METHODOLOGY ABSTRACT

Greenhouse Gas Emissions

Version 1.0

November 2023

Anabel Ng
Manager,
Methodology & Product Architecture

Clark Barr
Director,
Methodology & Product Architecture

Giacomo Bressan
Associate Director,
Methodology & Product Architecture

Dr. Hendrik Garz
Global Head of ESG,
Methodology & Product Architecture

Greenhouse Gas Emissions

Methodology Abstract

10 November 2023

Methodology & Product Architecture

Authors:[1]

Anabel Ng
Manager, Methodology & Product Architecture
anabel.ng@morningstar.com

Clark Barr
Director, Methodology & Product Architecture
clark.barr@morningstar.com

Giacomo Bressan
Methodology Manager, Methodology & Product Architecture
giacomo.bressan@morningstar.com

Dr. Hendrik Garz
Global Head of ESG, Methodology & Product Architecture
hendrik.garz@morningstar.com

Global greenhouse gas (GHG) emissions have been on a steady upward trend in the last decade. Considering several global agreements to mitigate these emissions, including the 2015 Paris Agreement, it is critical to assess the GHG emissions of companies. Company reporting of emissions has improved but is still limited. To bridge this gap, Morningstar Sustainalytics has created proprietary Scope 1, Scope 2, and Scope 3 GHG Emissions estimation models that assess a company's overall carbon footprint. These models are expected to evolve over time as the reporting landscape continues to change.[2]

Highlights

- Morningstar Sustainalytics provides a comprehensive coverage of GHG emissions data through the collection of company reported data and fills remaining gaps with estimated data.
- To supplement the reported data, there are multiple estimation model techniques that are used depending on the type and availability of data (i.e., source type hierarchy).
- One of the approaches is our multi-factor estimation model, which estimates the **Scope 1**, **2** and **3 Emissions**.[3] It is based on factors related to size, along with factors specific to subindustry, activity, and country.
- Our common multi-factor model framework facilitates consistency and comparability for Scope 1, 2 and 3 emissions, while introducing refining factors to provide flexibility to adjust for structural differences in the underlying data and address limitations of data availability.
- The Scope 1, 2 and 3 models achieved an R-squared value ranging from 76% to 84%, depending on the category scope of emissions.

Introduction

Effectively measuring and monitoring emissions lays the foundation for decarbonization efforts

There has been an enhanced focus on decarbonization and commitment to climate action by governments and companies. This trend is further supported by stricter regulation on a global scale and increasing investor and consumer awareness. For most companies, the first step towards effectively reducing emissions is to measure and monitor their current stock of greenhouse gas (GHG) emissions. However, there continues to be limited disclosure rates, particularly for small companies with fewer resources.

GHG emissions estimation models were created to expand emissions coverage

Recognizing the need for more comprehensive GHG emissions-related information, Morningstar Sustainalytics has created multiple estimation model techniques. One of the techniques created is a proprietary multi-factor regression model to estimate Scope 1, 2 and 3 GHG emissions for companies that do not report these data yet. The model considers multiple size-related factors, as well as business-model-related characteristics, as reflected in industry- and country-specific characteristics. Given the challenges with data availability and reporting inconsistency, the results are compelling, with average R-squared values of 84%, 76% and 78% for Scope 1, 2 and 3 respectively.

Methodology Description

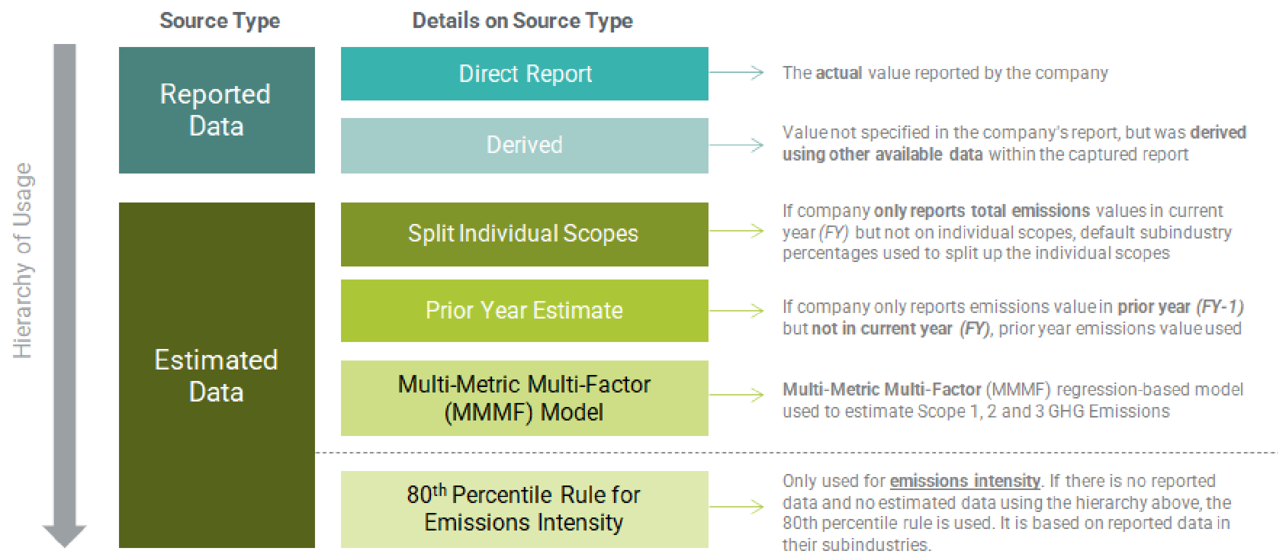
GHG Emissions Data

To determine what source type is used, the source type hierarchy is applied to the emissions dataset

In Morningstar Sustainalytics' GHG emissions dataset, there are two **Source Types** of data: Reported and estimated. As each company will only be assigned one emissions value, a source type hierarchy has been developed as depicted in Exhibit 1 below. This will depend on the availability and type of company data. More details will be explained in the following section.

Given that reported data falls high on the hierarchy and the estimation models are built on reported emissions data, it is important that the reported data is reliable and of high quality. For exceptional cases, Morningstar Sustainalytics may disqualify the reported data and an estimated value will be used instead. Reasons for disqualification include: Uncertain units of measurement, exceptionally high intensity values (e.g., contradictory unit references on different sections of the company reports); and inconsistent and erroneous reporting from year-to-year (e.g., frequent corrections and restatements).

Exhibit 1: Source Type Hierarchy



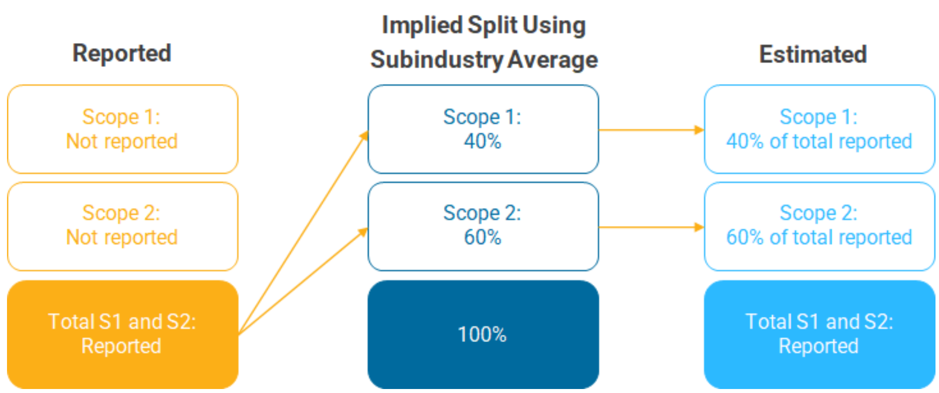
Source: Morningstar Sustainalytics

Split Individual Scopes

Estimating individual scopes when companies report aggregated emissions

The individual emissions scopes were estimated in cases where the company reports an aggregated emissions value but not the individual scopes. As Exhibit 2 illustrates, the implied split (e.g., company's Scope 1 or Scope 2 emissions, as a proportion of the total Scope 1 and 2 value) is based on the subindustry average split. The same approach is implemented for total Scope 1, 2 and 3.

Exhibit 2: Estimating Individual Scopes Using Subindustry Average Split



Source: Morningstar Sustainalytics

Prior Year Estimate

Minimizing the volatility in the historical GHG emissions data

To minimize the volatility in the historical GHG emissions data, a simpler approach is used to proxy the company's emissions. That is, if the company has a reported value in the previous year (FY-1) but not in the current year FY then (FY-1) emissions data are used as the second choice for estimation.

Multi-metric Multi-factor Model

The multi-metric multi-factor model is trained and calibrated on reported information

One of the more complex estimation model techniques employed is the size-related factors, along with subindustry-, activity- and country-specific factors. The dependent variable in the model is a company's Scope 1, 2 or 3 emissions for a given year. The model is trained and calibrated on reported information. The reported data is subject to rigorous cleaning procedure and statistical correction techniques to avoid inclusion of incomplete or inconsistent data as an input to our model calibration. The model outputs are analyzed for large deviations to further improve stability and performance.

GHG Emissions MMMF Models

The Factors

Relevance, data availability and quality support our estimation models

In general, the selection of factors for our estimation models is underpinned by the following criteria:

- Relevance: The factor is seen as a driver of and used to explain Scope 1, 2 or 3 emissions.
- Data availability and quality: The underlying data should be accessible, widely available, reliable, and as complete as possible (e.g., broad coverage over the carbon universe).

Size-related Factors

The company's size affects the amount of emissions produced

One of the main common drivers in all factor models is a company's size. The underlying assumption is that companies that are larger in size tend to have higher emissions. For instance, companies that own more buildings and machinery (captured in the relatively higher PP&E value) compared with their subindustry would require more electricity to run its day-to-day operations, thus resulting in higher Scope 2 emissions.

Addressing the different upstream and downstream emissions sources for Scope 3

Within the Scope 3 estimation model, the size-related variables also represent our attempt to address the different upstream- and downstream-related Scope 3 emissions sources. The company's revenues, for example, serve as a proxy for the volume of products and services sold. The cost of revenues, on the other hand, serves as a proxy for the more upstream-related value chain activities.

Subindustry Factor

Capturing subindustry-specific regulations, trends, and risks

Across all three estimation models, the subindustry factor is used to capture the industry-specific regulations, trends, and risks related to GHG Emissions. It is based on the 135 different subindustries in the **Morningstar Sustainalytics' Industry Classification System**.

Relying on Activity-based Research

To account for the variation within subindustries and better capture the heterogeneity of the underlying GHG emissions, several subindustries have been broken down into smaller subsets, known as **Subindustry Segments**, which translates to additional dummy variables. The segments are based on the **Activity-based Research** (i.e., research conducted at the level of a companies' business activities). This includes company's activities that have higher emissions, also known as 'brown' activities (e.g., involvement in Oil Sands Extraction activities) and company's activities that have lower emissions, also known as 'green' activities (e.g., involvement in Manufacture of Other Low Carbon Technologies activities).

To create these Subindustry Segments, several criteria were fulfilled:

- Relevant to metric under examination: Activities that could impact a company's emissions (positively or negatively).
- Meaningful distinction: Activities that could meaningfully distinguish a company's emissions relative to its subindustry.
- Reasonable data size: Sufficient reporting companies within each subset.

Country Factor

Company-level emissions show country-specific emissions features

One model assumption is that company-level emissions can display country-specific emissions characteristics, mainly driven by regulation and policies. To account for these characteristics, a country index was created using several country-level factors, such as GHG emissions by country and Morningstar Sustainalytics Country Risk Ratings. The country factors were standardized and consolidated based on the pre-defined weights.

Allocating the country level data down to the company level

To allocate the country level data down to the company level—assigning an individual company to a country or set of countries—we selected the most appropriate geographical data point, using the following prioritization:

- Asset location (i.e., country breakdown of total assets)
- Primary listing location
- Headquarters location

Subsequently, country index outcomes were divided into quintiles, with the bottom quintile corresponding to low pollution geographical presence and the top quintile corresponding to a high pollution geographical presence. Next, companies were assigned to one of these quintiles, depending on the available data.

The Regression Model

The input factors as shown in Exhibit 3 below serve as explanatory variables in a classical multi-factor regression model. Reported Scope 1, 2 or 3 emissions—expressed in metric tonnes—serve as the dependent variable.

Exhibit 3: Scope 1, 2 and 3 Regression Model

No.	Factors	Emissions Scope Applicability
1	Total Revenues	Scope 1, 2 and 3
2	Number of Employees	Scope 1, 2 and 3
3	Gross Plant, Property & Equipment	Scope 1, 2 and 3
4	Cost of Revenues (aka Cost of Good Sold)	Scope 3
5	Subindustry	Scope 1, 2 and 3
6	Subindustry Segments	Scope 1 and 2
7	Country	Scope 1, 2 and 3

Source: Morningstar Sustainalytics

A standardized approach to ensure scalability

One important consideration in the model selection process was the need for scalability. To fill the missing values in our carbon universe with modelled estimates, a standardized approach with common factors was employed. Although these generalizations can dilute differences at the subindustry-level, the standardized approach facilitates consistency and creates a clear, structured process that can be replicated for large sets of data across different industries.

A flexible model to capture any improvements in disclosure over time

Notably, with the increasing stakeholder and regulatory pressure to disclose total emissions, this model provides the flexibility to capture any improvements in disclosure over time.

Data Preparation and Cleaning

Training and testing the model with the reported Scope 1, 2 and 3 emissions

After establishing the relevant inputs for the model, the required data points were collected for all companies in the carbon product universe. As the model requires a full set of input data to estimate a Scope 1, 2 or 3 emissions outcomes, missing values for any one of the size-related input factors—such as Operating Revenues, Number of Employees, PP&E and Cost of Revenues—are imputed values derived from those ones where data is available.[4]

A stable and reasonably sized dataset for building the model

Although the quality of emissions data is improving, companies still face difficulties in collecting and tracking relevant data, leading to limited disclosure and errors in reporting. A rigorous cleaning process is employed to ensure that only reliable data is used for the model calibration. These steps include removing outliers on a subindustry-level and removing datapoints if there was an inconsistent historical reporting pattern. The purpose of the data cleaning exercise is to provide a stable and reasonably sized dataset to build the regression model, not to remove all the inaccurate or incomplete reported data points.

Data Adjustments

Reducing the underreported bias in Scope 3 data

As a next step in containing underreporting bias, an oversampling algorithm is applied to the cleaned Scope 3 dataset, with the following characteristics:

- Oversamples the minority class (i.e., companies with higher emissions).
- Randomly undersamples the majority class (i.e., companies with lower emissions).

The underlying assumption to this approach is that companies tend to underreport their emissions and hence higher values tend to be underrepresented in the dataset. By removing these underreported values while synthetically creating data points on the higher end, the underreporting bias is expected to be reduced.

Ensuring data quality and fulfilling assumptions

Several further adjustments were introduced to the underlying data to fulfill the assumptions, given that the model is a linear regression, as follows:

- All the input factors, except for the country and subindustry factor, are log-transformed to reduce skewness and create a more normal distribution.
- Binary variables are created for each of the factor quintiles to which companies have been assigned, based on factor outcomes as described above.
- The cost of revenues factor has been orthogonalized to minimize multicollinearity between revenues and cost of revenues in the Scope 3 model.
- Subindustry averages that are used as an explanatory variable in our model have been z-scored to minimize spurious correlation caused by having a Scope 3 variable on both sides of the equation.[5]

Training and Testing the Model and Quality Controls

Following data adjustments, the dataset was broken up into training and test sets. The training set that is used to calibrate and fit the model represents 80% of the data, while the test set that is used to evaluate the model represents the remaining 20%. In this way, issues around overfitting are identified and controlled. The Ordinary Least Squares method is used to estimate the model parameters. A Robust Standard Errors approach is applied to control for heteroskedasticity.

A feedback loop to improve performance and stability

To further stabilize the model, a feedback loop has been implemented, which involves identifying points of influence with large residuals and removing these if appropriate. This resulted in more stable models with improved performance.

Empirical Results

Training and Testing Results

Strong results, with the input factors explaining the reported emissions relatively well

To train and calibrate the models, we used the most complete and comprehensive emissions dataset available at the point of the models' development. For the other models' inputs, we used the data from the same fiscal year as the emissions data to maintain consistency. In cases where the data for the same fiscal year were not available, the latest available information was used, only if it can be reasonably assumed that there is sufficient stability in the data over time.

For our FY2020 dataset, the models obtained an R-squared of 84%, 76% and 78% for Scope 1, 2 and 3 respectively. The coefficients were significant at a 5% level across all three models, with emissions displaying an especially strong and positive relationship with the Revenue, PP&E and subindustry factors. The models' predictive power is expected to increase over time, as emissions reporting increases, as well as other relevant supporting data (e.g., financial or country data) for the estimation becomes available.

Rule of the 80th Percentile

80th percentile rule used for emissions intensity for companies with limited to no data available

Lastly, for companies that have limited data available, more specifically no reported or estimated emissions value, and no reported financial data, an 80th percentile rule is employed for emissions intensity. It is computed using its subindustry classification, if the number of companies in the subindustry is over or equal to ten. Otherwise, the peer group classification is used instead.

Conclusion

Arriving to a strongly grounded framework

Given that emissions disclosure rates remain low, several estimation model techniques have been employed to supplement the reported data. One of the most complex techniques developed is the multi-factor estimation model that is based on size-related factors, in combination with subindustry-, activity- and country-specific factors.

The models performed relatively well, with largely significant individual factors, especially given the data's nature and the limited emissions reporting. While acknowledging the limitations of the model regarding the generalizations of certain complexities, the framework employed is strongly grounded and it is easily adaptable to increased and improved reporting of emissions over time.

Glossary of Terms

Activity-based Research	Research conducted on the specific business activities of a given company.
Greenhouse Gas Protocol Corporate Accounting and Reporting Standard	The GHG Protocol Corporate Accounting and Reporting Standard provides requirements and guidance for companies and other organizations preparing a corporate-level GHG emissions inventory.
Morningstar Sustainalytics' Industry Classification System	Morningstar Sustainalytics has defined its own industry classification system to allow grouping companies in Morningstar Sustainalytics Ratings universe according to their business activities in three levels: Subindustries (bottom level: The highest granularity), Industry Groups (medium level: Medium granularity) and Sectors (top level: Lowest granularity).
Multi-metric Multi-factor Model Framework (MMMF)	The MMMF can be viewed as a master model designed to estimate multiple metrics related to pollution and environmental impacts. It distinguishes between two types of input factors, (a) the so-called common factors such as subindustry, size, and country of operations; and (b) the refining factors, which are used only for individual MMMF specifications, such as the Scope 3 Emissions Model. All models are specified as linear regressions and get trained and tested based on similar processes and standards.
Scope 1 Emissions	As defined by the GHG Protocol Accounting and Reporting Standard, Scope 1 refers to direct emissions that are from company-owned and controlled resources. It is typically reported in metric tonnes.
Scope 2 Emissions	As defined by the GHG Protocol Accounting and Reporting Standard, Scope 2 refers to indirect emissions that are from the generation of purchased energy, from a utility provider. It is typically reported in metric tonnes.
Scope 3 Emissions	As defined by the GHG Protocol Accounting and Reporting Standard, Scope 3 refers to all other indirect value chain emissions, beyond those covered in Scope 2. Scope 3 emissions are divided into 15 categories and cover both upstream (e.g., purchased goods and services) and downstream emissions (e.g., use of sold products). It is typically reported in metric tonnes.
Source Type	Refers to the origin of the company specific data point information. The Source Type can be either reported or estimated.
Subindustry Segments	In the context of the estimation models, subindustries have been broken down into smaller segments: Subindustry Segments, to reflect a company's business activities in a more granular manner and to account for within-subindustry variations. To determine this, activity-level research has been used. For example, the Integrated Oil & Gas subindustry can be broken down into two segments: One for higher intensity producers (with thermal coal, shale energy and/or oil sands activities), and other for lower intensity producers (without thermal coal, shale energy and oil sands activities).

Endnotes

[1] The authors would like to thank their Sustainalytics colleagues who helped in the preparation of the report. Sector research support was provided by Alex Osborne-Saponja. Quantitative modelling support was provided by Timo Schäfer. Cristina Zabalaga performed the editorial review.

[2] See the full methodology description here: Barr, Bressan, Garz, Ng (2023). “Methodology Description: Greenhouse Gas Emissions”. Morningstar Sustainalytics

[3] Text that is highlighted in bold teal indicates a term that is explained in the glossary of terms in the Appendix.

[4] Our approach requires that data for at least one of the four specified input factors must be available to complete an estimation. To obtain an estimate, the missing financial values were imputed using a standard K-Nearest Neighbor (KNN) approach. Using this approach, companies within the same subindustry are grouped together and the average value of the set of companies that are most similar in terms of the non-missing financial variable was computed.

[5] This means the subindustry averages have been standardized by subtracting the mean and dividing by the standard deviation.

Change Log

Version	Date	Initiator	Main items introduced/changed	Comment / Rationale
1.0	10.11.2023	Climate Solutions Methodology Team	▪ Creation of GHG Emissions Methodology Abstract.	N/A

About Morningstar Sustainalytics

Morningstar Sustainalytics is a leading ESG research, ratings, and data firm that supports investors around the world with the development and implementation of responsible investment strategies. For 30 years, the firm has been at the forefront of developing high-quality, innovative solutions to meet the evolving needs of global investors. Today, Morningstar Sustainalytics works with hundreds of the world's leading asset managers and pension funds who incorporate ESG and corporate governance information and assessments into their investment processes. The firm also works with hundreds of companies and their financial intermediaries to help them consider sustainability in policies, practices, and capital projects. With 17 offices globally, Morningstar Sustainalytics has more than 1,800 staff members, including more than 800 research analysts with varied multidisciplinary expertise across more than 40 industry groups. For more information, visit www.sustainalytics.com.