Bias assessment and mitigation for ESG scoring models

Mathieu Joubrel

Under the supervision of:

Nicolas Mottis

19/07/2021 - Oral presentation

Public report

Executive summary

This paper investigates bias detection methods for ESG scoring models. We show how to assess and quantify the model biases both relatively to a group of peers and in absolute value. We separate endogenous bias factors used as input in the model and can be modified by analysts, and exogenous factors such as the company size, activity sector, and main geographical area. We propose a methodology to mitigate these biases and compute a score that is independent of the chosen factors. Finally, we propose a framework to detect the companies whose intrinsic performances outperform their peers independently of the exogenous biases of the model.

Preface

One of the key drivers of the finance industry over the past few years has been socially responsible investment. From a few pioneers 15 years ago, 4000 financial institutions are now UN PRI signatories and the total assets managed by signatories reach more than 120T$. This trend shows the growing public commitment of the finance industry to ESG and responsible investment.

The rapidly evolving ESG landscape and growing market demand for ESG products entail new risks for all stakeholders in the ecosystem. Standards, frameworks and initiatives have multiplied to the point that claims from investors are almost impossible to monitor and compare properly. Many asset managers develop internal proprietary scoring models to make up for this opacity and justify their ESG commitment.

As available ESG scores from rating agencies already show low correlation (Aaron K. Chatterji, Rodolphe Durand, David I. Levine, Samuel Touboul, 2015), the multiplication of proprietary models is likely to amplify this phenomenon. The raw data underlying these scores also still lacks the quality to support efficient decision-making. Classic issues for designers of ESG scoring models include data availability, transparency and update frequency, or even the lack of a consensual proxy to measure a chosen indicator (biodiversity, happiness…).

The wide range of scoring models flooding the market is not harmful as such, as every investor may be concerned by different aspects of the ESG performances of a company and therefore when attributing it a score. The issue rather comes from the lack of transparency on the methodologies, biases and interests of the raters that only use their scores internally.

Scoring models are algorithms developed by finance professionals to aggregate raw ESG data into ESG scores that can support investment decisions along with a financial analysis. As there is no global standard or regulation regarding these algorithms' structure and features, major biases can be introduced in the company and market assessments of asset managers through their ESG models.

The main sources of divergence in ESG ratings have been studied by Florian Berg, Julian Kölbel and Roberto Rigobon in a 2019 paper (updated in 2022) on scores provided by scoring agencies. They identified three of them:

- Scope: what do you measure?

Scope refers to the set of indicators involved in the ESG score computation.
It accounts for 38% of the scores divergence.

- Measurement: how do we measure the issues and indicators?

Measurement refers to the exact metrics used to evaluate the chosen scope and how to measure them. It accounts for 56% of the scores divergence.

- Weights: how do the selected issues and indicators compare to each other?

Weights refer to the importance given by the model to each metric.
It accounts for 6% of the scores divergence.

The scope and weight-related divergences are merely the expressions of the personal preferences of the raters on extra-financial scoring. It cannot be addressed through regulation and standardisation alone, as it would amount to restricting the expression of their responsible investment thesis. Scope and weights are the levers investors can use to build a differentiating ESG scoring model to express their views and strategy.

To justify their extra-financial decisions, they need to be able to validate the conformity of their model with their scoring thesis. It involves knowing how they compare to their peers to assess exactly what their model captures and how they differentiate from them. It would bring a clearer structure to the ESG market and make it possible for all stakeholders to identify relevant signals and levers to improve their ESG performances, be they corporates or financial institutions.

Methodologically, this report is divided into three parts. First, we give practical methods and implementation details to compare a model with the market or a peer group. We do so to identify biases in any direction compared to a benchmark. Second, we analyse a given model regardless of the rest of the market to assess its endogenous biases, that is to say, the biases arising from its input data. Finally, we link the model biases to exogenous features, namely the company size, country and sector of activity, and explain how to transform the model's outputs to correct those biases and enrich the resulting analysis.

Table of content

Table of figures

Figures

Tables

I. Bias assessment by comparison

A. The data

1. Description

The results presented in this paper are based on the analysis of a set of 2903 listed companies worldwide. We use different data sources to build a realistic market view of these companies:

- The aggregate ESG scores of three major data providers,
- The aggregate ESG scores and quantiles of European asset managers and asset owners, extracted from the proprietary database of Valuecometrics[1],
- Metrics and raw data from Refinitiv[2].

Regarding the metrics used in parts II and III to train the weights estimation models, we extract all the available ESG data for all the companies in our universe. We then discard the metrics disclosed by less than 30% of the companies to end up with a list of 92 metrics (Annexe 1). We use the same set of metrics from the same data provider throughout the paper. It allows us to focus on the ESG models' scope and weight biases as stated in the preface.

We pick at random one of the nine proprietary models from the asset managers who contributed their data to be the « reference model » used in the following comparisons and analyses. All the other models are considered to belong to anonymous market players against which we benchmark the performances of the reference model.

2. Preprocessing

We preprocess the scores to get rid of systematic under- or over-scorings that would not amount to any relevant bias in the models. For instance, if a model A

[1] ©2022 VALUECOMETRICS™, https://en.valuecometrics.com

[2] https://www.refinitiv.com/en

gives the same score as another model B minus 10 points out of 100 to every company, the actual scores are radically different even though the biases are about the same.

We therefore choose to normalise the scores given by each model on the investment universe. We transform the output of the models so that its average score on the investment universe is 0 and its standard deviation is 1. We then compute the mean μ and the standard deviation σ of the distribution of the scores on the market for each company in our investment universe:

$$x_{norm} = \frac{x - \mu}{\sigma}$$

In this case, the scores can be both positive and negative. This normalised score keeps the shape of the scores distribution of the model as well as the relative score differences between the companies in the investment universe. As a result, the underlying biases are still present and we did not lose any important information with this transformation.

Many investors look at ESG scores, but also at the resulting company ranking in its investment universe. However, being ranked 100th by a model that scores 300 companies is not the same as being ranked 100th by one that scores 2000. As a consequence, we use the quantiles provided in the Valuecometrics database. They can be considered as a proxy for ranks that do not depend on the size of the investment universe. Being in the q-quantile means that a company scores better than q% of all the other companies in the investment universe.

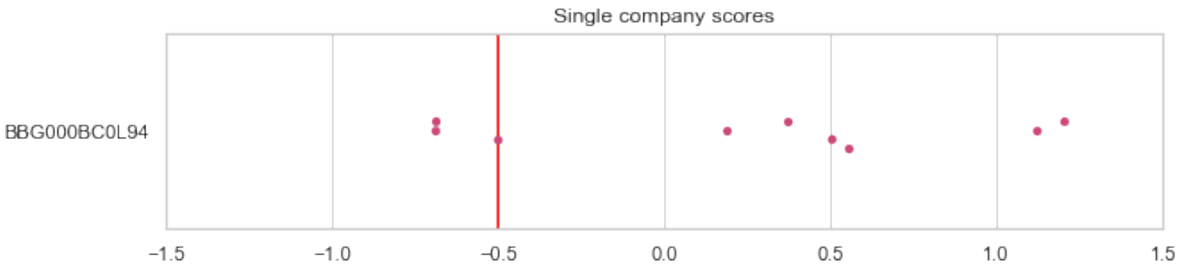
Regarding the metrics gathered from the Refinitiv platform, we do not modify the boolean (Yes/No) values and use one-hot encoding for the four categorical metrics available for at least 30% of the companies: Country, Sector, Board Structure Type and CO2 estimation method. We use the log value of the latest declared yearly turnover to account for company size. We then divide the dataset into a training and a testing set and normalise all the numerical values to improve our weights estimation model's performances. The point of dividing the sets before normalising is to avoid any information leakage from the testing to the training set through a normalisation that takes into account the mean and standard deviation of the scores of the testing set. We also map the countries of activity provided by Refinitiv to a more relevant set of countries, as China alone has 23 different categories under its name (such as ‘China’, ‘People’s Republic of

China' or 'PRC'). This operation brought down the number of countries represented from 206 to 74.

B. Comparison levels

1. Company-level visualisation

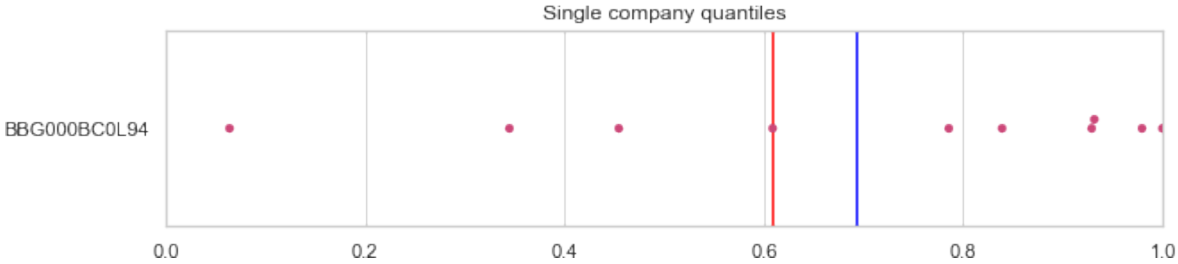
We begin by visualising the models' results at company level to compare the reference model to the rest of the market. A first simple plot consists in visualising all the normalised scores assigned to a specific company by each of the market players to see where the reference model stands compared to them. Each point on this plot is a score given by a model to the company. We use the normalised score for this plot to make comparisons more relevant:



Plot 1: Normalised ESG scoring distribution of a single company

We specify the score of the reference model with a red vertical line. The denser zones are those where more models find a consensus on the ESG score of the company. A score of zero is average on the normalised scores, so this company is considered above average by most market players, but below average (by half of a standard deviation) by the reference model. Almost all the scores are contained within one standard deviation of the score distribution, so the market consensus for this company is rather strong and no model in this sample can be considered to be a strong outlier of the distribution.

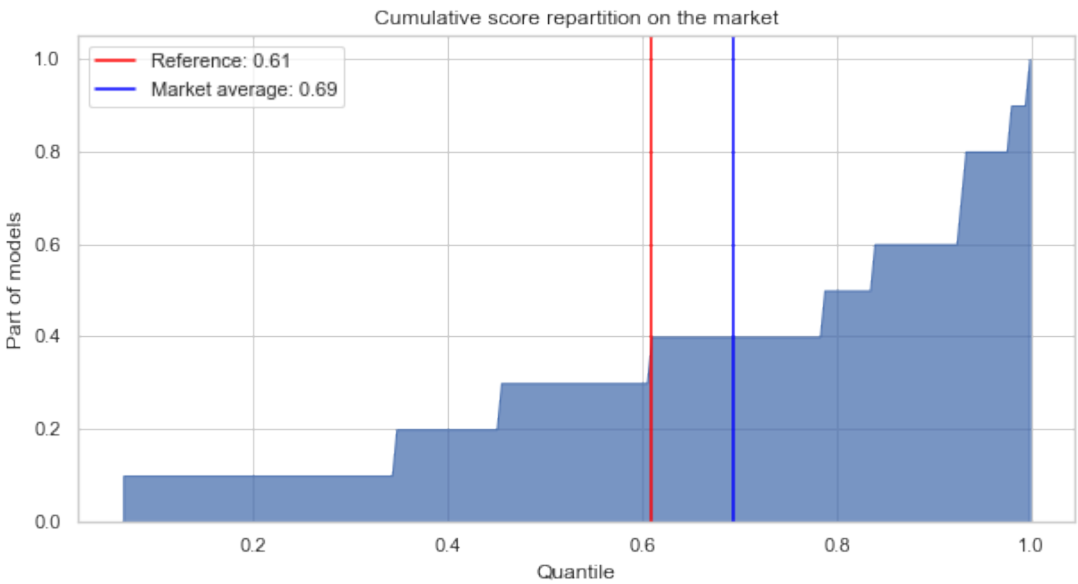
Another way to visualise this result is by looking at the quantiles. Here is the bee swarm plot summing them up:



Plot 2: Quantile distribution of a single company

This plot is interesting for investors who have more interest in the ranking of the companies than in the scores themselves. We can read that this company is in the top 22 % of companies in the investment universe of the raters on average. It reaches the top 39 % of companies in the investment universe of the reference model.

To have a more precise view of the quantile analysis, we can take a look at the cumulative histogram based on the same data. Here, we associate for each quantile on the x-axis the part of investors that gives the company a quantile that is equal or worse. If the curve is rather convex, it means that the company is badly ranked by a large proportion of the models. If it is rather concave, it means on the contrary that the company is well-ranked by most models. Comparing the area under this curve for different companies gives a quantitative measure to detect the ones that are more broadly praised by the market:



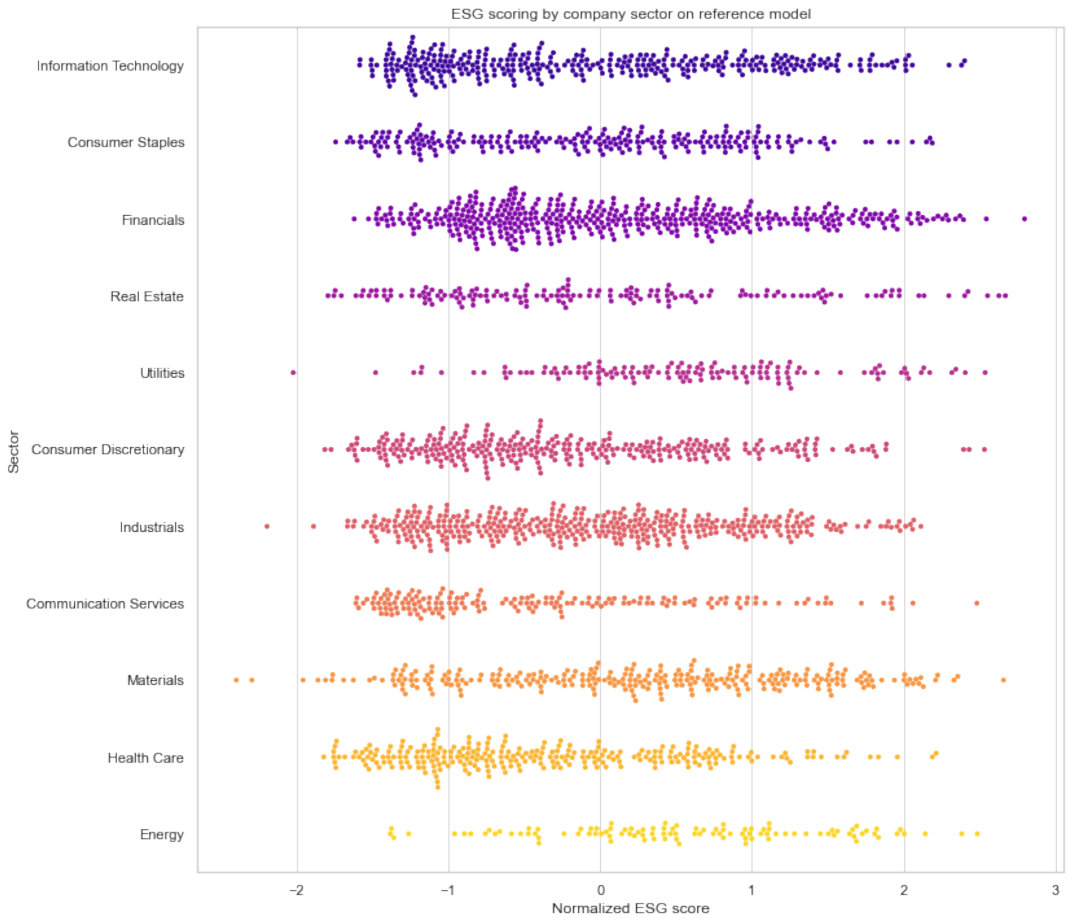
Plot 3: Cumulative quantile distribution of a single company

The quantiles indicated for the market and the reference model are the same as in the previous plot. This company is rather above average in both score and rank, but the convexity of the curve cannot be concluded with precision. More data points would be required to compare this company with a peer with a high level of certainty. The additional information we can find on this plot is that about 40 % of all market players associate this firm with a worse score than the reference model, and 40 % of them rank it worse than the market average.

These plots make it possible to assess how a company is perceived by the market and to compare it to the results of the reference model. They can be useful to make an investment decision involving a few companies, but we need to analyse the results of the reference model on a bigger scale to detect its biases. In the next paragraph, we propose similar plots averaged over activity sectors.

2. Sector-level visualisation

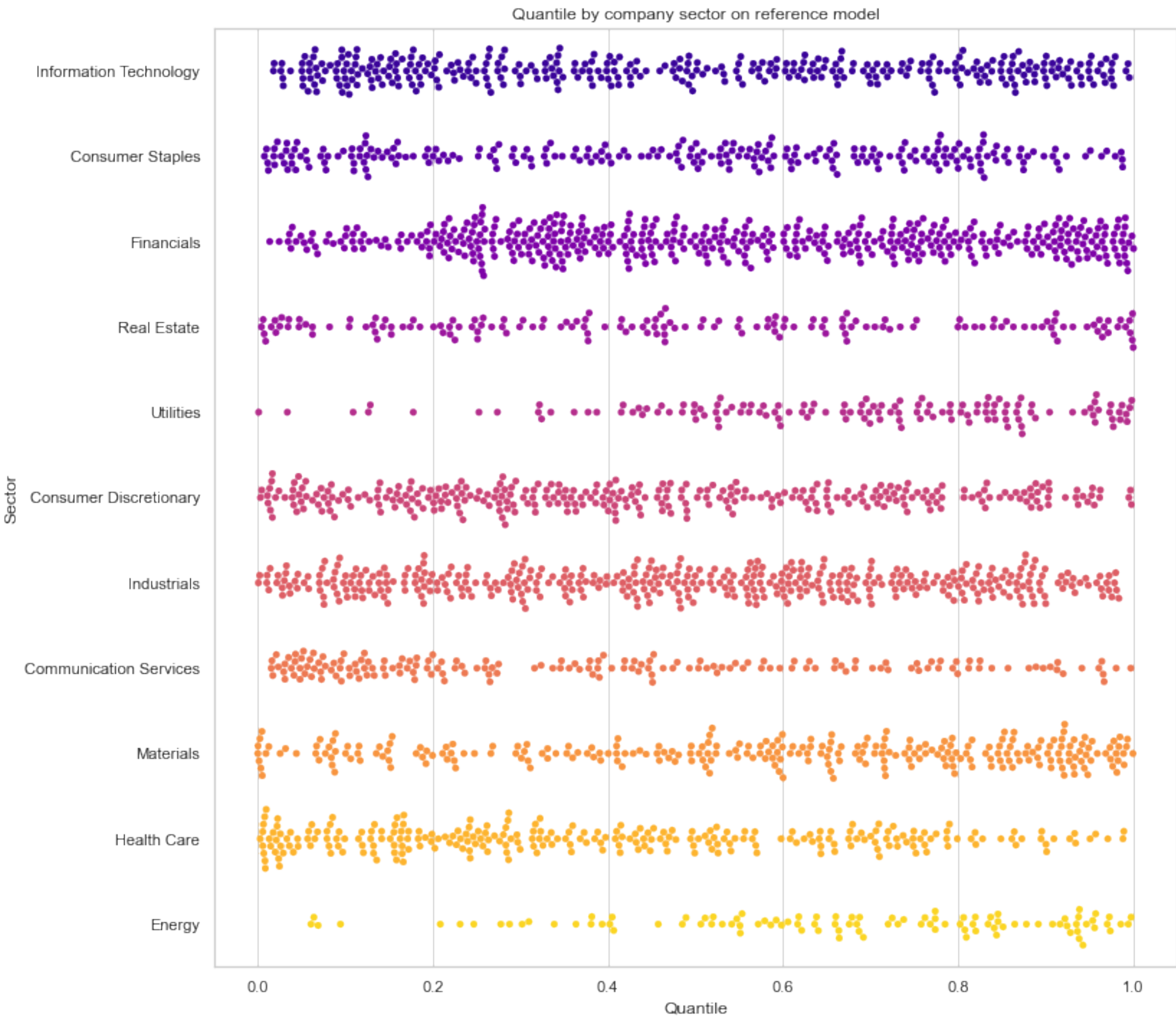
The first simple way to try to visualise the reference model's biases is to plot the normalised scores of all the companies in the investment universe. Unlike the previous plots, each data point here is a different company, all graded by the reference model, instead of different models grading the same company:



Plot 4: Normalised ESG scoring distribution of the reference model by sector

We can see that some sectors tend to be systematically underrated, such as Information technology or Healthcare. Some others are overrated, such as Utilities or Energy. With an unbiased best-in-class approach, all the categories should have a similar score distribution, no matter the number of companies in each category. None should be advantaged by the model compared to the others.

Once again, we can plot the same data with the quantiles instead of the normalised scores. This plot is easier to read as quantiles are always reported from 0 to 1: data points are more widespread across the plot. Moreover, the quantile computation takes into account the score of the company, but also that of the rest of the investment universe. As a consequence, the insight we get from this plot is more relevant than the previous one to detect biases :

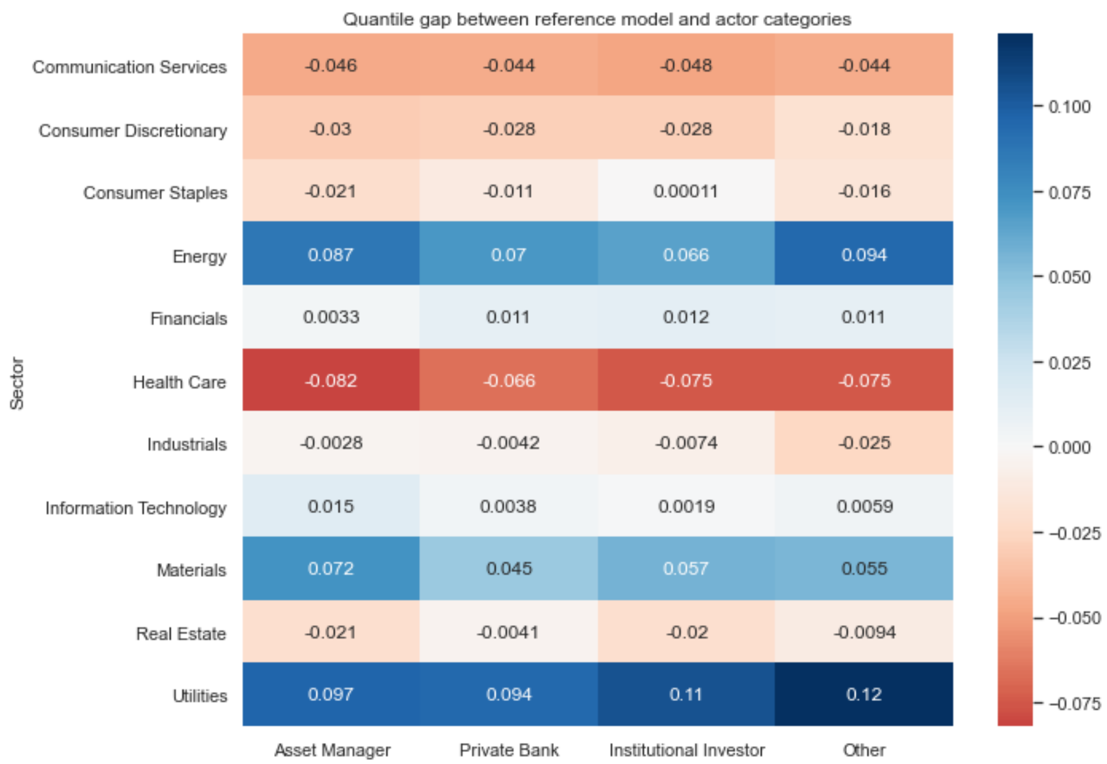


Plot 5: Quantile distribution of the reference model by sector

We can see on this plot that the rankings of Information Technology companies are actually rather well-balanced, and those in Healthcare are indeed disadvantaged. The trends are much easier to see, with Utilities, Material and Energy being put forward while Communication Services are lagging behind.

This plot makes it possible to stress the internal biases of the reference model, which are relevant for finance professionals who adopt a best-in-class strategy. They want to make sure all the activity sectors are well-balanced in order not to promote a specific sector. This is not as important for other investment strategies.

To have a more quantitative view of these biases, we can visualise the difference in quantile between the reference model and the market, on average for all the companies in an activity sector. We divided the market into four categories to analyse the differences between them: asset managers, private banks, institutional investors and others. If the owners of the models are known, we can choose any other categories to sort them. The point here is to demonstrate how we can cross two categorical variables to analyse our dataset. In this specific example, we compare the reference model to different categories of competitor models in each sector of activity:



<u>Plot 6: Sector/model owner segmentation of quantile differences between the reference model and the market average</u>

We can read that the reference model ranks companies from the Energy sector on average 4.1% higher than asset managers, and 4.6% higher the private bankers. We coloured in blue the cases where the reference model over-ranks companies compared to the market and in red those where it under-ranks them.

The biases that we conjectured with the previous plot are confirmed when compared with the market averages. The sectorial biases are consistent throughout the different market player categories.

We computed this heat map using two categorical variables: activity sector and model owner. To have granular comparison metrics, we can choose any other two categorical variables and plot the same heat map with one on each axis: company country, board type, listed company or not… Even numerical data can be turned categorical by defining relevant boundaries. For instance, if we decide to choose the number of employees to account for the company size, we can create categories by sorting companies into the 1 to 499, 500 to 4,999 and 5,000+ employees categories. This variable can then be used to plot a heat map with another categorical variable.

Plotting all the possible heat maps given the model input data would be possible, but we do not have access to all the relevant data here. This work can however be done by the model owner to detect the synergies between the input variables that can lead to major biases. For instance, a model can be biased positively towards companies that have good results in gender equity, but mostly in the tech sector.

The problem with this simple plot is that it only involves two variables, and some biases can occur only in specific cases involving more than two dimensions. We will investigate deeper in part II how to overcome this issue.

C. Quantile analysis

1. Motivation

A basic way to assess the relationship between two sets of scores is to compute the Pearson correlation between them. It measures the linear relationship between two distributions with a correlation that varies between -1 and +1. Positive correlations imply that as one score increases, so does the other. Negative correlations imply that as one score increases, the other decreases. At 0, the two sets are independent, and at 1 or -1 there is an exact linear relationship between them. The two sets we chose are the scores of the reference model and the average score of each company as seen by the other raters. We compute the Pearson correlation between them and average the results for each sector of activity.

However, correlation only gives a general idea of the degree of consensus between two model outputs. If the correlation is weak, we do not know whether it comes from small differences in a large number of companies or major disagreements in a few of them only. In order not to overlook company-level differences we also computed the Mean Absolute Deviation (MAD) with the normalised score, which is the average absolute distance between the average score and the individual scores of the models. Once again, we computed it for all the sectors of activity covered by the reference model:

ACTIVITY SECTOR	CORRELATION	MEAN ABSOLUTE DEVIATION
Information Technology	88.7 %	0.386
Consumer Staples	86.4 %	0.387
Financials	85.9 %	0.389
Real Estate	86.1 %	0.443
Utilities	86.3 %	0.41
Consumer Discretionary	83.6 %	0.426
Industrials	86.5 %	0.374
Communication Services	89.7 %	0.34
Materials	89.0 %	0.399
Health Care	85.8 %	0.413
Energy	84.8 %	0.407
Average	86.7 %	0.395
Minimum	/	0.125
Maximum	/	1.155

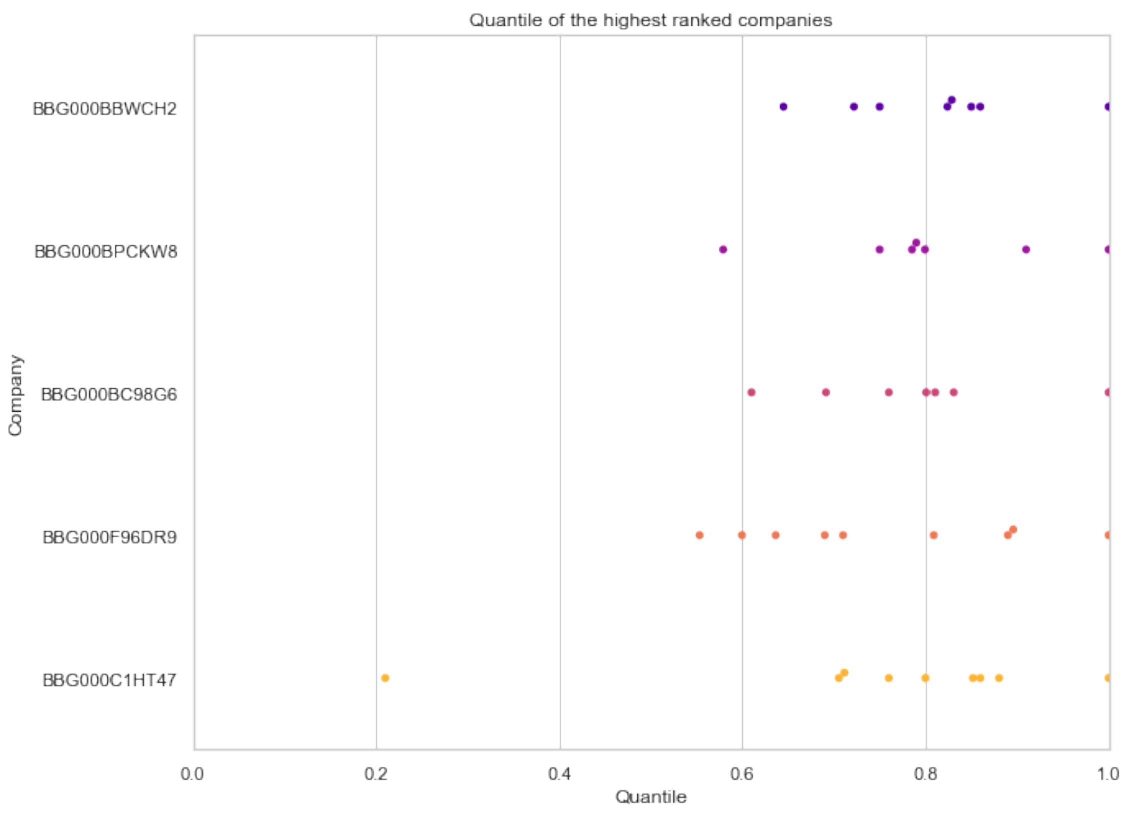
Table 1: Overview of Pearson correlation and MAD on the normalised scores

We observe that there is no significant variation between sectors, both for the correlation and the MAD. To analyse further the drivers of the homogeneity or the heterogeneity of the models, we plot the scores of the companies for which there is the most and the least disagreement among the available models. We measure this disagreement with the MAD, as it accounts for the average difference between an individual score and the average score on the market. These plots are available in Annexe 2.

The 5 companies triggering the least disagreement operate in 5 different sectors (Financials, Materials, Health Care, Information Technology and Communication Services), with normalised scores ranging from -1.7 to +1.9. They have an average MAD of 0.23. The 5 companies triggering the most disagreement operate in 4 different sectors (Consumer Staples, Real Estate twice, Information Technology and Communication Services), with normalised scores ranging from -3.7 to +1.8. They have an average MAD of 1.28.

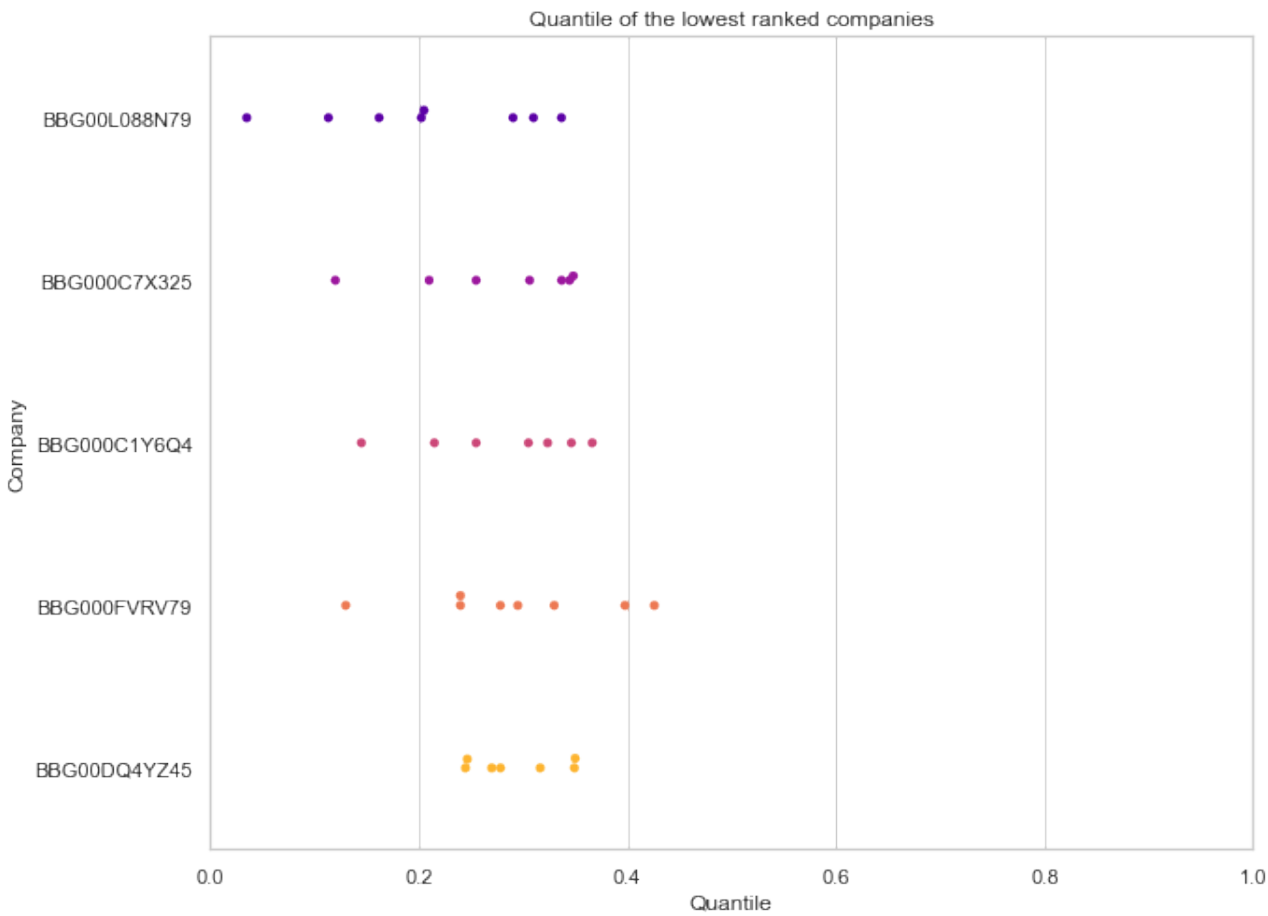
The is no obvious driver for the disagreements between models. In this small sample, the sector of activity and the average rating does not seem to be linked to an extreme agreement or disagreement. The same goes for the country of activity and market capitalisation.

As many investors are interested in the rankings of companies rather than their absolute scores, we now analyse the quantiles distribution to try and find market-wide biases. We plot the quantiles distribution of the companies that are the most systematically ranked in the top quantiles by the models available:



Plot 7: Quantiles distribution of the 5 best-ranked companies on the market

The 5 best-ranked companies are almost always ranked in the top 40% of the investment universe of all the models. They come from 5 different activity sectors (Information Technology, Materials, Financials, Energy and Real Estate) and have 5 different countries of activity (France, South Africa, Italy, USA, Australia). Once again, there is no obvious criterion that seems to drive the rankings of these companies up. Let us now plot the companies that are the most systematically ranked in the bottom quantiles by the models available:



Plot 8: Quantile distribution of the 5 worst-ranked companies on the market

These companies are hardly ever ranked above the 40% quantile of the investment universe of all the models. They only come from three activity sectors (Consumer Discretionary, Health Care and Materials three times) and two countries (the USA and China for four of them). Even though we are only analysing a very small sample of companies, the models seem to show negative biases towards certain countries and activity sectors.

The measure of the MAD can be useful for asset managers as it accounts for the strength of the consensus of the market on the average company scores. For a single company, this consensus strength can have different interpretations:

- If the consensus is strong (low MAD), market players tend to agree on the extra-financial score of a company and the risk is limited for all raters close to the consensus score,
- If the consensus is weak (high MAD), it means market players disagree on the score of the company, which can have several explanations:
  - The activities, policies or performances of the company involve a risk regarding which investors need to position themselves, splitting the market view and thus the distribution of the scores,
  - The company sends contradictory signals to the market. If the issue is merely related to poor investors relations rather than ESG practices and results, there is an investment opportunity for managers adopting a value investment strategy, as the company's scores are likely to improve in the future. If the signals are the result of poor management and policies, we can expect the opposite.

2. Quantile ranking count

This paragraph aims at analysing more thoroughly the hypothesis made after the previous quantile analysis. Even though their level of agreement or disagreement does not seem to follow a clear pattern, the models may be biased (at least negatively) when it comes to rankings.

One of the issues with the previous approach (only looking at the top or bottom companies) is that the results of our analysis may depend on the number of companies we choose to visualise (here, only the 5 best- or worst-ranked). To overcome this limitation, we use a measure presented in a 2020 paper by Florian Berg, Julian Koelbel and Roberto Rigobon: the quantile ranking count (QRC).

The idea here is to count the companies ranked in the same quantile or less by all the models, and divide this number by the total number of companies in the investment universe:

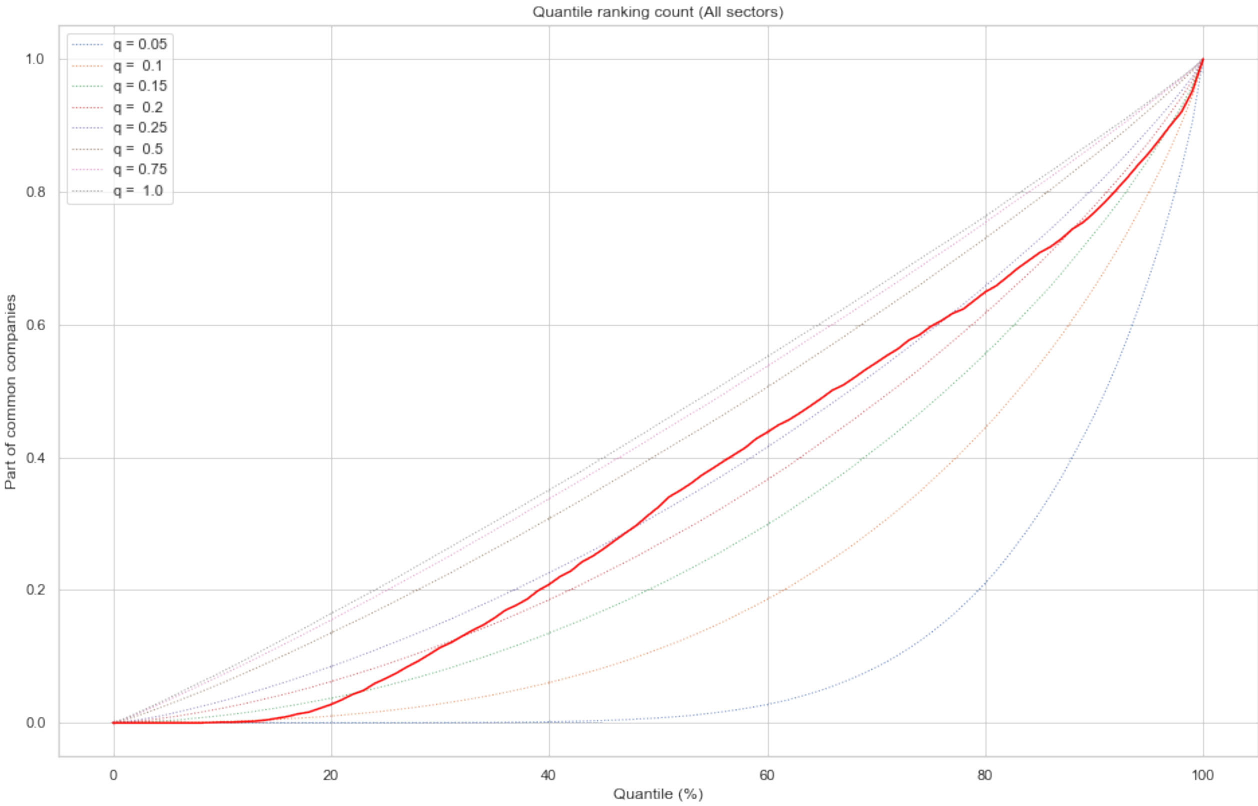
$$QRC_q = \frac{\text{Number of companies in the lowest q quantile of all the models}}{\text{Total number of companies}}$$

If the models were perfectly correlated, we would find the same companies in the lowest q% of every ranking, given the investment universe is the same for all the models. We would then get a QRC of q% for every quantile. The models being imperfectly correlated, we will find a lower number of companies that we can study to detect biases in the market assessment of some companies.

To have something to analyse and compare this lower number to, we will plot the QRC for several known distributions. As our dataset is composed of 2903 companies graded by 12 models, we generate 2903 artificial rankings 12 times in such a way that we control the correlation between the different artificial datasets thus simulated. These will be used as references for the QRC of the actual models. In order to be statistically relevant, we compute this artificial distribution 100 times for different correlation values: 5%, 10%, 15%, 20%, 25%, 50%, 75% and 100%.

We choose to implement a tolerance of 5% on the quantiles, meaning we consider that the models reach a consensus on the quantile of a company if at least 95% of them ranked the company in this quantile or lower. We do not require a perfect consensus to protect our analysis against outliers that may mitigate relevant trends in the results.

The first plot takes into account all the companies in the investment universe. The light dotted lines are the results of the simulated distributions. Their correlations are displayed in the legend. The red curve is the quantile ranking count of the reference model:



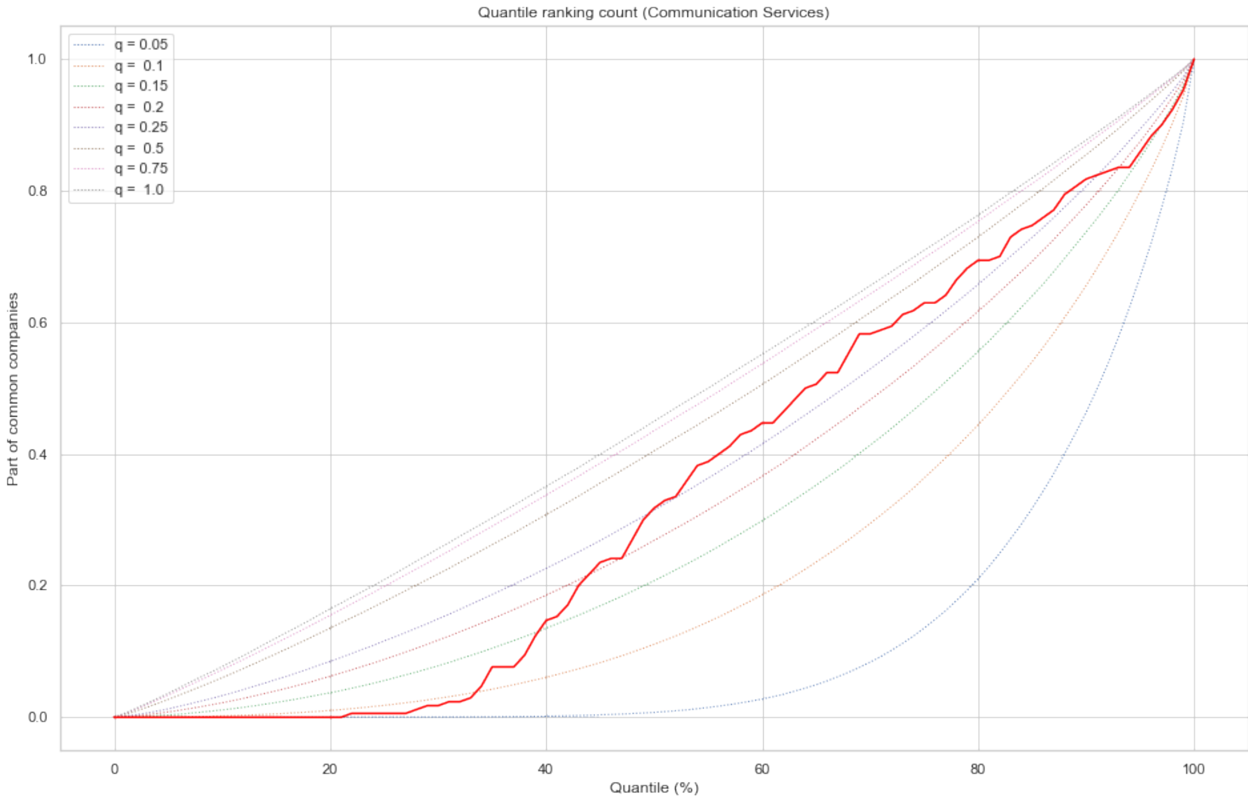
Plot 9: Quantile ranking count of the reference model on the full investment universe

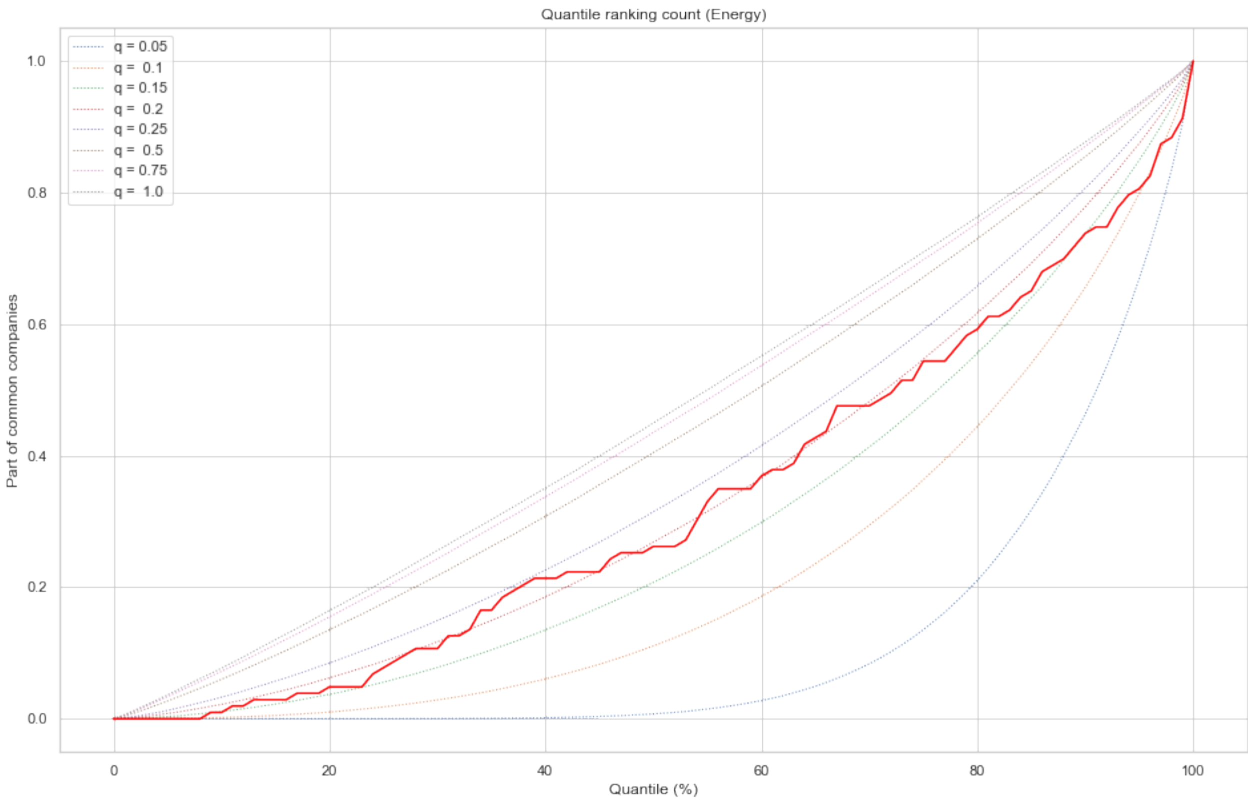
We can observe that the red line begins between the second and third dotted lines, which means the correlation of the different models before the 20% quantile is between 10% and 15%. It means that the QRC of the models available is similar to that of a set of models whose correlation is 10% to 15%. The red line then crosses several dotted lines to reach more than a 25% correlation between the 50% and the 75% quantiles. The correlation then seems to decrease again to reach about 10% for the highest quantiles.

This plot shows that there is more disagreement overall for the top and bottom companies, and more agreement on the average ones. Even though the reference model is correlated on average at 86.7% with the rest of the market, the market itself is not that well-correlated. It reaches about 30% correlation at best between the 50% and the 60% quantiles.

This plot gives us an overview of how we can interpret the QRC analysis. To discover biases drivers in the market, we can plot this analysis for subsamples of the original dataset. The plots of the different activity sectors can be found in

Annexe 3. We report the plots of the Communication Services and the Energy sectors to analyse them:





Plot 10: Quantile ranking counts of the reference model on the Communication services (top) and Energy (bottom) sectors

The Communication Services curve is flat until about the 30% quantile whereas the Energy one takes off at the 10% quantile. It means that the market tends to find a consensus by ranking companies from the Energy sector low, beginning at the 10% quantile. On the contrary, very few Communication Services companies are ranked by all the models under the 30% quantile. We can interpret from these plots that the market has a negative bias toward Energy companies. These firms tend to be ranked lower than others because of their activity sector.

Similarly, Communication Services companies have a QRC similar to a score distribution with a 15% correlation in the highest quantiles. Energy companies are similar to a distribution with a 5% correlation in the same conditions. It means that the different models tend to disagree more on the top-rated companies of the Energy sector than on those of the Communication Services one. It is harder for an Energy company than for a Communication Services one to be consistently ranked in the top quantiles of most of the available models.

Using the graphs in Annexe 3, we can analyse with precision the biases of the market towards any activity sector and compare them. Like with the heat maps in paragraph I.B.2., we can run this analysis for any other segmentation of the original dataset: by country, company size, growth rate…

The different tools available to compare the reference model to the rest of the market made it necessary to study the general biases of the market. This analysis can be achieved with a QRC applied to relevant subsamples of the market scores. We will now explore the biases of the reference model independently from the rest of the market.

II. Endogenous biases

A. Input categories

In this paragraph, we explore the biases of the reference model regarding its input data, regardless of the rest of the market. The aim is to understand in-depth the origins of the biases and the synergies between them and to correct them to obtain a more balanced model.

Before analysing the reference model itself, we need to transform our input data. Many metrics available to us are indeed collinear, which can mitigate the influence of each one of them on the biases of the model. In order not to overlook any interesting effects, we create input categories in which we will map the available metrics.

We choose to adopt a top-down approach by setting the framework and the list of categories independently of the available metrics. We use the set of essential extra-financial ESG indicators proposed by the AFG.

As stated in paragraph I.A.2., all the numerical metrics are already normalised, which means they have all been reduced to a distribution with a mean of 0 and a standard deviation of 1. As a consequence, metrics that were not on the same scale before are now, such as the number of controversies and the fines paid after controversies.

We then take all the available metrics and label them with a category from our framework. We make sure all the normalised scores are positively correlated with a good ESG performance so that different metrics do not cancel each other out. Missing metrics are replaced with average values. Then, we average all the metrics with the same label to obtain the category score of each company. After adding the remaining categorical features and removing the categories for which no metric is available, we obtain a dataset of 23 normalised input categories. The list of categories is available in Annexe 4.

The advantage of this new input data is that the interpretation of the results of the next paragraphs will be much easier to read, as all the influence of collinear metrics is supposed to be gathered in the same category. Moreover, we use a standardised framework to explain the results of the reference model, thus decreasing the effects of the biases coming from the choice of the metrics. Indeed, we want to focus on the structural biases of the score computation method (which depends on the model owner) independently of the choice of the metrics (which rather depends on the data provider).

B. Weights estimation

1. Models training

We now want to find a regressor to explain the scores of the reference model from the input data. A paper from Aaron K. Chatterji, Rodolphe Durand, David I. Levine, and Samuel Touboul (2015) showed that major data providers are on average correlated at 54%. Since we do not know the origin of the data used by the reference model, we can expect to explain at most 54% of the variations of the reference model with our regressor.

We use independent training and testing sets to fit and evaluate our models. As normalised scores are computed using both raw input data and the scores of all the other companies attributed by the reference model, they are harder to predict using only our category scores. As a consequence, we train our models to predict the original ESG score for better precision.

We test several methods, both linear and non-linear, to predict the final ESG scores of the reference model. More complex methods are likely to get better results, but they are also harder to interpret once trained. To strike a balance between accuracy and interpretability, we train increasingly complex models: linear, decision trees, support vector machines, ensemble methods and neural networks.

We assess the performance of the models using several metrics: the R-squared, the mean squared error and the absolute deviation. The R-squared is a measure of how much the variations of the target score are explained by the model, given the input data. It ranges from 0% to 100%, with an R-squared of 100% meaning that all the variations of the target score can be explained by the model:

$$R^2 = 1 - \frac{\text{Unexplained variations}}{\text{Total variations}}$$

The absolute deviation is the average absolute difference between the predicted score and the actual score, and the mean square error is the average square of this difference. This measure is especially useful as it puts more weight on the scores that are the most poorly predicted by the regressor. These scores are likely to modify the results of our analysis, so we want to penalise the regressors that create outliers. The results of the different regressors are summed up in this table, from the least to the most efficient:

REGRESSOR	AJUSTED R-SQUARED	MEAN SQUARED ERROR	ABSOLUTE DEVIATION
Linear support vector	21.2 %	99.806	8.07
Nu support vector	23.2 %	97.287	8.26
Stochastic gradient descent	24.4 %	95.773	8.041
Stacking regressor	27.3 %	92.13	7.652
Histogram gradient boosting	36.6 %	80.366	7.324
Random forest	40.2 %	75.836	6.672
Gradient boosting	42.7 %	72.621	6.954
Bagging regressor	43.4 %	71.775	7.031
Adaboost	44.4 %	70.44	6.69

Table 2: Overview of the regressors' performances on the test dataset

As we expected to have an R-squared below 54% (Berg et al., 2019), reaching 44.4% is satisfactory. The Adaboost regressor is a regressor that fits weaker regressors on the input data and sequentially learns how to boost their performances. After a prediction has been computed by an ensemble of weak regressors, each of them votes for the final combination. The Adaboost regressor then modifies the data so that the samples that have been incorrectly predicted have more weight in the next vote. As a consequence, the next regressors can focus on the hardest samples during the next training iteration to improve their overall performances.

The parameters used for the best performing regressor have been chosen through a Grid Search CV analysis to try and test many hyperparameter configurations, and the chosen ones are as follows:

- Base estimator (the weak regressor): random forest regressor using the mean absolute error as error function,
- Learning rate (the contribution of each base estimator to the final combination): 0.1,
- Number of estimators (the number of estimators voting for the final combination): 100,
- A linear loss to compare the predicted scores with the actual ones.

The issue we now have to face is the lack of explainability of this regressor. As it is composed of 100 weaker regressors voting for the results of weighted samples, there is no straightforward way to interpret the weights of the regressor as with a linear model.

2. Shapley values

We choose to assess the influence of each category using Shapley values. This method developed by the Economics Nobel prize laureate Lloyd S. Shapley is derived from game theory. It aims at quantifying the participation of each input data in the final prediction of a black-box model. We will use these Shapley values to analyse the biases of the reference model in the next paragraphs.

To explain how the Shapley values work, we first define a coalition game G(P, f). It is composed of a set of players P and a characteristic function f. The function f assigns a score to each subset of player S in the form of a real number. The Shapley value distributes the participation to the final score to each player. In our case, the players are the different instances of the categories used as input by the Adaboost regressor. The amount assigned to a player i belonging to the set P is given by the formula:

$$Shapley_i = \sum_{S \subseteq P \setminus \{i\}} \frac{|S|!(|P| - |S| - 1)!}{|P|!} (f(S \cup \{i\}) - f(S))$$

In this equation, S represents successively all the subsets of P that do not include the player i. |S|!(|P| - |S| - 1)! is the number of permutations of players in S and outside S (excluding i), for which the marginal contribution of player i to the final score is the same. It is divided by |P| to take the weighted average contribution over all the coalitions that can be formed without player i. Finally, f(S U {i}) - f(S) represents the marginal contribution of player i in coalition S.

The application of Shapley values to machine learning models explanation is called Shapley additive explanation (SHAP), as developed by Scott M. Lundberg, Su-In Lee (2017). We compute Shapley values to build an explainer function. This function is an approximation of the original model that is linear in binary vectors based on the input data. The coefficients are the Shapley values associated with these input data. We denote by g this explainer function, ϕ_i the Shapley value associated with the i-th category and ϕ_0 a base value for the explainer model. Given an input vector v, the output of the explainer model is:

$$g(v) = \phi_{v,0} + \sum_{i=1}^{n} \phi_{v,i} v_i$$

In this equation, v_i either equals 1 if the feature i is present and 0 otherwise. If we want to explain the score associated with an input vector w by the reference model, we just need to compute the corresponding binary vector v and compute the value of g(v) using the previous formula and the Shapley values associated with w by the Adaboost regressor. The Shapley values can be interpreted as the weights associated with each category in the decision of the model. The difference between the prediction of the regressor and the actual score will make it possible for us to assess the relevance of our conclusions. Indeed, our regressor does not perfectly fit the reference model and we can only interpret the results of our regressor, so we should be careful when interpreting the results. This process is summed up in the next chart:

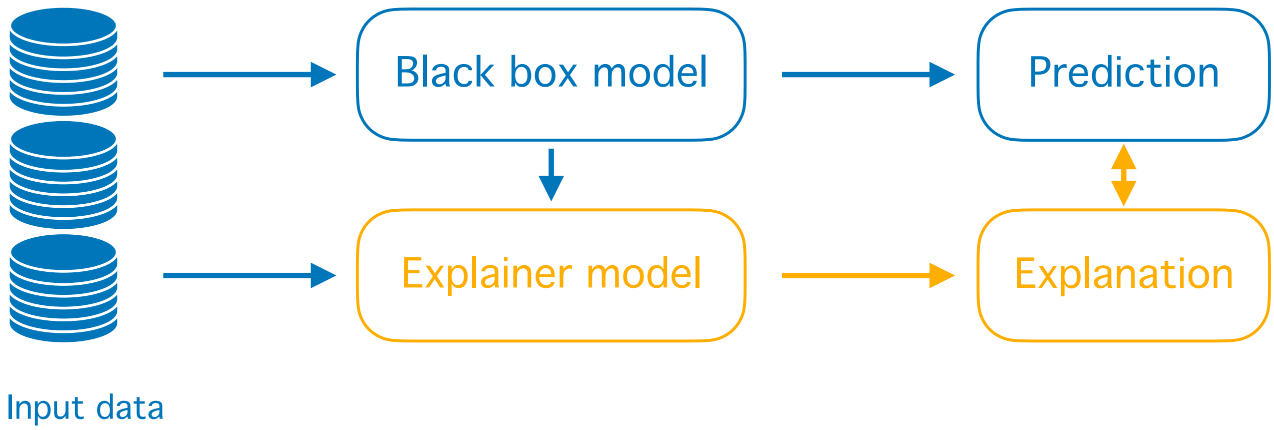


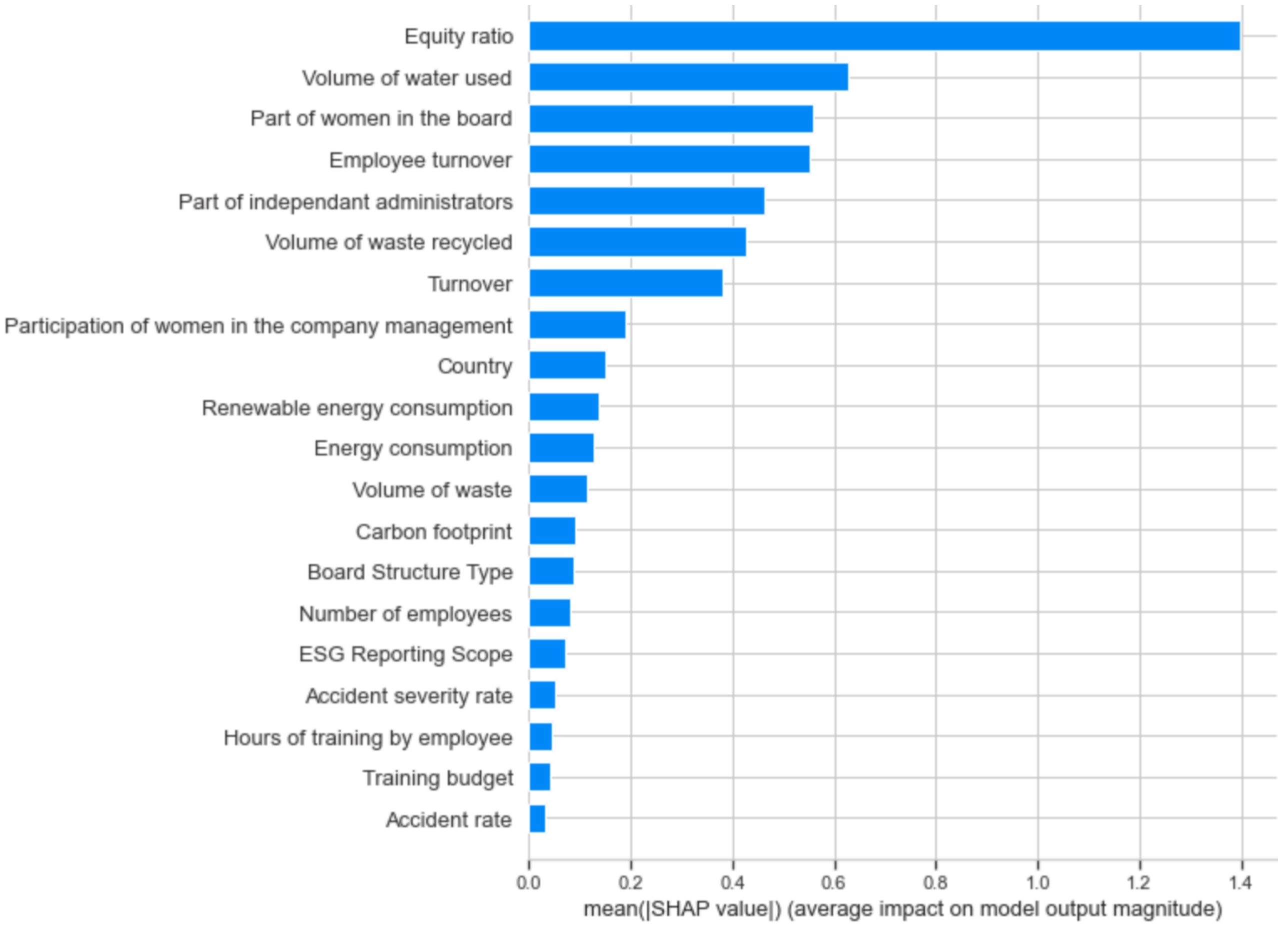
Chart 1: Construction of the explainer model

We use this explainer model to analyse our Adaboost regressor and discuss the results in the next paragraph.

C. Categorical biases analysis

1. Category weights

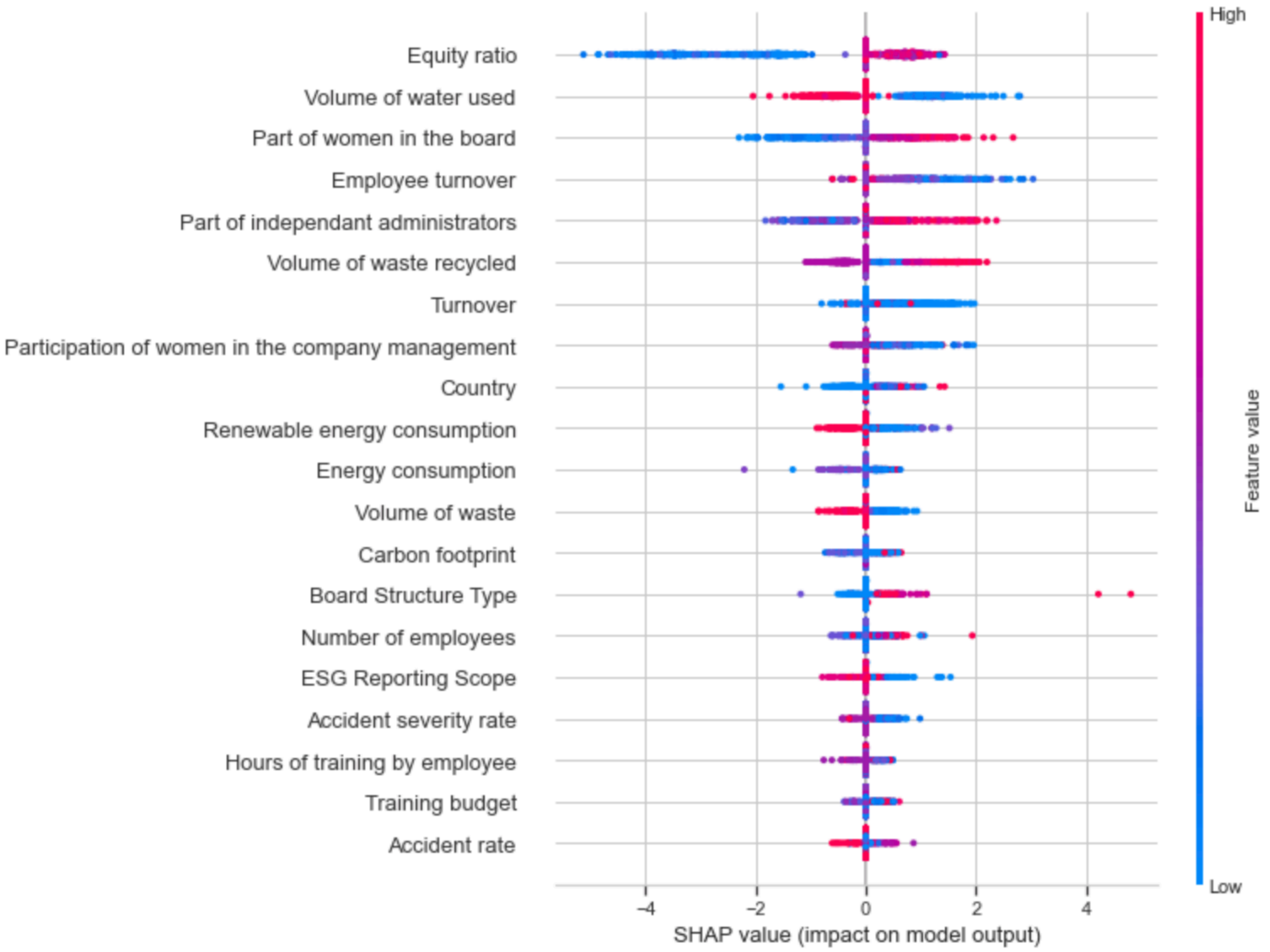
Even though SHAP values are computed for each instance of each category, we can assess their relative importance in the model computations by looking at the average absolute value of the SHAP value associated with each input category. We plotted the 20 most important categories:



Plot 11: Average absolute SHAP values associated with the 20 most important input categories

The equity ratio, which accounts for the salary discrepancies between the upper management and the average employee in the company, is by far the most important category. The part of women on the board and the employee turnover also have a major influence on the regressor's output. We can guess from these observations that the social pillar is overweighted in the reference model compared to the environmental or the governance ones. Most of the regressor's output can be explained with 7 input categories, from the equity ratio to the company turnover. The turnover is probably not taken into account as such by the reference model, but it can still be biased regarding this feature (which can account for company size). We will investigate the importance of exogenous biases further in the third part of this document.

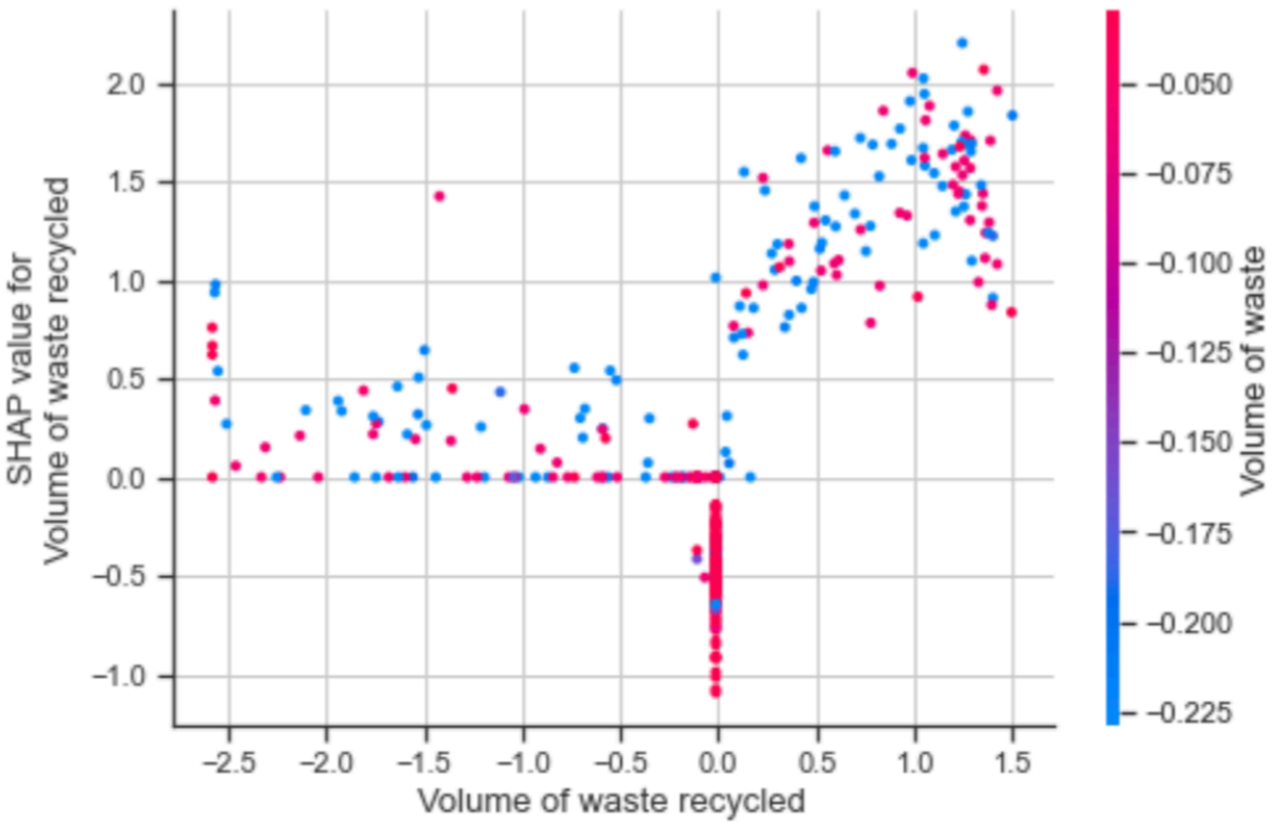
To have a more nuanced vision of the SHAP values of the different categories, we plot the repartition of positive and negative SHAP values for each category. Each point is the SHAP value of a company in the investment universe, and the colour of the point represents the value of the corresponding category:



Plot 12: Repartition of the SHAP values associated with the 20 most important input categories

We observe that companies with a low equity ratio category score are heavily penalised while those with a high score are moderately advantaged. Since the original input data has been preprocessed so that each category score is positively correlated with a performance leading to a better ESG score, a high category score here is associated with a low equity ratio. Indeed, a low equity ratio demonstrates more equality in the company salaries. Other criteria such as the renewable energy consumption cannot be associated with a good or a bad ESG performance without context, which is why we can see higher values associated with negative SHAP values. A more relevant metric would be the ratio of renewable energy consumption over total energy consumption, whose synergies could be interpreted thanks to the analysis presented in the next paragraph.

Some categories are more complex to interpret: high category scores in the volume of waste recycled are associated with high positive SHAP values, low values are associated with low positive SHAP values, and average values are associated with negative SHAP values. To understand better this output of the explainer model, we zoom in on the volume of waste recycled category. We plot the same SHAP values vertically this time and spread the datapoint horizontally in function of the underlying category score. The colour of the points represents the value of the most closely correlated category, the volume of waste:



Plot 13: Zoom on the SHAP values repartition of the Volume of waste recycled category

We notice three main groups of points. The first group is composed of companies with an average category score of 0. Since we fill empty data points with average values, these companies are likely not to have disclosed any metric for the volume of waste recycled category. The colour indicates that the companies in this group are mostly those which generate a large volume of waste. The resulting SHAP value is negative, which seems appropriate for companies that generate a lot of waste but lack transparency in their recycling processes.

The second group is composed of companies with a negative volume of waste recycled category score. It means they disclose the metrics related to this category but fall below the average. Their corresponding SHAP values are slightly positive, mostly lower than 0.5. It means the model rewards moderately the companies that are transparent about their recycling processes, but still have poor performances.

The third group in the top right corner is composed of the companies that disclose good results in the waste recycling category. They are rewarded with a high positive SHAP value, as they are both transparent and efficient in this category.

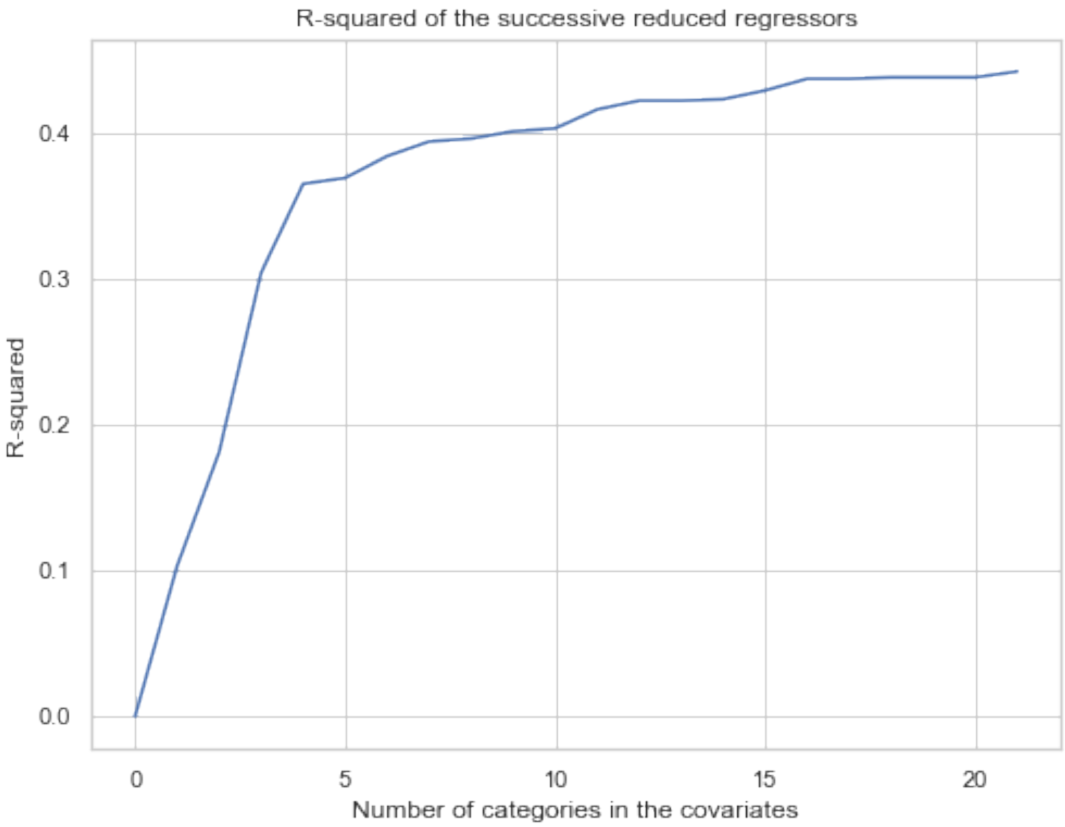
Using this particular example, we showed how the behaviour of the reference model can be explained in detail with the SHAP analysis. In addition to identifying the most influential categories, we can have a detailed view of the synergies between the different input features to explain the regressor's decisions. The same analysis can be run on subsamples of the total dataset, for instance on each activity sector or geographical area. The conclusions drawn on these various subsamples can help understand the biases of the reference model in specific situations.

One limitation of this method is that it relies on the categories we computed to train the regressor. The choice of these categories can introduce biases, as we aggregated all the raw metrics into these categories regardless of the number of metrics in each category or of the original input data of the reference model. This choice is necessary as we do not have access to this input data, but a different framework may lead to different conclusions regarding the reference model biases.

2. Reduced regression

We now want to quantify how much the Adaboost regressor relies on each category to make its predictions. This will help us assess to what extent the reference model is biased by the most influential categories. This bias is different than the ones we studied in the previous paragraph: we do not analyse the influence of each category score on the final ESG score, but the influence of category scores on each other. We do not want the performances in a category to influence the reference model's interpretation of another category. As many scores are computed or adjusted by analysts, they may introduce bias by over-or underrating companies on which they already have a view because of another category score.

For this analysis, we compute reduced regressors and evaluate their R-squared. The first regressor only uses the most influential category as a variable, the second adds the second most influential one, all the way until the 23rd and least influential variable. If all the categories are indeed taken into account independently in the final ESG score computation, then the explanatory power of the model should increase when new categories are added. If the category scores are increasingly collinear, the explanatory power of the successive models should almost stop increasing at some point. We plot the R-squared of the reduced regressor in function of the number of categories included in the explanatory variables:



Plot 14: R-squared in function of the number of explanatory variables

The steeper the curve, the more the regressor gains in explanatory power with the addition of new explanatory variables. We can see that the first 4 categories explain almost all of the predictive power of the model (36.5%) before the curve flattens. The last 15 categories hardly add any explanatory power at all. With all the variables included, the regressor reaches an R-squared of 44.4%, as stated in Table 2.

This high cross-category correlation can be explained by the way companies are analysed by most model owners. Analysts often tend to specialise in companies rather than in specific metrics and categories. As a consequence, the attributed scores influence each other and are all biased in the same way because they are estimated by the same person.

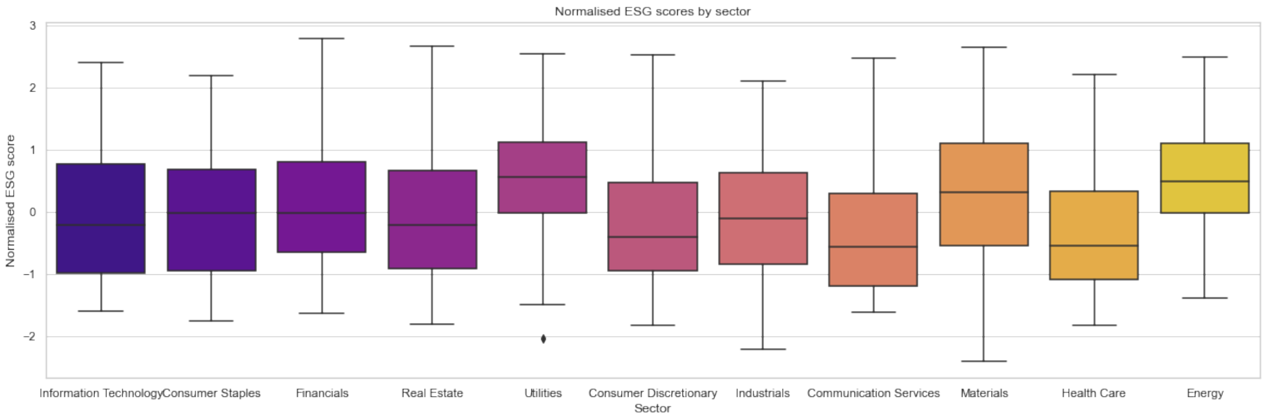
Moreover, we fill the empty metrics with average values. The metrics that are not disclosed by a company are likely to be correlated across categories. As a consequence, the disclosure scheme of a specific company probably strengthens the correlation between our explanatory variables, leaving only a few clusters of independent categories. With our reference model, only 4 clusters seem to explain most of the variations of the score, all the categories being strongly correlated to one of these clusters: Equity ratio, Volume of water used, Part of women on the board and Employee turnover.

III. Exogenous biases

A. Motivation

In addition to the biases coming from the input data, the reference model can have exogenous biases. It means it can be biased towards external features that are not directly present as input. The most common exogenous factors are the company size, sector, and geographical activity area. Amir Amel-Zadeh and George Serafeim (2018) showed that investors could benefit from a better standardisation of ESG scoring through a deeper understanding of the effect of these factors.

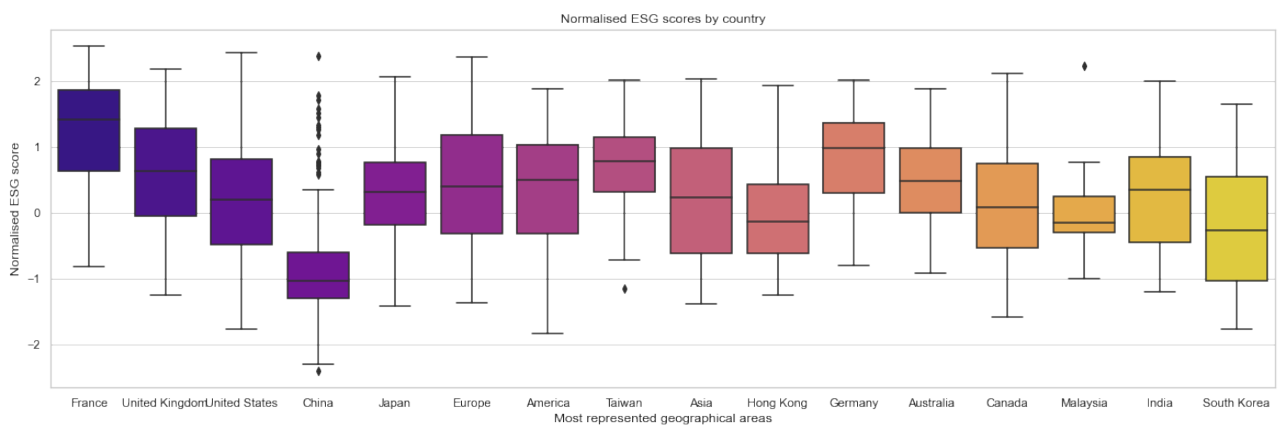
We first demonstrate that the reference model does have exogenous biases by box plotting the average ESG scores of different subsamples of the investment universe. In this representation, the horizontal line represents the mean of the distribution, the two boxes are the first and third quartiles and the vertical whiskers indicate the total range of normalised ESG scores in this activity sector. Any value outside this range (computed as a function of the interquartile range) is represented with an individual point and considered an outlier. Here is a box plot of the normalised ESG scores distribution by activity sector:



Plot 15: Normalised ESG scores by sector

We observe that the average values are different from one sector to the other. Some sectors have the same average as the investment universe taken as a whole (Consumer staples, Financials) while others have more than half of a standard deviation of difference with it (Utilities, Health care). The interquartile ranges also range from roughly one to two standard deviations, whereas the total range is rather consistent throughout the different sectors. Considering many model owners adopt a best-in-class strategy, their assessment of activity sectors must be balanced and as similar as possible. The reference model is not well-suited for this kind of investment strategy, as most companies in the Utility and Energy sector will be ranked higher than those in the Health Care sector.

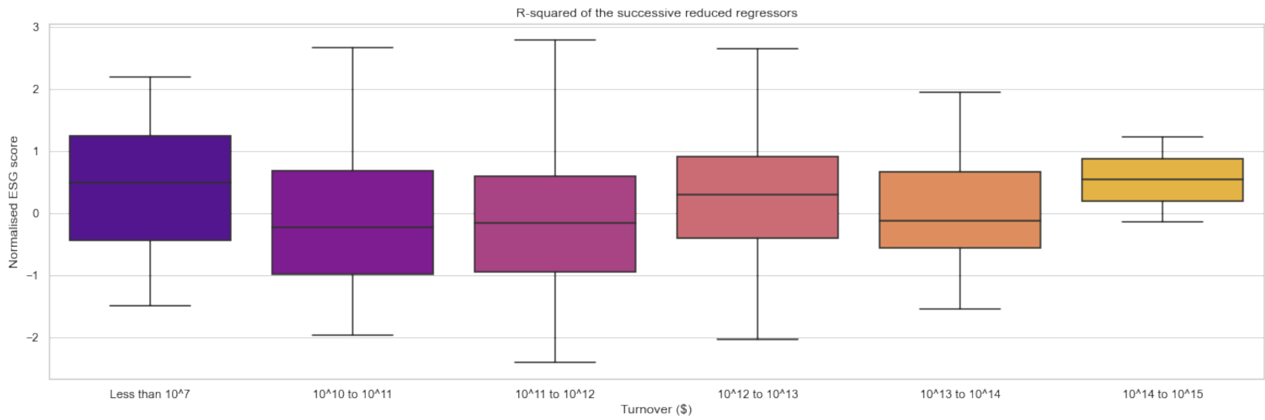
We now visualise the same plot using the most represented countries and geographical areas:



Plot 16: Normalised ESG scores by country

Similarly, we can observe wide differences in the distribution of the most represented geographical areas. These differences are also critical to take into account, as they may result from ESG dumping from companies in wealthier countries to companies in cheaper and less regulated ones. These differences may also come from the different parts of each activity sector in each country. If a model is biased regarding this variable (as it is with the reference model), it may add to the bias towards geographical areas. With a model unbiased towards activity sectors taking properly into account the whole value chain of the companies in the investment universe, the biases towards geographical areas should be minimal.

Finally, we plot the normalised ESG scores distribution by company size. We use the yearly turnover on a logarithmic scale as a proxy to create groups of companies of different sizes:



Plot 17: Normalised ESG scores by company size

As ESG models mostly rely on disclosed corporate data, the company size factor is often a source of biases in ESG scores. Wealthier companies tend to have more resources to dedicate to marketing, greenwashing and seemingly green side-projects. Even though some of them do have good ESG performances, the overall better scores of bigger companies can partially be linked to their greater disclosure capabilities. An efficient model should be able to sort the communication efforts from the actual effects of the company's CSR policies. The reference model does not seem to be too prone to giving significantly better scores to bigger companies, but on the opposite, the smallest companies in the investment universe stand out. We need to investigate why they are advantaged this way.

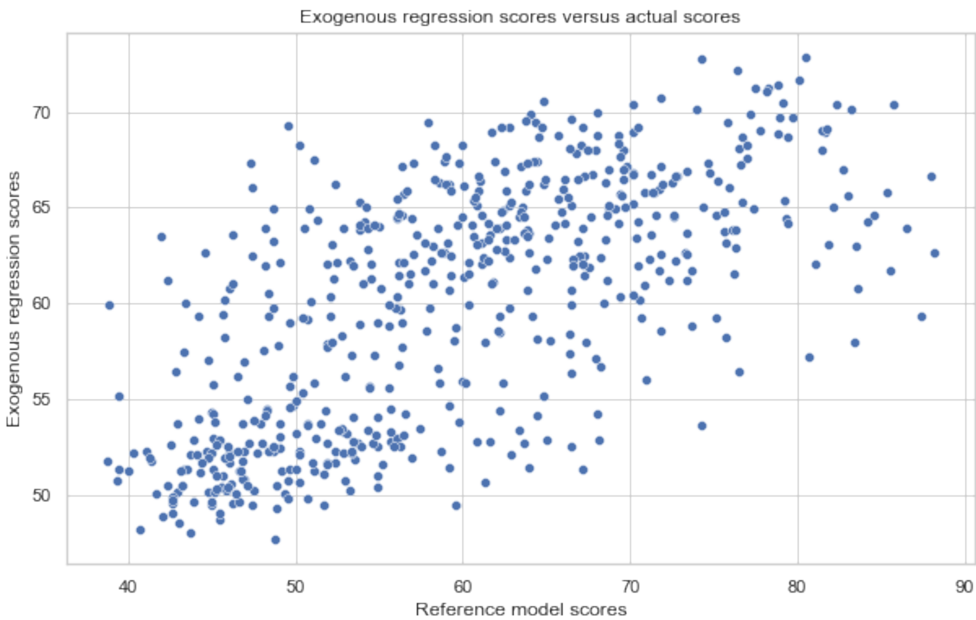
We showed that the reference model is biased towards 3 exogenous variables: activity sector, geographical area and size. Even though there must be many other factors towards which the model may be biased, we explained why an efficient model should be well-balanced regarding these three specifically. As a consequence, we will analyse in the next paragraphs how to mitigate and control these biases to improve the reference model's performance.

B. Biases offsetting

We attempt to mitigate the biases linked to exogenous factors by performing a regression using only these three factors as input variables. The aim is to isolate the contribution of these factors to remove them from the ESG score of the reference model, as suggested by Kevin Ratsimiveh et al. (2020).

However, a major hypothesis to extract the individual contribution of each exogenous factor is their mutual independence. As we explained in the previous paragraph, they are interconnected, which is likely to modify the results. As a consequence, we use as regressor a Partial Least Square (PLS) regressor. This regressor first aggregates the input variables into independent variables before trying to fit multiple linear regressors on this new dataset.

We reach an R-squared of 18.3% and an MSE of 104.226. Here is the plot of the predicted scores against the actual ones:

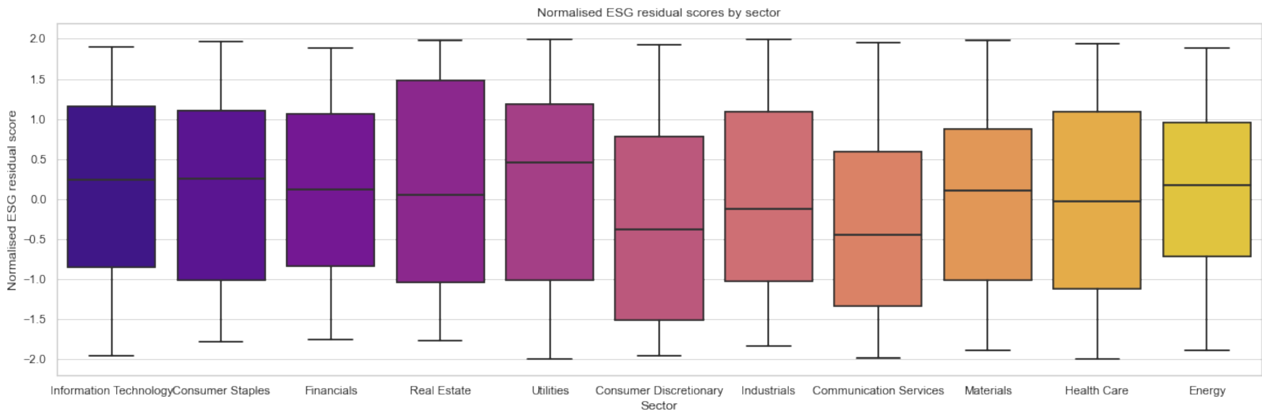


Plot 18: Exogenous regression scores versus actual scores

The PLS regressor is not as accurate as the Adaboost one, but it only relies on three variables to explain the variations of the reference model's ESG score. To extract the biases linked to these factors from the ESG scores of the reference model, we study the differences between the predictions of the PLS regressor and the target scores, also called residuals. The residuals represent the part of the score that cannot be explained by the input values. All the information contained in the residuals can only be explained by factors that are independent of them. It means all the biases linked to the exogenous factors should be removed from the residual scores, only keeping the unbiased part of the reference model's ESG score. We, therefore, normalise the vector of residual scores we just computed to compare it with the actual normalised scores. We denote by R the vector of residuals and compute a normalised residual score with the formula:

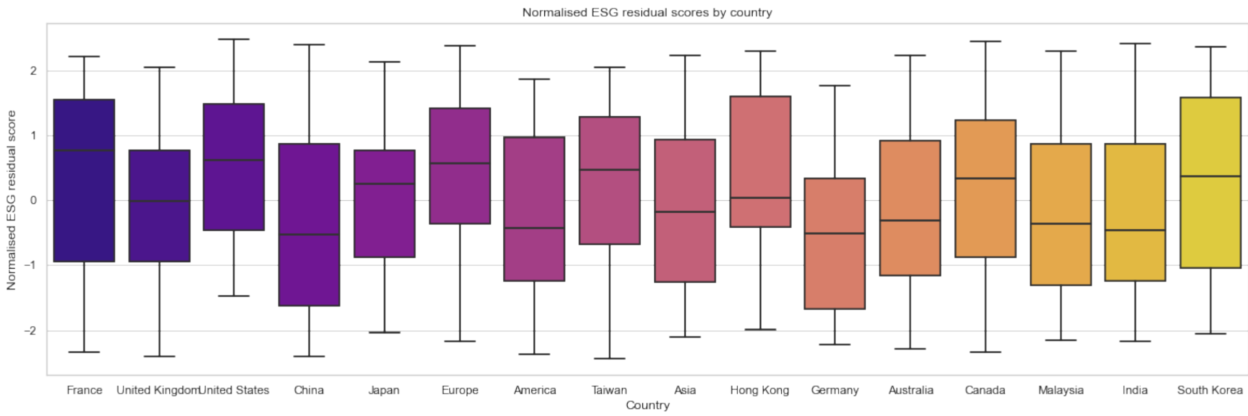
$$\text{Residual score}_i = \frac{R_i - mean(R)}{std(R)}$$

This new score represents the ESG performances of the companies independently of the three exogenous factors. To assess the improvement of these scores compared to the normalised ESG scores of the reference model, we compute the box plots by sector:

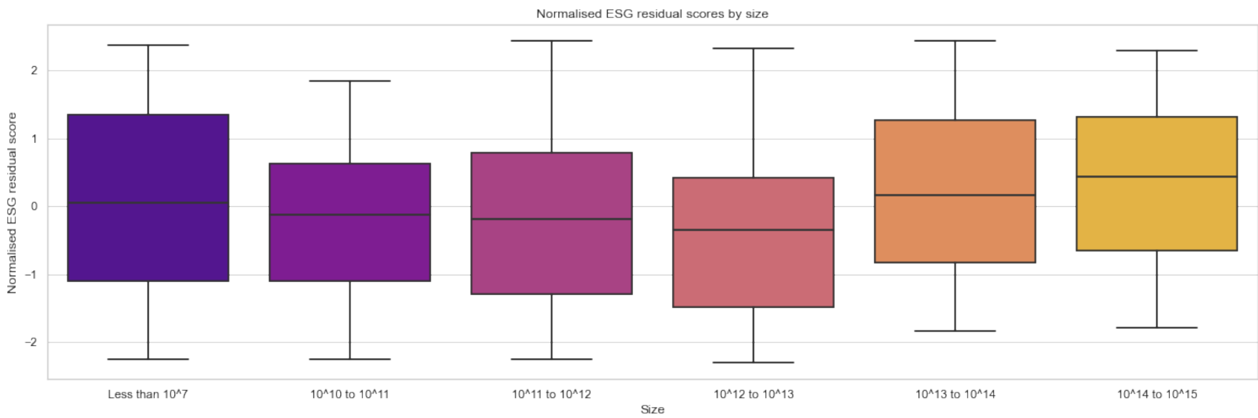


Plot 19: Residual scores by activity sector

Even though the different sectors are still not perfectly equivalent, we notice an improvement in the repartition of ESG scores. We observe similar results in the country and company size segmentation:



Plot 20: Residual scores by geographical area



Plot 21: Residual scores by company size

The biases are attenuated though still present, which confirms the residual scores are a new version of the original ESG scores that is more independent of the three identified exogenous factors than the original ESG score. The remaining biases could be mitigated further with a better fit between the exogenous factors and the ESG scores of the reference model. As our dataset is based on poorly disclosed data, an even better independence level could be achieved by improving the completeness of our dataset. Only 34% of companies disclose all three of the exogenous factors, which leaves much room for improvement.

C. Residuals analysis

We computed in the previous paragraph two new scores:

- A regression of the reference model's ESG score on the three exogenous factors,
- A residual score that is independent of these factors.

The first score represents the part of the final scoring explained by the exogenous factors and the second explains the rest of the score independently from them. We visualise the relationships between them by plotting the residual scores against the exogenous factors-based scores:

Plot 22: Residual scores versus exogenous factors-based scores

Each point represents a company from the investment universe. The residual score is on the y-axis, so the higher the company is on the plot the higher its ESG score (as explained independently from the exogenous factors). The exogenous factors-based score is on the x-axis, so the more on the right the company the most advantaged by its exogenous factors. From the plot, we can distinguish four main groups of companies:

- In the top right corner: these companies have activity sectors, geographical areas and sizes that favour high ESG scores, and do have good intrinsic ESG performances as evaluated by the reference model,
- In the bottom left corner: these companies have activity sectors, geographical areas and sizes that favour low ESG scores, and do have poor intrinsic ESG performances as evaluated by the reference model,
- In the top left corner: these companies have activity sectors, geographical areas and sizes that favour low ESG scores, but still have good intrinsic ESG performances as evaluated by the reference model,

- In the bottom right corner: these companies have activity sectors, geographical areas and sizes that favour high ESG scores, but still have poor intrinsic ESG performances as evaluated by the reference model.

We observe that most companies are either in the top right or in the bottom left corner, showing that they have an ESG score aligned with their exogenous factors. Most of the companies in the top left or bottom right corners remain rather close to the centre of the distribution, meaning only a few of them are perceived really differently from what their exogenous factors would suggest.

The proposed split of the reference model's ESG scores into residual and exogenous scores allows for a more advanced analysis of each company's performance. Lasse Heje Pedersen and Shaun Fitzgibbons (2020) suggested that this dual view would help ESG scores play their two main roles: assess a company's intrinsic performance and affect the investor's preferences.

While designing a bias-free ESG scoring model is impossible (due, for instance, to variable disclosure requirements between countries), this methodology shows what part of the ESG score can be imputed to the three exogenous factors defined in this document. Once the individual performances of the companies are isolated, best-in-class investors can make more relevant decisions on their investment universe while keeping in mind the highest-scoring sectors, sizes and geographical areas.

Conclusions

This document aims to explore different methods to detect and quantify the biases of ESG scoring models. The wide variety of ESG scores on the market is the consequence of both the low quality of ESG data and the variety of views among investors. It is legitimate to have different takes on ESG, but not to disagree on the very definition of indicators and measurement methods.

Once the measurement issues are addressed through standardisation and regimentation, the remaining divergence in extra-financial scoring is associated with the ESG strategy of each rater. They need to disentangle the differentiating choices they made regarding scope and weights for two main reasons:

1. Explain and compare their scoring model to their peers,

2. Improve the matching between their quantitative models and their qualitative responsible investment strategy.

Even though many investors take into account meta-factors such as disclosure rate or dynamic data such as controversies, we chose to focus on models that simply aggregate public metrics declared by companies into comprehensive ESG scores. A similar analysis could be carried out with models predicting scores on a more granular level, such as E, S or G individually.

We begin with the analysis of the relative biases of a model compared to another model or set of models. The methods described could be used to compare a specific model to the rest of the market. We show the necessity of normalising the scores to make relevant comparisons and demonstrate how quantiles can be used to benefit investors to run firm-level comparisons. As biases are based on factors that are not perfectly independent from each other, we present a methodology to detect the biases that are amplified by the synergies of two factors, be they endogenous or exogenous.

In the second part, we focus on the model itself and study its biases in absolute value and not relative to another model. We build a new set of input data to have more independent features using an AFG-developed framework. We use these features to train an Adaboost regressor and explain its results with a Shapley analysis. The resulting Shapley values help us assess the importance of each input feature in the final result on a firm-level basis and on average for the model. We understand what values or intervals of values drive the score up or down and can analyse the synergies between input features. Compared to the methodology in part 1, this one is more precise but only works with features used as input by the model.

Finally, we try to understand if exogenous factors can be an efficient proxy for model biases. We choose to focus on three (activity sector, company size and geographical area), as they are widely recognised as biases vectors and can significantly impact investment decisions. We train a model to create two independent scores. They represent the part of the ESG score explained by the exogenous factors alone and the part explained by the actual intrinsic performances of the company. After showing that we indeed achieve better independence from exogenous factors using this model, we propose a dual view of the original ESG scoring showing how the companies of the investment universe can outperform or not their peers, as defined by the exogenous factors.

The results presented in this document could be improved with a better input dataset, as only 7 of the 92 features used for the regressions had a disclosure rate of more than 50%. The others were replaced by average values, which undoubtedly mitigated the prediction and explanatory power of the regressors.

Our results can be useful to ESG analysts, as bias analysis is mandatory to check if the performances of an ESG scoring model match their investment strategy. The comparison to a group of peers is the most relevant here, as investors want to control their biases compared to the market. As the necessary data to run this analysis is not likely to be available to analysts, we also provide tools to evaluate the absolute biases of a given model.

For investors, we show how the exogenous biases necessarily developed with their ESG models can be used to enhance their firm-level analysis instead of merely biasing their investment decisions.

Bibliography

Aaron K. Chatterji, Rodolphe Durand, David I. Levine, Samuel Touboul, Do ratings of firms converge? Implications for managers, investors and strategy researchers (June 12, 2015), https://onlinelibrary.wiley.com/doi/10.1002/smj.2407

Alix Faure, Essential extra-financial indicators to assess a company (June 19, 2020), https://www.afg.asso.fr/wp-content/uploads/2020/06/guidepro-esgeng200618web.pdf

Amir Amel-Zadeh, George Serafeim, Why and How Investors Use ESG Information: Evidence from a Global Survey (December 12, 2018), https://www.tandfonline.com/doi/abs/10.2469/faj.v74.n3.2?journalCode=ufaj20

Berg, Florian and Kölbel, Julian and Rigobon, Roberto, Aggregate Confusion: The Divergence of ESG Ratings (May 17, 2020), https://ssrn.com/abstract=3438533 or http://dx.doi.org/10.2139/ssrn.3438533

Berg, Florian and Kölbel, Julian and Pavlova, Anna and Rigobon, Roberto, ESG Confusion and Stock Returns: Tackling the Problem of Noise (October 12, 2021), https://ssrn.com/abstract=3941514

Berk, Jonathan B. and van Binsbergen, Jules H., The Impact of Impact Investing (August 21, 2021), https://ssrn.com/abstract=3909166

Dimson, Elroy and Karakaş, Oğuzhan and Li, Xi, Active Ownership (August 7, 2015), https://ssrn.com/abstract=2154724

European Securities and Market Authority (ESMA), Sustainable Finance Roadmap 2022-2024, (February 10, 2022), https://www.esma.europa.eu/sites/default/files/library/esma30-379-1051_sustainable_finance_roadmap.pdf

Fama, Eugene F. and French, Kenneth R., Disagreement, Tastes, and Asset Prices (November 2005), https://ssrn.com/abstract=502605

Gueant, Olivier and Peladan, Jean-Guillaume and Robert-Dautun, Alain and Tankov, Peter, Environmental transition alignment and portfolio performance (June 29, 2021), https://ssrn.com/abstract=3876731

Jin, Ick, ESG-Screening and Factor-Risk-Adjusted Performance: The Concentration Level of Screening Does Matter (October 17, 2020), https://ssrn.com/abstract=3722485

Kalesnik, Vitali and Wilkens, Marco and Zink, Jonas, Green Data or Greenwashing? Do Corporate Carbon Emissions Data Enable Investors to Mitigate Climate Change? (November 24, 2020), https://ssrn.com/abstract=3722973

Kevin Ratsimiveh, Patrick Hubert, Valéry Lucas-Leclin, Emeric Nicolas, ESG scores and beyond (July 7, 2020), https://content.ftserussell.com/sites/default/files/esg_scores_and_beyond_part_1_final.pdf

Krueger, Philipp, Corporate Goodness and Shareholder Wealth (July 7, 2014), https://ssrn.com/abstract=2287089

Philipp Krueger, Zacharias Sautner, Dragon Yongjun Tang and RUI Zhong, The Effects of Mandatory ESG Disclosure Around the World (November 30, 2021), https://ssrn.com/abstract=3832745

Lasse Heje Pedersen, Shaun Fitzgibbons, Lukasz Pomorski, Responsible investing: The ESG-efficient frontier (December 8, 2020), https://www.sciencedirect.com/science/article/pii/S0304405X20302853

Moinak Maiti (2021) Is ESG the succeeding risk factor?, Journal of Sustainable Finance & Investment, 11:3, 199-213, DOI: 10.1080/20430795.2020.1723380

N. C. Ashwin Kumar, Camille Smith, Leïla Badis, Nan Wang, Paz Ambrosy & Rodrigo Tavares (2016) ESG factors and risk-adjusted-performance: a new quantitative model, Journal of Sustainable Finance & Investment, 6:4, 292-300, DOI: 10.1080/20430795.2016.1234909

Pastor, Lubos and Stambaugh, Robert F. and Taylor, Lucian A., Sustainable Investing in Equilibrium (February 14, 2020), https://ssrn.com/abstract=3559432

Rajna Gibson Brandon, Philipp Krueger & Peter Steffen Schmidt (2021) ESG Rating Disagreement and Stock Returns, Financial Analysts Journal, 77:4, 104-127, DOI:10.1080/0015198X.2021.1963186

Scott M. Lundberg, Su-In Lee, A Unified Approach to Interpreting Model Predictions (November 25, 2017), https://papers.nips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

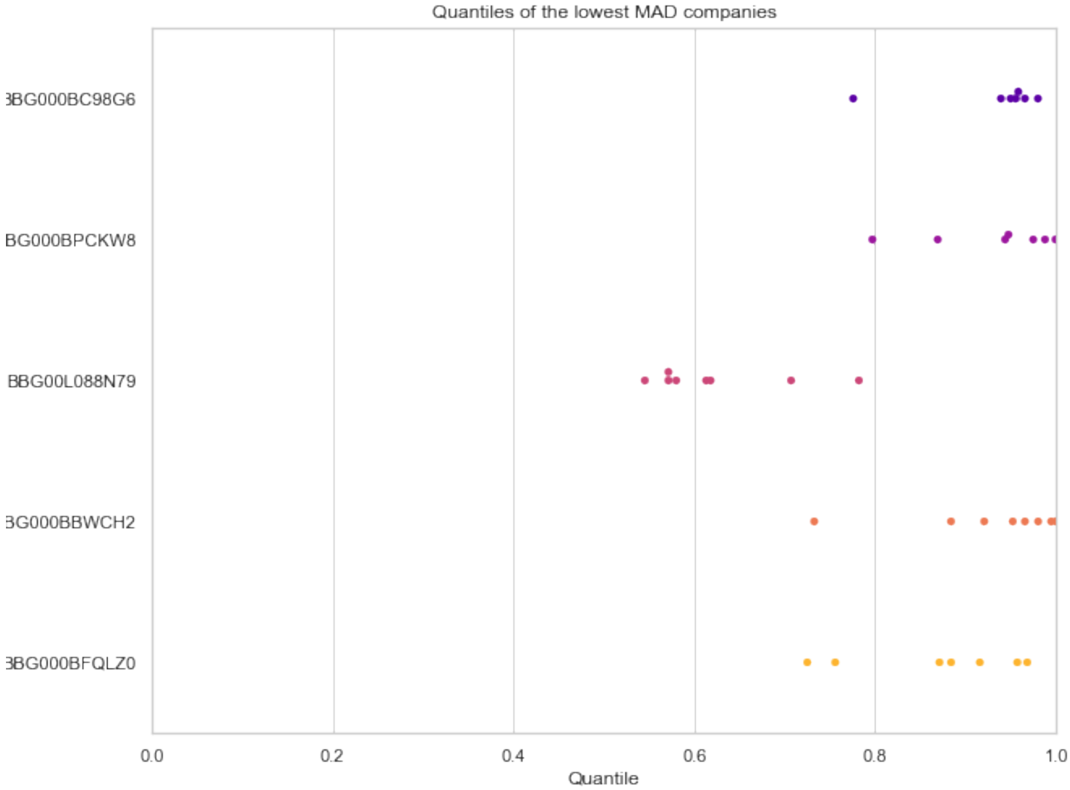
Annexes

ANNEXE 1: LIST OF BASIC FEATURES

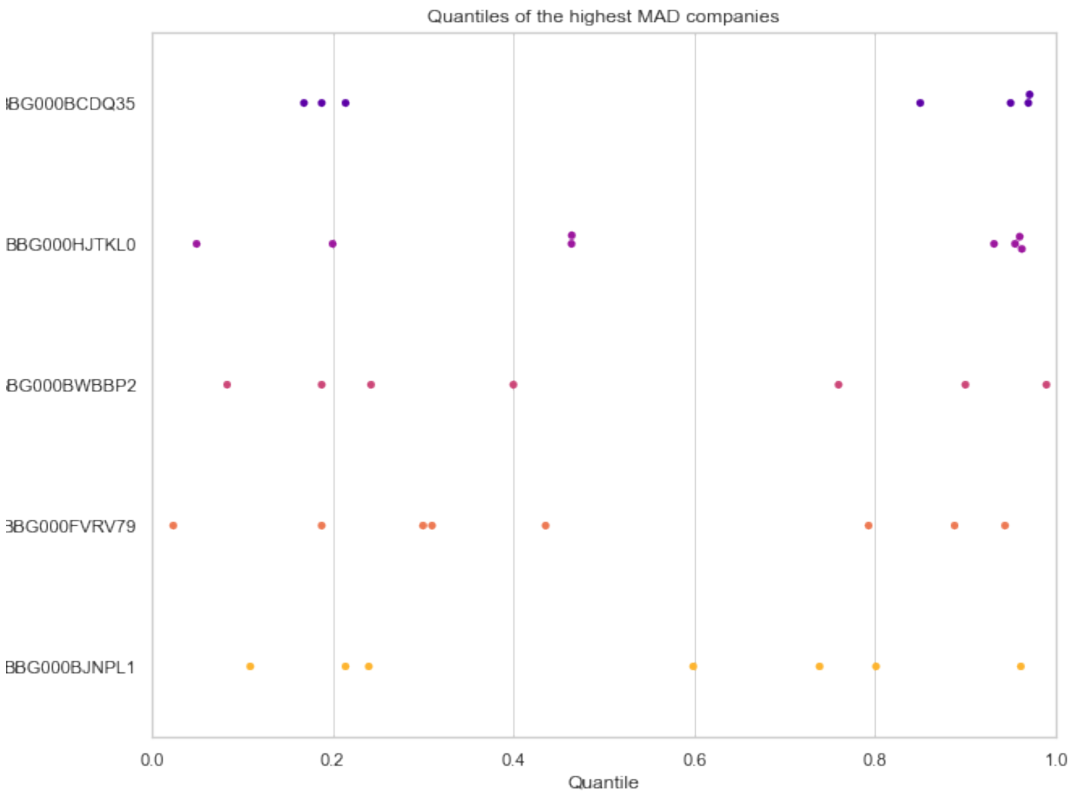
1	Accidents Total
2	Advance Notice Period Days
3	Audit Committee Expertise
4	Audit Committee Independence
5	Audit Committee Mgt Independence
6	Board Member Affiliations
7	Board Attendance
8	Board Background and Skills
9	Board Gender Diversity, Percent
10	Board Member Compensation
11	Board Individual Re-election
12	Board Size More Ten Less Eight
13	Board Specific Skills, Percent
14	CEO-Chairman Separation
15	Total CO2 Equivalent Emissions To Revenues USD in million
16	Compensation Committee Independence
17	Compensation Committee Mgt Independence
18	Net Employment Creation
19	Total Energy Use To Revenues USD in million
20	Estimated CO2 Equivalents Emission Total
21	Executive Members Gender Diversity, Percent
22	Average Board Tenure
23	Total Hazardous Waste To Revenues USD in million
24	Highest Remuneration Package
25	Independent Board Members
26	Injuries To Million Hours
27	Announced Layoffs To Total Employees
28	Nomination Committee Involvement
29	Nomination Committee Independence
30	Nomination Committee Mgt Independence
31	Non-audit to Audit Fees Ratio
32	Non-Executive Board Members
33	Salary Gap
34	Strictly Independent Board Members
35	Total Donations To Revenues in million
36	Total Waste To Revenues USD in million
37	Waste Recycled To Total Waste
38	Water Use To Revenues USD in million
39	Audit Committee NonExecutive Members
40	Auditor Tenure
41	Average Training Hours
42	Board Meeting Attendance Average
43	Number of Board Meetings
44	Board Member LT Compensation Incentives
45	Board Member Membership Limits
46	Board Size

47	Board Structure Type
48	Board Member Term Duration
49	CEO Board Member
50	Chairman is ex-CEO
51	Classified Board Structure
52	CO2 Equivalent Emissions Direct, Scope 1
53	CO2 Equivalent Emissions Total
54	CO2 Estimation Method
55	CO2 Equivalent Emissions Indirect, Scope 2
56	CO2 Equivalent Emissions Indirect, Scope 3
57	Committee Meetings Attendance Average
58	Compensation Committee NonExecutive Members
59	ESG Reporting Scope
60	Different Voting Right Share
61	Donations Total
62	Earnings Restatement
63	Electricity Purchased
64	Emission Reduction Target Percentage
65	Emission Reduction Target Year
66	Employee Accidents
67	Employee Fatalities
68	Employee Resource Groups
69	Number of Employees from CSR reporting
70	Energy Purchased Direct
71	Energy Use Total
72	Fresh Water Withdrawal Total
73	Hazardous Waste
74	Nomination Committee NonExecutive Members
75	Non-Hazardous Waste
76	Product Recall
77	Profit Warnings
78	Total Senior Executives Compensation
79	Staggered Board Structure
80	Total Injury Rate Employees
81	Total Injury Rate Total
82	Total Renewable Energy
83	Trade Union Representation
84	Training Hours Total
85	Turnover of Employees
86	Voting Cap Percentage
87	Waste Recycled Total
88	Waste Recycling Ratio
89	Waste Total
90	Water Withdrawal Total
91	Women Employees
92	Women Managers

ANNEXE 2: HIGH AND LOW DISAGREEMENT COMPANIES

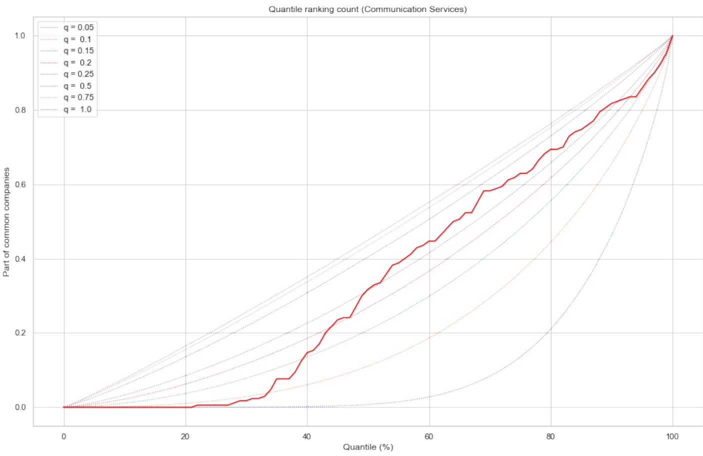


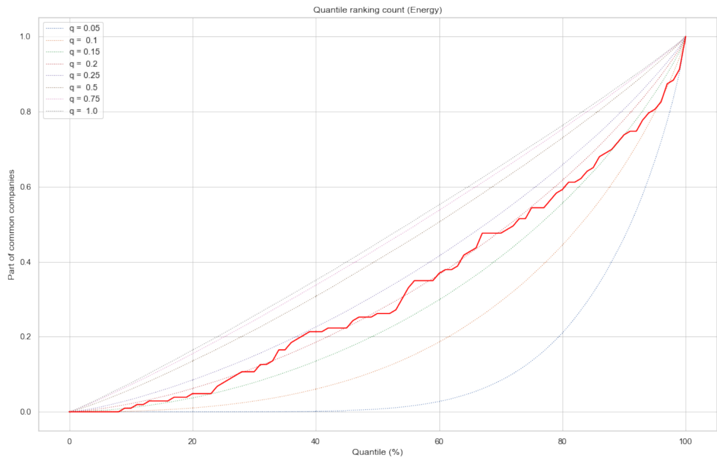
Normalised scores of companies with the lowest MAD

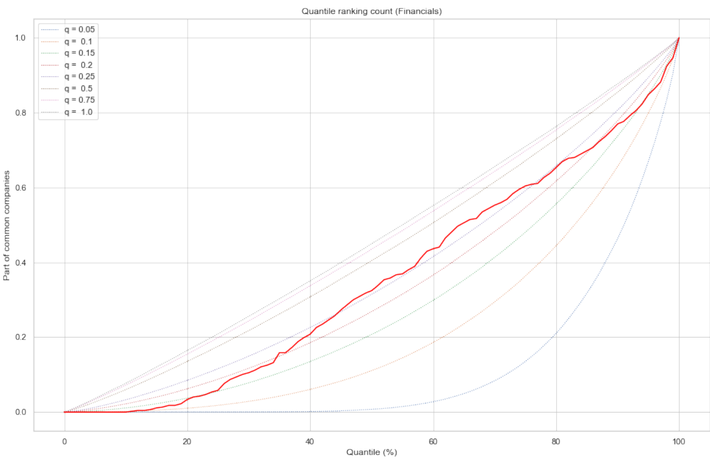


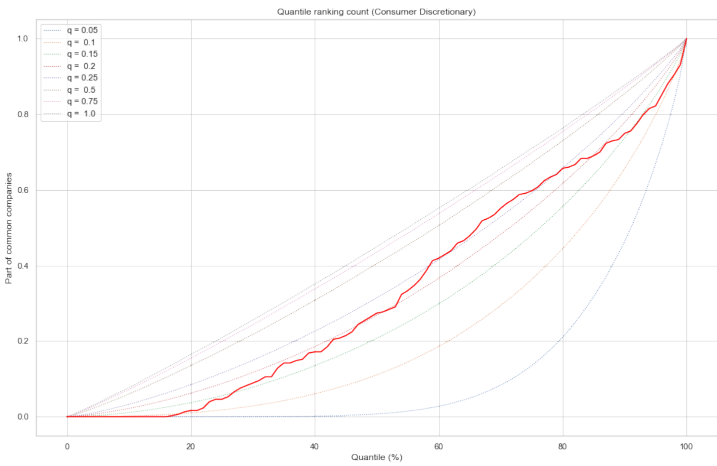
Normalised scores of companies with the highest MAD

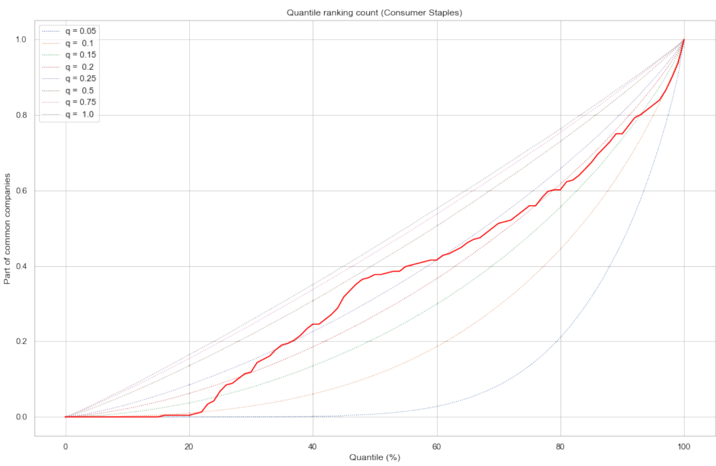
ANNEX 3: QUANTILE RANKING COUNT BY ACTIVITY SECTOR

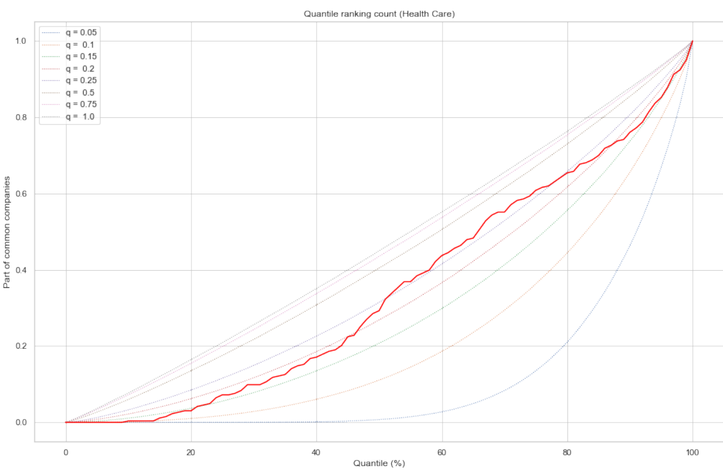


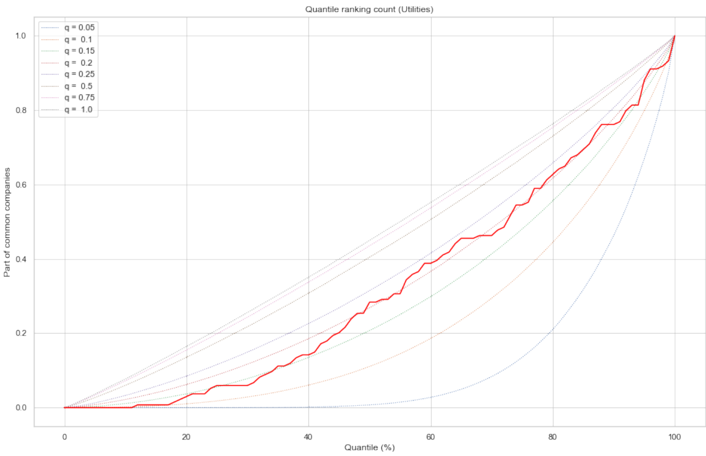


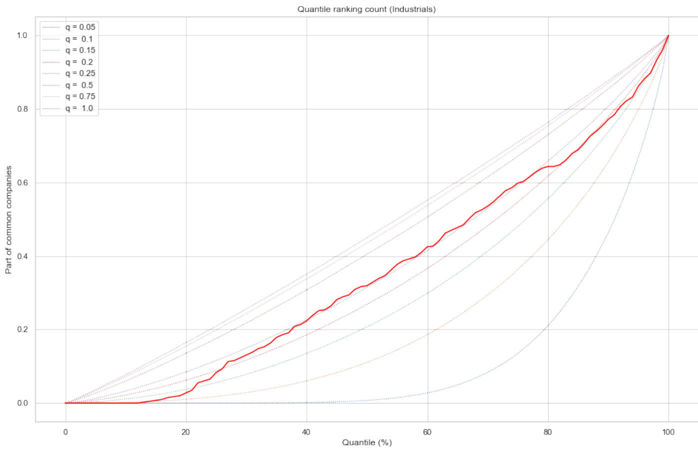


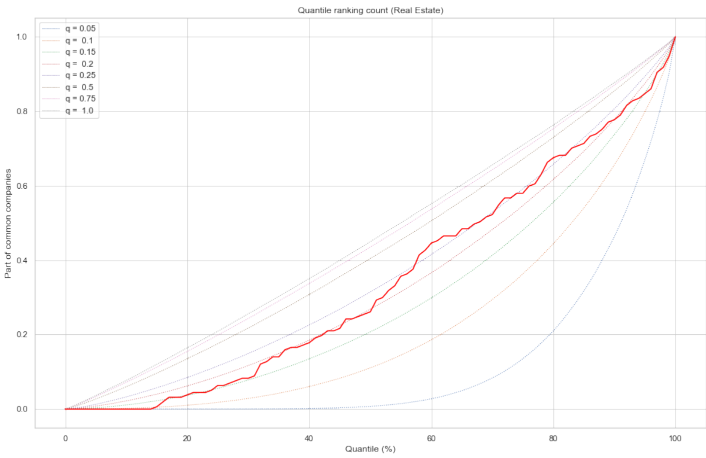


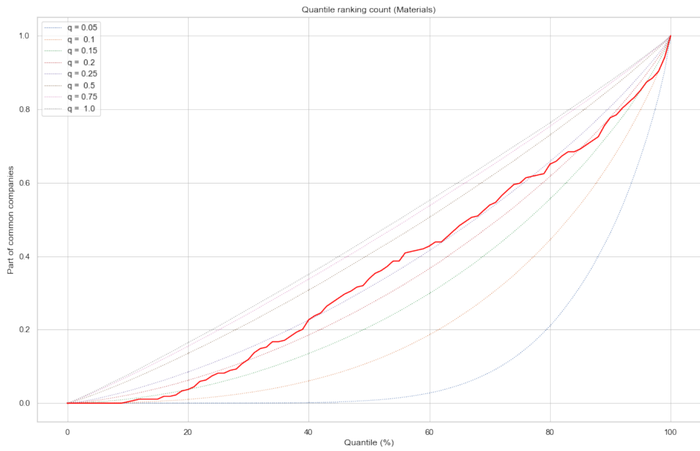


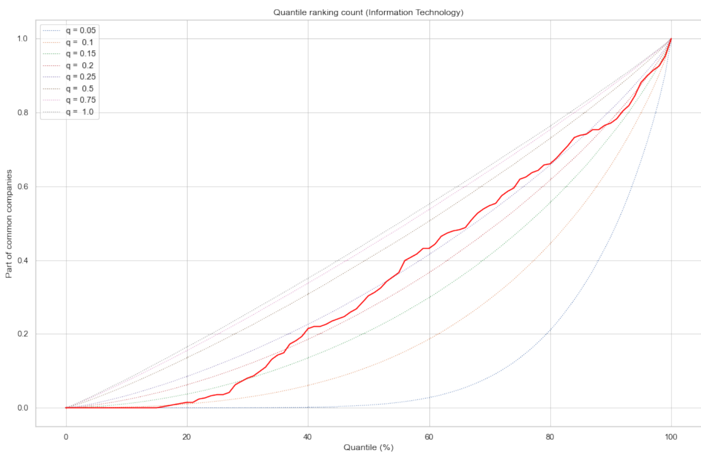



Quantile ranking count (Utilities)
q = 0.05
q = 0.1
q = 0.15
q = 0.2
q = 0.25
q = 0.5
q = 0.75
q = 1.0
Part of common companies
Quantile (%)


Quantile ranking count (Industrials)
q = 0.05
q = 0.1
q = 0.15
q = 0.2
q = 0.25
q = 0.5
q = 0.75
q = 1.0
Part of common companies
Quantile (%)


Quantile ranking count (Real Estate)
q = 0.05
q = 0.1
q = 0.15
q = 0.2
q = 0.25
q = 0.5
q = 0.75
q = 1.0
Part of common companies
Quantile (%)


Quantile ranking count (Materials)
q = 0.05
q = 0.1
q = 0.15
q = 0.2
q = 0.25
q = 0.5
q = 0.75
q = 1.0
Part of common companies
Quantile (%)


Quantile ranking count (Information Technology)
q = 0.05
q = 0.1
q = 0.15
q = 0.2
q = 0.25
q = 0.5
q = 0.75
q = 1.0
Part of common companies
Quantile (%)

ANNEXE 4: CATEGORIES BASED ON THE AFG ESSENTIAL METRICS

0	Carbon footprint
1	Volume of waste
2	Volume of waste recycled
3	Volume of water used
4	Energy consumption
5	Renewable energy consumption
6	Employee turnover
7	Training budget
8	Hours of training by employee
9	Participation of women in the company management
10	Accident rate
11	Accident severity rate
12	Number of fatal accidents
13	Patronage/charity
14	Part of independant administrators
15	Part of women in the board
16	Equity ratio
17	Country
18	CO2 Estimation Method
19	Turnover
20	ESG Reporting Scope
21	Number of employees
22	Board Structure Type