A Bloomberg Professional Services Offering

Content & Data Solutions

Sustainability Data

Bloomberg's Greenhouse Gas Emissions Estimates Model

A Summary of Challenges and Modeling Solutions

Bloomberg

Contents

04 Coverage

06 Model

10 Model Evaluation

Introduction

Governments, citizens, and companies around the world are increasingly taking action to reduce greenhouse gas (GHG) emissions. For investors, monitoring the GHG emissions of their portfolio companies is becoming an important part of the investment process. However, the availability of reported GHG emissions data varies tremendously across countries and business sectors, and many companies do not report their emissions at all.

In order to bridge this gap, Bloomberg developed a machine learning-based model to estimate the GHG emissions of companies. The Bloomberg GHG Model estimates direct (scope 1) and indirect (scope 2 and scope 3) emissions for companies with a sufficient amount of available data.

Scope 1 are GHG emissions directly related to a company's operating activities. Scope 2 are indirect GHG emissions resulting from purchased electricity, steam and heating/cooling. Finally, scope 3 are other indirect GHG emissions not captured by scope 2 that occur in the value chain of the reporting company, including both upstream and downstream emissions.

Using the model, we are able to provide robust estimates of scope 1 and scope 2 emissions for approximately 50,000 companies globally in 2020, compared to approximately 4,000 companies that self-reported their emissions in the Bloomberg ESG universe. Coverage for scope 3 includes the oil & gas and mining sectors, or around 4,000 companies. However, the scope 3 model will continue to expand to other sectors.

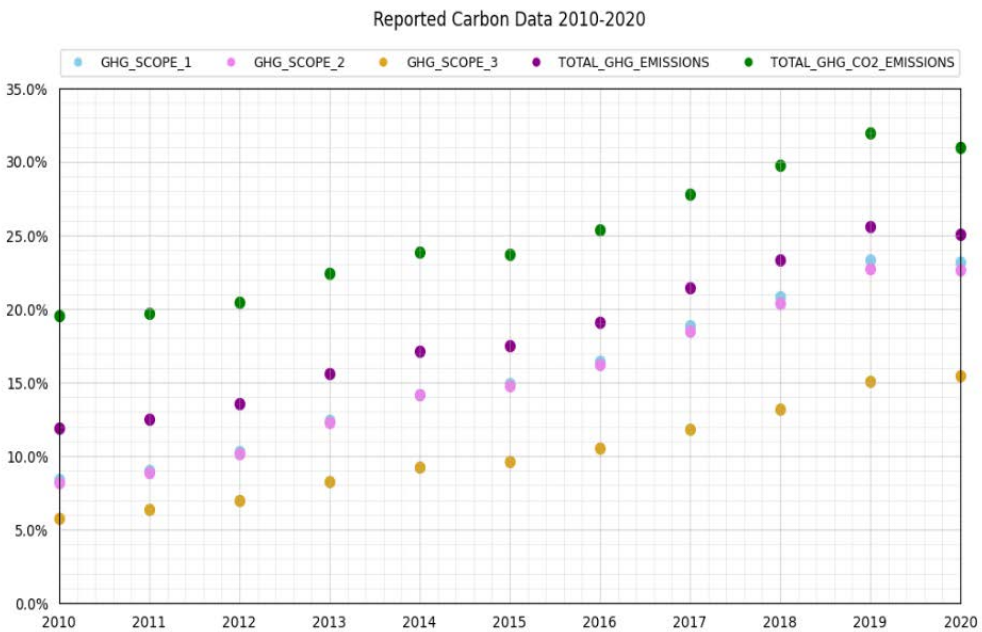

Figure 1. Percentage of companies reporting GHG emissions data in the Bloomberg ESG universe (currently composed of 11,700 global companies) which report ESG data.

Coverage

What companies are covered by the scope 1 & 2 model?

The total number of companies covered since 2010 is over 50,000 companies globally. 39,000 of them are publicly traded companies and the remaining 11,000 companies are private. The following tables show coverage broken down by index, region and country.

Table 1. Scope 1 & 2 coverage per index

Index Name	Coverage with reported GHG data	Coverage with reported & estimated GHG data
MSCI World Index	55.2%	100%
MSCI ACWI Index	54.11%	99.83%
MSCI Emerging Markets Index	41.11%	99.72%
MSCI AC Asia Pacific Index	33.7%	99.87%
CSI 300 Index	23.67%	100%
Russell 3000 Index	19.4%	94.4%
STOXX Europe 600	81.17%	99.33%
Bloomberg Global Agg Corporate Index	51.53%	95.46%

Table 2. Scope 1 & 2 coverage per region

Region	Number of companies covered
Americas	12,270
APAC	27,490
EMEA	10,765

Table 3. Scope 1 & 2 coverage. Top 10 countries

Country	Number of companies covered
United States	7,658
China	6,396
Japan	4,016
India	3,195
Canada	2,612
South Korea	2,453
Taiwan	1,979
Hong Kong	1,842
Australia	1,691
United Kingdom	1,601

What companies are covered by the scope 3 model?

The coverage for scope 3 GHG emissions estimates is over 4,000 companies globally. Around 90% of those companies are publicly listed and the rest are private companies.

Table 4. Scope 3 coverage per index

Index Name	Coverage with reported GHG data	Coverage with reported & estimated GHG data
Bloomberg World Oil & Gas Index	51.61%	93.54%
MSCI World Metals & Mining Index	54.05%	75.68%

Table 5. Scope 3 coverage. Top 10 countries

Ranking	Country	Number of companies covered
1	Canada	1,316
2	Australia	666
3	United States	364
4	China	300
5	United Kingdom	129
6	India	108
7	Russia	104
8	Vietnam	84
9	Hong Kong	79
10	Indonesia	66

Model

How did we decide on the right model to estimate GHG emissions?

Estimating the carbon footprint of companies is a complex task in itself; additionally, the data required to perform the estimation is noisy and often missing for the companies that must be estimated. Linear models offer a high degree of explainability, but they struggle when the underlying data contains interdependent relationships, missing values and categorical information – precisely the issues faced when producing GHG estimates. More intricate machine learning models, such as regression trees, can naturally learn complex relationships in the data, handle missing values, process categorical data, and model the inherent noise of GHG emissions. Based on these considerations, we decided to use a machine learning model based on regression trees. A regression tree is a type of decision tree model that delivers a numerical value. Decision tree models are similar to flowcharts, so a regression tree is conceptually just a flowchart where the final outputs are all continuously varying numerical values, such as price, temperature, or in this case, GHG emissions.

What data goes into the model?

The quality of GHG emissions estimates greatly depends on the quality of the data being used to generate them, and this is where Bloomberg excels. The Bloomberg GHG model uses multiple datasets, such as company location; size; and financial, environmental, social and governance (ESG) data; the breakdown of revenue by industry sectors; and industry-specific company data. Examples of industry specific data are the energy source (e.g., fossil fuel, solar, wind) used by utilities to produce electricity, or production data by cement, steel and oil & gas companies. In total, the model leverages over 800 individual features.

What does the model output?

The model produces estimates for scope 1 and 2 emissions for all industries and scope 3 for the oil & gas and mining sectors. Every estimate will have a unique distribution based on comparable companies. This allows users to select different percentiles in the distribution and use a more aggressive or conservative estimate than the one provided by the mean of the distribution.

Another element of the Bloomberg solution is the GHG Confidence Score, which is a measure of the depth and relevance of the data points available for the calculation of the greenhouse gas emission estimate for a particular company. The GHG Confidence Score is based on comparing the available data points for a given company and the most relevant data features for all companies in that same industry.

Finally, Bloomberg has created two sets of derived fields that complement the information provided by the estimate fields:

1. **Waterfall fields.** These fields are populated with reported GHG data, when available, and with estimated emissions for non-reporting companies.
2. **Intensity ratios.** These fields provide the amount of GHG emissions per unit of sales or enterprise value including cash (EVIC) to facilitate the comparison across companies.

How does the model come up with its estimates?

We train the model to learn the relationship between the data features of a company and the distribution of GHG emissions for companies with similar sets of features. The model training consists of applying a number of machine learning techniques, which are able to handle the complexity and challenges found in the data, to generate distributions. Finally, the model is able to apply these learned relationships to other companies.

Scope 3 model

Scope 3 emissions are estimated using a different model to the one employed to estimate scope 1 and 2 emissions due to several reasons:

- **Availability of reported data.** Scope 3 data is not as widely reported as scope 1 and 2.
- **Inconsistency in the scope 3 time series.** Due to the complexity in measuring scope 3 emissions, many companies have recently updated their methodology. This has resulted in the time series of reported data often showing large year over year variations, which are due to changing methodology, not changing emissions values.

Because of those challenges, there will not be a single scope 3 model that works across all companies. Therefore, industries or groups of industries may have their own specific models.

In-depth details about each scope 3 model can be found in the appendix section. Coverage of scope 3 is limited to oil & gas and mining companies at present, but coverage will increase as new industries are added in the near future.

Model distributions

Bloomberg's GHG emissions estimates model does not produce a single estimate but a full distribution of the estimated emissions. For each company and a given emissions metric, we provide the estimated expected value, or average of the distribution, and the emissions for percentiles between 1 and 99.

As an example, let's take a look at what the 75th percentile of the estimate means:

The 75th percentile provides the level of GHG emissions that will be equal to or larger than the emissions of 75% of comparable companies within the same industry. Naturally, the larger the percentile, the more conservative the estimate will be, with the 99th percentile being the most conservative estimate in the distribution.

This is better than producing a single estimate for two reasons. First, this allows us to precisely quantify all of the uncertainty within a GHG emissions estimate. Intuitively, the more data is available for a company and the more relevant that data is to emissions, the less uncertainty there should be about its estimated GHG emissions. In these cases, the distribution will be narrow, with values closely placed near the average. Companies with less data will display wider distributions with values further away from the average.

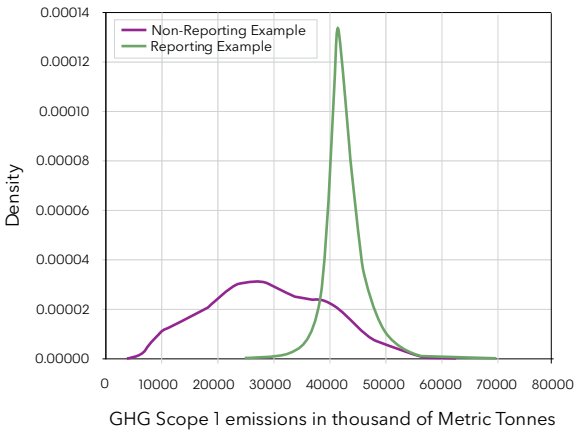

Figure 2. Examples of the scope 1 distributions for two companies. The green line shows a company with good data, low uncertainty and a high confidence score. The purple line shows a company with less data, high uncertainty and a low confidence score.

The second reason that a distributional estimate is more useful is that it allows users to adhere to the precautionary principle, as recommended by the European Union. This principle holds that when there is uncertainty in estimated emissions, it is more responsible to err on the side of protecting the planet, and hence take an estimate at a high percentile. Using estimates at the 75th percentile when carbon footprinting, for example, will help incentivise companies to report their emissions.

In summary, the distributions empower users to be as cautious as they feel is appropriate. The larger the percentile used, the more conservative the estimate will be.

Table 6. Scope 1 and 2 estimates across key percentiles for Spanish company Acciona SA.

Company Name	Acciona SA
GHG_SCOPE_1_ESTIMATE	100.476
GHG_SCOPE_1_ESTIMATE_25TH_PCTL	88.48
GHG_SCOPE_1_ESTIMATE_50TH_PCTL	98.394
GHG_SCOPE_1_ESTIMATE_75TH_PCTL	108.481
GHG_SCOPE_1_ESTIMATE_99TH_PCTL	194.808
GHG_SCOPE_2_ESTIMATE	107.914
GHG_SCOPE_2_ESTIMATE_25TH_PCTL	95.354
GHG_SCOPE_2_ESTIMATE_50TH_PCTL	103.544
GHG_SCOPE_2_ESTIMATE_75TH_PCTL	112.431
GHG_SCOPE_2_ESTIMATE_99TH_PCTL	244.951

Aren't companies that report GHG different from those that do not? Does this negatively affect the model?

This phenomenon is called distribution shift, and it is potentially a serious problem. One major source of distribution shift is the difference in *the amount of data reported:* companies that report GHG emissions also report many other related pieces of information, like energy consumption. To solve this problem, we train our model on data that is masked by applying the patterns of missing data from the companies that do not report their emissions to the ones that do.

For example, in the universe of carbon emissions-reporting companies, around 80% also report energy consumption – but in the universe of non-reporting companies, only 5% do. This is a challenge that can introduce reporting bias. In order to prevent that issue, we create a "mask" that essentially blots out the energy consumption of reporting companies until that also falls to a 5% level, and then train the model on that masked data. The result is that the model learns how to estimate the carbon emissions of companies that do not report much data.

Model Evaluation

How do we evaluate our model?

To evaluate our model, we use a technique called cross-validation.

Cross-validation is a technique in which we remove some of the reported data from the model training. Specifically, we split the reported data into ten sections, use nine of them to train the model, and then evaluate the remaining section; we then repeat this procedure nine more times.

Cross-validation gives us estimates on data that was not used for model training. We then analyze the performance on those out-of-sample predictions over multiple subsets of the data (i.e., does the model perform consistently across different industries and for differently sized companies?), and compare that performance against other baseline methods.

We have collaborated with subject matter experts inside and outside Bloomberg to discuss the model and review the estimates it produces. At Bloomberg, we worked with BloombergNEF (BNEF) and Bloomberg Industries experts, as well as the ESG team at large. Outside Bloomberg, we consulted with Professor Andreas Hoepner, Ph.D, at the University College Dublin, who advised on the importance of incorporating precautionary principles in the use of GHG estimated data.

Comparison of accuracy against baseline methods

We have compared our model against two other methods:

- **Sector Intensity.** This widely used model consists of calculating a carbon intensity metric (e.g., Scope 1 Emissions divided by Sales) for reporting companies, aggregating the data on an industry level and then taking the mean or the median as the "industry intensity ratio." We can then estimate GHG emissions for non-reporting companies as long as we know the company's industry segment and its revenue.
- **Linear Model per Industry.** In this approach, we create a linear model per industry to estimate the GHG emissions of companies. In each linear model, we select features that are relevant to the carbon footprint of companies in that industry and try to find those that best explain the reported carbon emissions. Examples of relevant features are the industry classification, revenue, net fixed assets, energy consumption and number of employees. When the reported value of a feature is missing, it's replaced by the industry average.

The following charts and tables show how each of the two models above perform against the Bloomberg GHG emissions model. We break down the analysis into two separate groups.

1) **Firms with good disclosure.**
   Companies in this group disclose company financials, industry segmentation and other relevant datasets that are related to their carbon footprint. This group primarily includes companies based in markets with good reporting standards and large, international companies.

2) **Firms with average or poor disclosure.**
   These companies tend to omit some of the features available for companies in the first group. In general, they are smaller in size and often located in emerging markets or markets with weaker disclosure standards. Many private companies will be included in this group as well.

For firms with good disclosure, both the linear approach and Bloomberg's model outperform the sector intensity approach, which confirms that not all companies in the same sector have a similar carbon intensity profile and points out that using industry averages or medians is not a good approach to estimate the carbon footprint of these companies. The linear model performs well but below Bloomberg's model as indicated by both the R-squared (R^2) and the root-mean-square error (RMSE).

Both R^2 and RMSE provide a measure of how well the estimated GHG emissions match the actual GHG emissions reported by companies. In other words, they provide a measure of how well the model performs at predicting GHG emissions.

More specifically, R^2 shows how much of the reported carbon emissions can be explained by the model, Therefore, the larger the R^2, the better the model is at predicting the carbon emissions of companies. By contrast, the RMSE measures the difference between values predicted by a model and the values observed. In this case, the lower the RMSE, the better the model is at predicting carbon emissions.

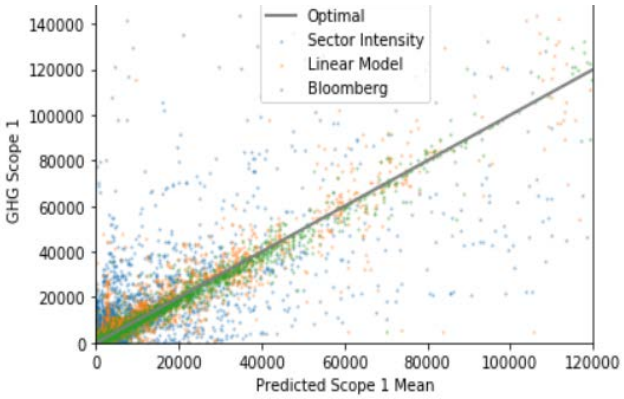

Figure 3. Scatter plot showing scope 1 predicted values versus observed values for companies with good data.. Ideally, both values should be equal and all dots should be on the optimal line, but this is not always achievable due to missing features and nosiness in the data.

Table 7. R^2 and RMSE for the sector intensity, linear model and the Bloomberg model for companies with good data. A large R^2 and a small RMSE indicates a strong model performance.

Model	R^2	RMSE
Sector Intensity	0.2997	7657.5
Linear Model	0.682	3464.8
Bloomberg GHG Model	0.8441	1703.9

For firms with average or poor disclosure,

Bloomberg's model continues to outperform both the linear model and the sector intensity approach. In this case, the linear model cannot handle the amount of missing data and performs very poorly. The sector intensity model is dependent on industry classification and revenue, so as long as that data is available its performance won't be affected.

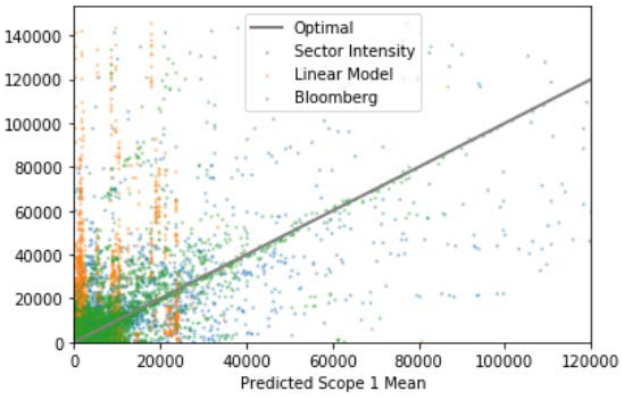

Figure 4. Scatter plot showing scope 1 predicted values versus observed values for companies with sub-optimal data. Ideally, both values should be equal and all dots should be on the optimal line, but this is not always achievable due to missing features and noisiness in the data.

Table 8. R^2 and RMSE for the sector intensity, linear model and the Bloomberg model for companies with sub-optimal data as seen in Figure 4. A large R^2 and a small RMSE indicates a strong model performance.

Model	R^2	RMSE
Sector Intensity	0.3181	7474.2
Linear Model	0.110591	9690.5
Bloomberg GHG Model	0.4108	6453.8

How do we know how well the model performs for companies that do not report?

In order to be able to fairly evaluate the models on non-reporting companies, we did a separate training pass in which we held out some companies with reported data from the training set, and also removed features from them that we would anticipate to be missing from non-reporting companies. This allowed us to get model predictions as close as possible to the situation of non-reporting companies, while still having reported values to compare against.

For example, say that we were to hold out Apple Inc., which does report its emissions. We would remove it from the training set entirely (so the model never sees it until evaluation time), and then mask out its data so that the remaining data is similar to what we would see for a non-reporting company. Then when we get the model's estimate of Apple's emissions using that masked data, we can actually compare to Apple's reported values.

Additional Reading

Want to learn more about Bloomberg's model?

Read the extended white paper, "Distributional Greenhouse Gas Emissions Estimates." In addition to the information in this report, that white paper includes additional data and exhaustive information on our research approach.

If you have any questions or would like to speak to a representative, please email eprise@bloomberg.net.

Appendix.

Scope 3 for Oil & Gas / Mining companies

The scope 3 model for oil & gas and mining companies combines a bottom-up model with a top-down machine learning model.

The bottom-up model uses companies' sales and production numbers on oil, gas, natural gas liquids, coal, iron ore, and more alongside carbon emission factors, i.e., the amount of CO2 equivalent emitted per unit of product. It then calculates the indirect emissions produced when using or processing those products. The top-down machine learning model sits on top of the bottom-up model and estimates carbon emissions by learning the relationship between calculated scope 3 emissions, revenue per industry and other key factors.

Using sales and production metrics for these two sectors works well because the most significant contribution to scope 3 emissions for oil & gas and mining companies comes from the downstream processing and use of their products. Other scope 3 components contribute in a much lesser way to the overall scope 3 footprint of firms in these two sectors.

There are a few exceptions within these sectors. They are 'Midstream - Oil & Gas', 'Oilfield Services & Equipment', 'Drilling & Drilling Support' and 'Mining Services', which may not have any production numbers. Those companies will leverage a similar model to the one explained for scope 1 & 2 emissions.

About Bloomberg.

We are the central nervous system of global finance. Born in 1981, Bloomberg is a forward-looking company focused on building products and solutions that are needed for the 21st century. As a global information and technology company, we connect decision makers to a dynamic network of data, people and ideas – **accurately delivering business and financial information,** news and insights to customers around the world.

Take the next step.

For additional information, press the <HELP> key twice on the Bloomberg Terminal®.

Beijing
+86 10 6649 7500

Dubai
+971 4 364 1000

Frankfurt
+49 69 9204 1210

Hong Kong
+852 2977 6000

London
+44 20 7330 7500

Mumbai
+91 22 6120 3600

New York
+1 212 318 2000

San Francisco
+1 415 912 2960

São Paulo
+55 11 2395 9000

Singapore
+65 6212 1000

Sydney
+61 2 9777 8600

Tokyo
+81 3 4565 8900

bloomberg.com/professional