determine how many observations one needs in order to produce reasonable coefficient estimates and with

*Regression-based Imputation:*
*Fitted values of Y estimated from*
*the chosen regression model are*
*used to impute missing values.*

reasonable power.[14] However, a rule of thumb suggests that in extreme cases of data sparsity like the ones observed with the majority of ESG metrics, and without a consistent time series, regression-based single imputations become tricky and suboptimal.
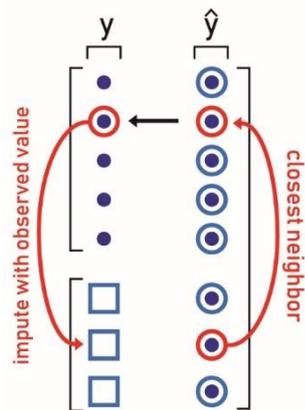
Predictive Mean Matching (PMM)

Predictive mean matching (PMM) represents an attractive machine learning alternative to regression imputations, both in the context of single and multiple imputation. PMM for single imputation still employs linear regressions in its algorithm, but instead of using directly the fitted value from the regression to replace a missing data point, it "imputes an observed value

---

[13] Van Buuren (2012) provides guidelines for choosing the set of independent variables.
[14] Frank Harrel: "Regression Modelling Strategies"

which is closest to the predicted value from the simulated regression model for each missing value" (Yuan, 2005).



*Predictive Mean Matching: Fitted values of Y from the regression algorithm are used to identify the donor observation with the closest fitted value*

In the context of multiple imputation, PMM can be used in place of simple regression models. The advantage of using PMM lies in safeguarding the nature of the observed data and thus dominating regression-based imputations whenever variables are not normally distributed (Horton and Lipsitz, 2001). PMM is gaining traction in the imputation space, as it is able to impute all kinds of data based on its observed distribution, making it an implicit model and thus less subject to misspecification problems (van Buuren, 2012). Despite its advantages, limitations need to be considered before applying the method blindly to the data. Firstly, the main assumption behind PMM is that data is missing at random (MAR). Data is MAR when a variable's *real unobserved data* is not correlated with the same variable's observed data, but can be correlated with observations on other variables. For example, data on *percentage of female employees* is *not* MAR if the reason for its missingness is that companies that are not doing great on employee diversity systematically fail to disclose that information. As there is no objective way to determine the reason for the missingness of data, this assumption already poses a big obstacle. By design, PMM imputes values drawing from the range of observed data: wrongfully assuming data to be MAR and imputing it via PMM would lead to significantly biased estimations. In the instance of ESG data, specifically, the lack of thorough regulation and mandatory ESG fields to disclose makes it hard for the MAR assumption to hold in practice. Another challenge of PMM is posed by the choice of neighbours to look for, as it embeds a trade-off between designing biased estimators by choosing too many neighbours and producing high standard errors by choosing too few.