

Company Emissions Estimates for Private Markets

A Machine Learning Approach for Investors

Dr. Ben McNeil
Dr. Nicholas Pittman



EMMI

24 FEBRUARY 2025



Summary

Private equity (PE) is a significant asset class for investors worldwide. It can be up to 20% of a multi-asset portfolio and has reached \$8.2 trillion in total. However, the lack of emissions reporting requirements in the unlisted company sector has made it difficult to estimate the carbon footprint of private equity assets, and the sector as a whole. This is an increasing problem for investors and other funders as their reporting requirements ramp up globally in 2025.

In this paper, we demonstrate an emissions modelling approach that overcomes this issue of data scarcity for unlisted companies, and significantly improves the accuracy of emissions modelling for private markets.

We present a machine learning (ML) approach to estimating Scope 1, 2 & 3 emissions for any private company using as little as one financial metric (e.g. revenue or total assets). We also compare our method to the traditional

approach of using Industry Emission Factors from Input-Output Models.

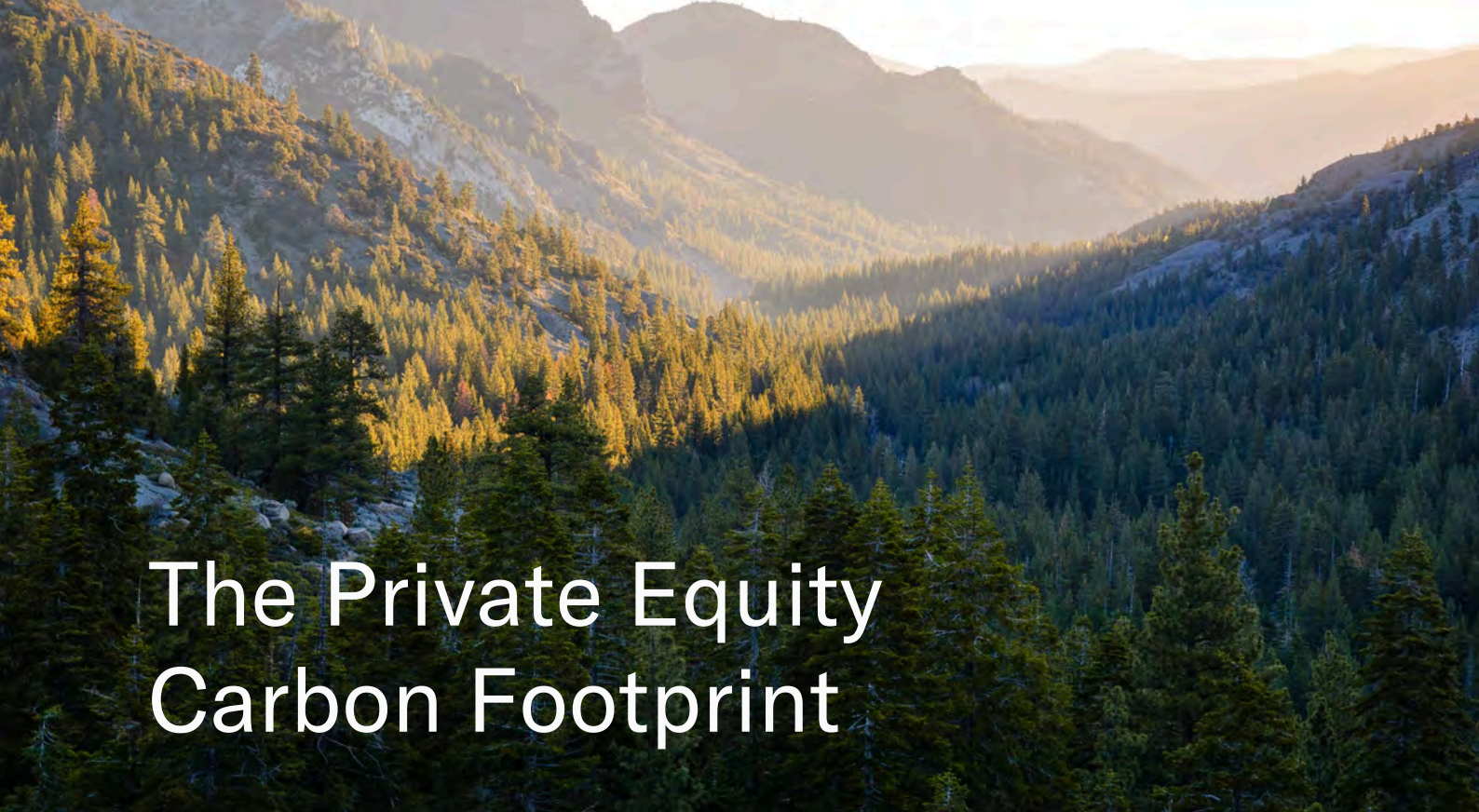
We find our ML approach to be up to seven times more accurate than the traditional industry factor approach. This outperformance is achieved by accurately capturing non-linear relationships in real world data, instead of relying on generalised industry averages within a theoretical model.

Private equity data are often scarce, but with the ML method, even sparse input data is used more intelligently to greatly improve the accuracy of quantifying emissions for private markets.

The new ML approach also allows uncertainties to be quantified at the company level for the first time. This approach now gives investors accurate but comprehensive emissions estimates for private markets to assess materiality, transition risk and reporting.

Contents

Summary	1
The Private Equity Carbon Footprint	3
Estimating Private Companies Emissions	5
Limitations of Industry Factor Approach	6
New Machine Learning Approach to Estimate Private Company Emissions	7
The Emmi Machine Learning Approach for Private Equity	8
Financial Data	9
Carbon & Financial Input Data	9
Model Selection	10
Private Equity Data Constraints	10
A Customer-Centric Framework for Model Evaluation	10
Testing and Validation	10
Accuracy	10
Predictability	11
Robustness	11
Noise Injection Test (MdAPE*)	11
Performance	12
Scope 1 Performance	12
Revenue & Employees Model	12
Revenue Only Model	12
Total Assets Model	12
Enterprise Value Model	12
Practical Implications	13
Scope 2 Performance	14
Comparison with Traditional Industry Factor Approach	16
Conclusion	18
About Emmi	19



The Private Equity Carbon Footprint

Private equity is an asset class that involves investing directly in businesses that are not publicly traded on any stock exchange. This allows investors to access an alternate pool of opportunities such as early-stage startups, growth-oriented companies, and established businesses seeking expansion or restructuring. As such, they are often long-term commitments, allowing a private equity firm time to create value through considered strategies and plans. These more patient investors trade any temporary business volatility for higher potential returns in due course.

PE can make up as much as 20% of an institutional investor's multi-asset portfolio.

The total value of global PE assets reached a record high of \$8.2 trillion in 2023.¹ That year, industrial carbon emissions were 37.2 billion tonnes.² Assuming PE is roughly 8% of global Gross Domestic Product (GDP) (~\$105 trillion), then the sector's total carbon footprint was about 3 billion tonnes. This is not insignificant in comparison to the public equity footprint of ~11.4 billion tonnes.³

”

The total value of global private equity assets has reached a record-high \$8.2 trillion by 2023.

¹ <https://www.mckinsey.com/industries/private-capital/our-insights/mckinseys-private-markets-annual-review>

² <https://essd.copernicuhttps://essd.copernicus.org/articles/14/4811/2022/s.org/articles/14/4811/2022/>

³ McNeil, B., Pittman, N., Kitto, A. and Greven, H. *Estimating the Corporate Carbon Footprint: A Meta-Model Machine Learning Approach*, (2024), <https://www.emmi.io/newsroom/white-paper-nov-2024>

Why It's Important?

Climate Mandates Now Include Private Companies

Governments have primarily targeted public companies in their efforts to regulate carbon emissions. However, private companies are now also being held accountable. California recently passed legislation requiring both public and private American companies, that 'do business in California' and have >\$500 million annual revenue, to report their greenhouse gas (GHG) emissions and potential financial risks related to climate change. Similarly, the United Kingdom, Australia, Singapore, and New Zealand are also requiring private companies to disclose information on their climate impact.

Exits Will Need Carbon Due Diligence

Private equity often comes at a pivotal juncture in a company's development, and can be integral in assisting plans for major company investments, including to lower carbon emissions. A growing portion of PE firms are incorporating climate-related opportunities into their investment strategies and decisions, as these will likely have better 'exit' feasibility - the potential buyer group are now more than likely to have made climate commitments, and will also require an understanding of the financial cost of a carbon price or regulation on the valuation of the company.

Net Zero Commitments Growing

The number of Net Zero commitments for private equity is on the rise. The Initiative Climat International (iCI) is a global community of private equity firms and investors dedicated to understanding and managing the risks associated with climate change. Originally launched as iC20 in 2015 by French private equity firms, it has now grown to include over 212 firms representing combined assets under management (AUM) of \$3.2 trillion (as of October 2022).

Estimating Private Companies Emissions

Investors seeking to quantify their portfolio's 'financed emissions' typically only have access to financial data from their investee companies like revenue, employees, liabilities, net assets, capital expenditure, and sometimes not even that.

The most common approach to estimating private company emissions is Spend-based Industry Emission factors, as they are somewhat easy to apply. These directly translate financial values into CO₂e per dollar spent or earned. They rely on Environmentally-extended Input-Output (EEIO) models that map the flow of goods and services in the economy. However, the EEIO models only use sector averages and do not account for any company-specific differences in suppliers, processes, or efficiencies.

The Exiobase model,⁴ for example, includes 163 sectors across 44 countries and five 'rest of the world' regions, with estimates of economic output and total emissions for each combination of sector and country.

For example, if we assume the model suggests a revenue of \$100 billion for the United States steel manufacturing sector which produces 500 million tonnes of CO₂e, the resulting EEIO emission factor is 5 tonnes of CO₂e per \$1000 revenue. Using this industry factor, you can estimate emissions for a specific steel company (Company A) by multiplying their revenue by the factor (e.g. \$2 billion x 5 tonnes/\$1000 = 10 million tonnes CO₂e).

⁴ <https://www.exiobase.eu>



The most common approach to estimating emissions for private companies is via Industry Emission Factors

Limitations of Industry Factor Approach

No Breakdown of Scope 1, 2, or 3 Emissions

The approach uses an economic model, so there is no breakdown in Scope 1, 2 or 3 emissions. Financed emissions should only take into account the Scope 1 & 2 emissions of the portfolio - while Scope 3 can be included separately to understand the full carbon transition risk impact of the portfolio.

The Largest Components of Scope 3 are Missing

The largest components of Scope 3 are missing from EEIO factors, which has serious implications for accurately assessing carbon transition risk. Our research has shown the downstream Scope 3 emissions categories 'Product Use' and 'Product Processing' are by far the most material. 'Product Use' contributed ~60% of Scope 3, while 'upstream' categories only contributed ~10%. Some data providers use an 'uplift factor' to try and account for the large missing downstream component. However, this likely introduces even more unsystematic biases.

Results are Averages

Results are averages, meaning that all companies in the same sector receive the same emissions intensity factor per million of revenue (this is a major limitation and explored in detail later).

No Understanding of Uncertainties

The output of EEIO factors are taken as given, which means there is no systematic quantification and understanding of emission uncertainties for each public company. This black-box approach means it is therefore difficult to assess sensitivity and materiality of holdings within a portfolio.

New Machine Learning Approach to Estimate Private Companies Emissions

Machine Learning models are sophisticated algorithms that learn from both successes and failures by 'training' and categorising patterns. They can effectively analyse large, non-linear datasets, such as financial markets, in order to identify complex patterns. While machine learning has a long history of use in finance, its application in the field of carbon is relatively recent.

Introducing Machine Learning to Estimate Carbon Emissions

Nguyen et al (2021) were the first to present a new approach using ML to estimate corporate GHG emissions using publicly available information. Their most successful model was a two-step process that utilised a meta-learner to combine predictions from six different base-learners: Ordinary Least Squares, Elastic Net, Neural Network, K-Nearest Neighbours, Random Forest, and Extreme Gradient Boosting. They demonstrated that their ML was up to 30% more accurate than traditional revenue-based emission factors in predicting total emissions (Scope 1 + Scope 2).

Since 2021, Emmi has partnered with these pioneering climate finance researchers from the University of Otago and Griffith University to extend and apply those ML methods to estimating corporate carbon emissions for all publicly-listed companies and private companies.

Nguyen et al (2023) was a follow-up study between Emmi, the University of Otago and Griffith University that researched the data quality and predictability of Scope 3 using ML methods. They showed significant under-reporting of Scope 3 reported data. Machine learning techniques were found to improve predictions of Scope 3 by up to 25% when compared to traditional linear regression approaches. This research formed the basis to develop and extend prediction capabilities that allow Emmi to generate 'now-casted' emissions estimates for all public companies - irrespective of whether they have reported carbon emissions.

Nguyen Q, Diaz-Rainey I, Kuruppuarachchi D, (2021) Predicting corporate carbon footprints for climate finance risk analyses: A machine learning approach, Energy Economics, Volume 95, 105129.

Nguyen Q, Diaz-Rainey I, Kitto A, McNeil BI, Pittman NA, Zhang R, (2023) Scope 3 emissions: Data quality and machine learning prediction accuracy. PLOS Clim 2(11): e0000208. <https://doi.org/10.1371/journal.pclm.0000208>

The Emmi Machine Learning Approach for Private Equity

In regions like the EU, UK, and even the US, public companies are required to disclose their GHG emissions and climate-related risks. On the other hand, private companies have fewer mandatory reporting obligations and often only disclose their carbon data voluntarily or at the request of certain stakeholders (such as private equity owners, major customers, or lenders). Without an independent private equity carbon dataset, assessing model performance is hard.

In order to build and assess private equity emissions models, we use the public equities carbon dataset and assess models with limited data inputs across a typical range of private companies globally.

In essence, we assume that private companies operate in a similar way to those in public markets. By matching the public dataset with limitations found in the private equity markets, we can explore different models and optimise the performance.

(Emmi's unique meta-model machine learning framework for public companies is designed for quality, flexibility, coverage, and timeliness (see white-paper⁵).

⁵ McNeil, B., Pittman, N., Kitto, A., and Greven, H (2024) Estimating The Corporate Carbon Footprint: A Meta-Model Machine Learning Approach. <https://www.emmi.io/newsroom/white-paper-nov-2024>



By matching the public dataset with data limitations found in the private equity markets, we can explore different models and optimise the performance.

Financial Data

Each year, we gather the financial and business information in Table 1 for all publicly listed companies. These data provide insight into the size, profitability, assets, and location of the company. Data were extracted for 2016 to 2024 using the FactSet Company Fundamentals database, totalling 50,236 companies.

Carbon & Financial Input Data

We source the reported carbon emissions of over 10,000 companies from FactSet ESG Solutions for 2021, 2022 and 2023. We only use data from 25 advanced countries, with more mature reporting standards.⁶ The data were collected from companies' extra-financial communications and annual reports and merged with financial data using FactSet. Before training our models, we performed data cleaning and validation on both the reported carbon emissions and financial input data. To identify anomalies in reported carbon emissions, we performed year-on-year difference checks, excluding carbon data that had variations of 100%+ or manually cross-checking annual reports. This cleaning resulted in about 300 companies or 5% being excluded from the training dataset.

Scope	Type of Emissions	Number of Companies	
		Training	Testing
Scope 1	Direct Emissions	8,983	1,885
Scope 2	Indirect Emissions (Electricity)	6,970	1,523
Scope 3	Other Indirect Emissions Upstream and Downstream	3,772	782

⁶ Australia, Austria, Belgium, Canada, Denmark, Finland, France, Germany, Ireland, Israel, Italy, Japan, Luxembourg, Netherlands, New Zealand, Norway, Poland, Portugal, Singapore, South Korea, Spain, Sweden, Switzerland, United Kingdom, United States

Model Selection and Performance

Private Equity Data Constraints

Private equity presents a unique challenge for building and evaluating machine learning prediction models. Available financial metrics for private companies are often limited to just one or two key indicators like:

Revenue

Enterprise
Value

Employee
Count

Basic Profit
Metrics
(e.g. EBITDA)

A Customer-Centric Framework for Model Evaluation

To address the data constraints, we developed a systematic approach to test models across multiple combinations of limited metrics, mimicking real world customer PE data constraints.

Testing and Validation

We use 80% of the carbon data for training and 20% to independently validate the results (hold-out set). We analyse model performance for the validation set across several dimensions:

- How does performance vary with the number of available metrics?
- Which combinations of metrics provide the most reliable predictions?
- How consistent is model performance across different data scenarios?

Our evaluation criteria were split into four categories - accuracy, predictability, feature selection and robustness.

Accuracy

Typically, models are assessed for accuracy using Absolute Percentage Error (APE) for a given company (i):

$$APE_i = \left| \frac{y_i - \hat{y}_i}{y_i} \right| \times 100$$

Where: y_i is the actual value
 $\hat{y}_i$ is the predicted value

Mean Absolute Percentage Error (MAPE), calculated as the mean of APE across all data is often used to quantify accuracy across larger datasets. However, this can give a misleading representation of the typical errors if the distribution is skewed and sensitive to outliers.

We found MAPE to be heavily biased given the significant outliers within the reported carbon dataset. Instead, we found the Median Absolute Percentage Error (MdAPE) gave a more robust measure of the overall performance of the model at the company level.

Predictability

To assess how well a given model and feature set explains the variance in the data we use R-squared. This statistical measure indicates the 'goodness of fit' of the model, ranging from 0 to 1, where 0 indicates that the model does not explain any variance in the dependent variable, and 1 indicates that the model perfectly explains the variance.

R-squared provides a straightforward interpretation of model performance and easy comparison between models. However, it assumes linearity between features and does not provide information on whether the model's features are important or the stability of the model.

Robustness

Robustness is the ability of a model to maintain accuracy and performance when faced with variations, noise, or unexpected inputs that were not present during training. This is important for several reasons:

Handling Real World Data Variability

In real world applications, a robust model can generalise well to new, unseen data and maintain performance across a range of inputs.

Prevention of Overfitting

Overfitting occurs when a model learns the noise or specific details of the training data rather than the underlying patterns. A non-robust model may perform well on the training data but fail on new data due to overfitting.

Trust and Reliability

Robust models provide more consistent and reliable outputs, even in less-than-ideal conditions, which builds the user's trust, which is important for investment decision-making processes.

Noise Injection Test (MdAPE*)

In order to evaluate a model's ability to handle variations in data and detect potential overfitting issues, we introduce a test which adds 30% noise into our feature datasets. This approach allows for sensitivity analysis.

If the model maintains stable accuracy despite the added noise, it suggests that it can effectively generalise to new datasets. On the other hand, if the added noise decreases accuracy substantially, it likely indicates overfitting to the training data.

Scope 1 Performance

Data Inputs	Model / Approach	R ²	MdAPE (Tonnes)	MDPAE Noisy (pct)	MdAPE (pct)
Revenue	xgboost	0.61	13846	67.2	66.2
Revenue, Employees	linear tree	0.45	18757	70.3	70.0
Revenue, Employees	svr	0.33	16555	70.3	70.2
Revenue	svr	0.32	16285	71.1	71.1
Revenue	linear tree	0.53	15938	73.5	71.2
Total Assets	svr	0.49	17221	74.1	73.2
Revenue	linear forest	0.45	16652	74.4	73.2
Total Assets	linear tree	0.54	17551	74.4	73.3
Revenue	ols	0.43	17995	77.9	78.4
Enterprise Value	svr	0.32	19367	78.4	79.4
Enterprise Value	ols	0.29	21395	86.1	86.3

Revenue Models

Revenue-based models showed strong overall performance, with XGBoost achieving the highest R² of 0.613 and lowest MdAPE of 66.2%. The other revenue models (Linear Tree, Linear Forest, SVR, and OLS) were found with R² values ranging from 0.317 to 0.526.

Revenue + Employee Model

Adding employee data to revenue provided only moderate predictive power, with the Linear Tree approach (R² = 0.453, MdAPE = 69.95%). This combination performed better than some revenue-only models but couldn't match revenue with XGBoost.

Total Assets Model

Models using 'total assets' demonstrated solid mid-range performance, with Linear Tree (R² = 0.539) and SVR (R² = 0.488) both achieving respectable results. Despite decent explanatory power, these models had higher error rates (MdAPE 73.2-73.3%) compared to the best revenue-based approaches.

Enterprise Value Model

Enterprise value proved to be the least effective predictor, with both SVR (R² = 0.318, MdAPE = 79.39%) and OLS (R² = 0.288, MdAPE = 86.34%) showing the weakest performance. These models had the lowest R² values and highest error rates across all input types, making them the least reliable options.

Practical Implications

These results carry important practical implications for carbon emissions modelling for private markets, given the limited data availability. Often only one financial metric is available for privately held investments, and our modelling suggest this provides good predictive power to estimate emissions.

Total assets can be a viable alternative when revenue data are unavailable, but enterprise value should be used only when other metrics cannot be obtained.

In all tested combinations, XGBoost and Linear Tree algorithms demonstrate the best overall performances, while most models show impressive robustness to noise.

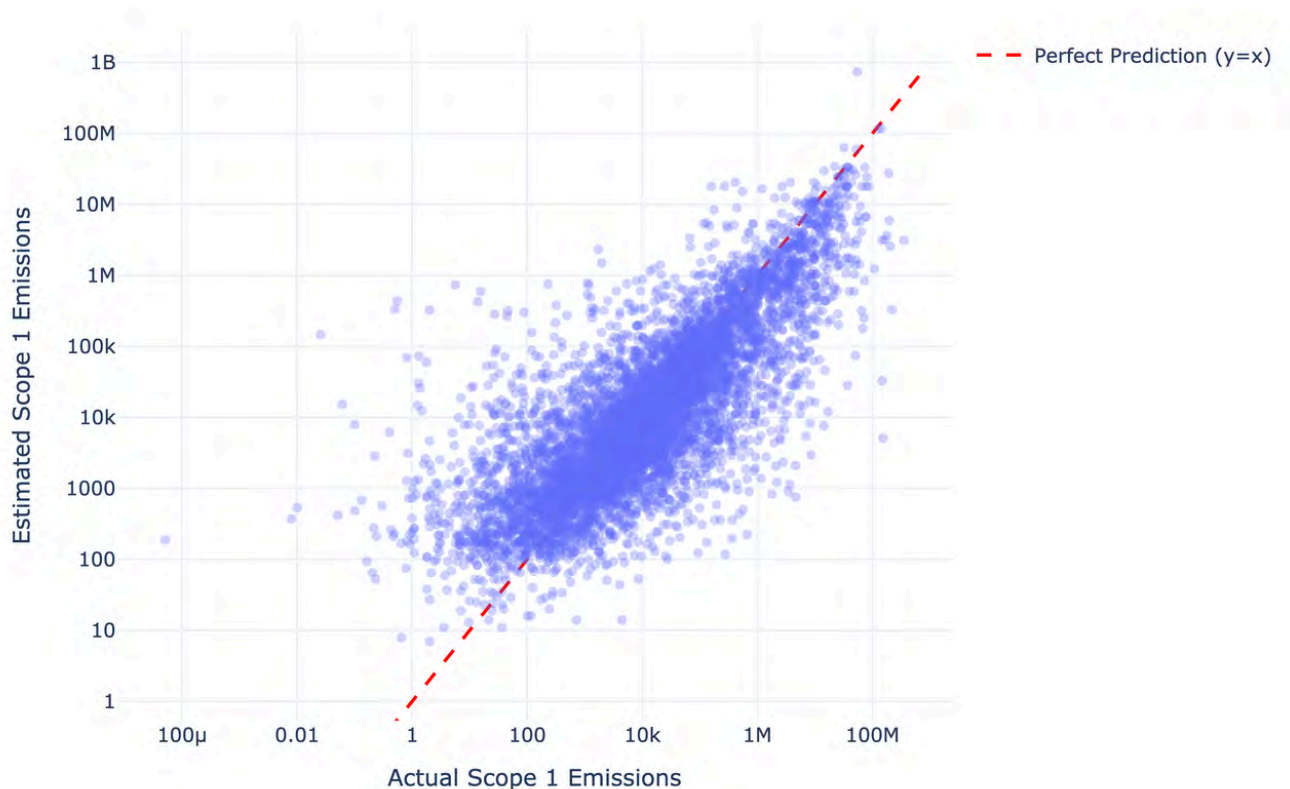


Chart: Comparison of Actual vs Estimated Scope 1 Emissions (2023)

Scope 2 Performance

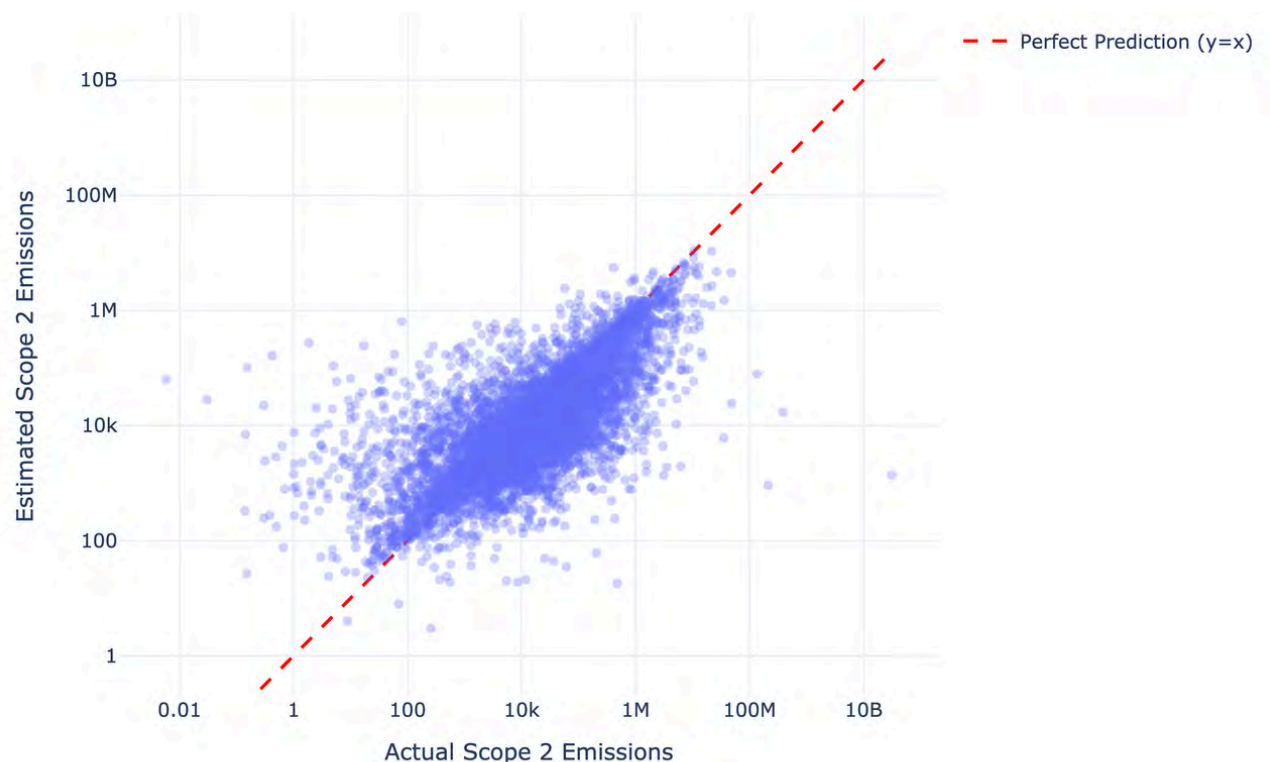
Data Inputs	Model / Approach	R2	MdAPE (Tonnes)	MDPAE Noisy (pct)	MdAPE (pct)
Revenue, Employees	svr	0.28	20424	67.5	66.2
Revenue, Employees	linear tree	0.43	19194	67.9	67.7
Revenue	linear tree	0.29	19530	67.9	67.3
Revenue	svr	0.20	19615	68.5	67.2
Total Assets	svr	0.35	20337	69.8	68.9
Revenue	ols	0.25	19525	71.6	70.3
Revenue	linear forest	0.41	19695	71.8	70.6
Total Assets	linear tree	0.26	19272	73.0	72.7
Enterprise Value	svr	0.29	22842	75.2	75.2
Enterprise Value	ols	0.08	23401	82.6	82.7

Scope 2 emission models show weaker predictive power than Scope 1 models, with R^2 values reaching only 0.425 (Linear Tree with revenue and employees). Despite the weaker explanatory power, Scope 2 models achieve comparable or slightly better percentage error rates, with the best model reaching 66.2% MdAPE compared to 66.2% for Scope 1. However, Scope 1 models consistently outperform in terms of absolute error in tonnes, with the best model achieving 13,846 tonnes MdAPE versus 19,194 tonnes for Scope 2.

The weaker predictive power for Scope 2 likely reflects the step changes that can occur from companies switching to renewable electricity, making it more difficult for the models to find generalised features in the reported dataset.

Scope 2 emissions may be more challenging to predict in absolute terms. However, the relative accuracy of predictions is similar across both emission types. Revenue-based features emerge as important predictors for both emission scopes, with XGBoost showing particular strength for Scope 1 predictions.

Comparison of Actual vs Estimated Scope 2 Emissions (2023)



Performance by Market Size

Market Size	Scope 1 Reported	Scope 1 Estimate	Scope 2 Reported	Scope 2 Estimate
Small Cap (<\$1B)	1,272	328	2,048	437
Mid Cap (\$1-10B)	20,547	13,203	19,191	17,085
Large Cap (>\$10B)	198,120	154,868	122,874	112,209

Private equity holdings are typically smaller than public equities, so we have compared the median emissions estimates with actual emissions across the different company sizes. Our method typically under-estimates 'small cap' company emissions by at least 50% for Scope 1 and Scope 2. However, with median emissions of ~1,200-2,000 tonnes, small caps are immaterial at the portfolio level.

Our estimated emissions compare well with reported emissions for Mid-sized and large-cap private companies, being within 9-36%. This is important for investors given their much higher absolute materiality of emissions (10x-70x) in comparison to small caps.

Comparison with Traditional Emission Factor Approaches

In private markets, spend-based emission factors are commonly used to estimate emissions for companies (with all the limitations previously noted).

Taking an example of the United States steel manufacturing sector, let's assume the model suggests a sector revenue of \$100 billion which produces 500 million tonnes of CO₂e. The resulting EEIO emission factor is therefore 5 tonnes CO₂e per \$1000 revenue. To estimate emissions for a specific steel company (Company A) you multiply their revenue by the factor (e.g. \$2 billion x 5 tonnes/\$1000 = 10 million tonnes CO₂e).

Let's say Company B is in Australia and has the same revenue as Company A, but uses electric arc furnaces for more efficient production. Using EEIO both companies will have the same emissions estimate.

On the other hand, the Emmi ML method uses actual emissions patterns from reported data. It learns that emissions may not always increase at the same rate as revenue within and across sectors and countries. For instance, in the steel manufacturing industry, while EEIO emission factor may assign a universal ratio of 5 tonnes per \$1000, Emmi ML might learn that emissions per revenue increase slowly up to \$500 million before increasing at a faster rate between \$500 million and \$2 billion, with diminishing increases beyond \$2 billion. These growth patterns would vary among different countries and sub-classifications within the steel industry.

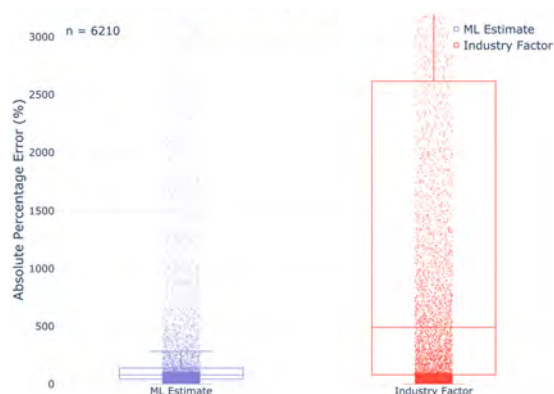
To demonstrate this concept, we present the Scope 1 emissions as a function of revenue for Utilities, which has the highest emissions (around 2.3 billion tonnes) of any industry. We show the reported 'actual' emissions compared to both the Industry Factor Estimate and the Emmi ML estimate (see chart next page).

This graph shows how non-linear the relationship is between revenue and emissions. While the ML estimates accurately follow this non-linearity, the average factor method consistently overestimates the carbon footprint of many companies.

Specifically for companies with revenues ranging from \$100 million to \$1 billion, the industry factor method results in 10-100 times higher emissions than actually reported.

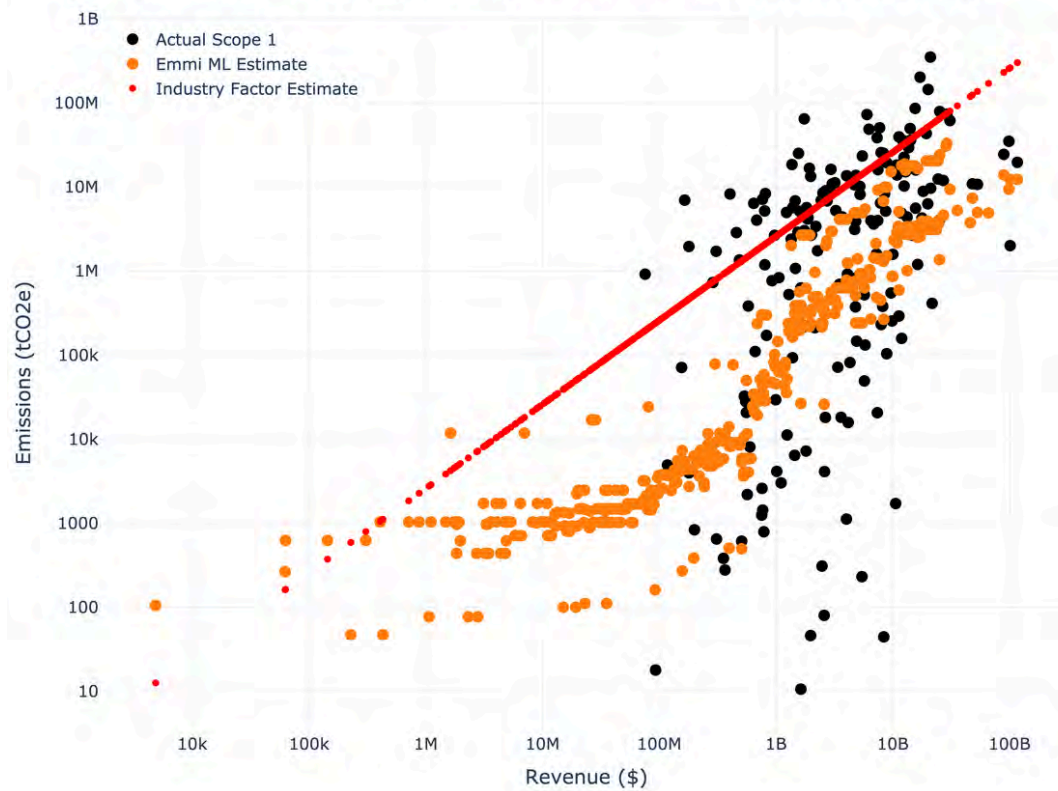
One limitation of the ML method is for companies with less than \$100 million in revenue, as there is not enough training data available to make accurate estimates; therefore, these footprints are assumed to be a constant ~1,000 tonnes. However, at an investor level, these smaller footprints have minimal impact on overall financed emissions calculations.

Absolute Percentage Errors (APE)

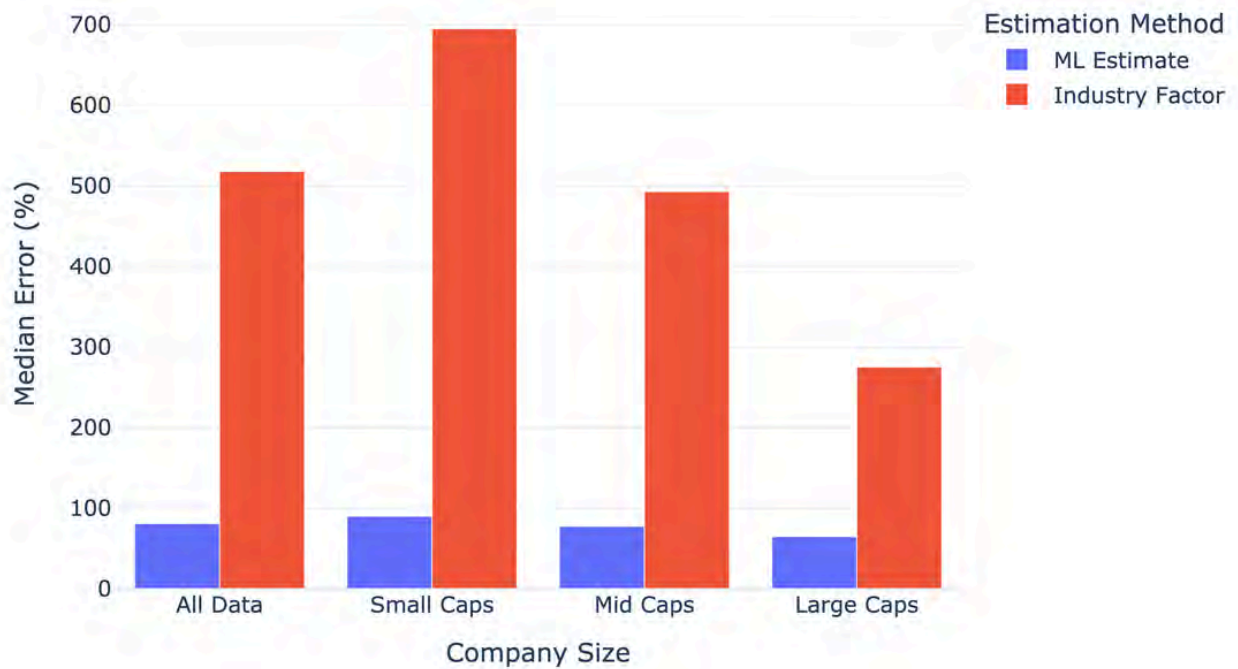


When comparing methods across the reported universe, we find the Emmi ML approach outperforms industry factors by nearly seven times, with an MDAPE of ~70% for the ML compared to ~480% for industry factors (see chart). We demonstrate that Emmi ML models have the ability to use reported data more intelligently, learning the complex relationships between classification, geography, scale, and emissions.

ML vs Factor Approach: Utilities



ML vs Factor Approach: Errors based on company size





Conclusion

Emmi's Machine Learning models provide a vastly more accurate and flexible approach to estimating private company emissions for investors in comparison to traditional industry factor methods. We demonstrate up to seven times the accuracy when using the same data and sparse set of input features.

This performance enhancement lies in the Emmi ML model's ability to use reported data more intelligently, learning the complex relationships between classification, geography, scale, and emissions.



About Emmi

Emmi provides financed emissions data and climate risk analysis across all major public and private asset classes. These support climate-related reporting, and analysis that feeds into investment management processes.

Using a combination of reported emissions and proprietary machine-learning models, Emmi translates emissions into financial implications, based on climate and pricing scenarios. This gives our clients actionable insights about their carbon exposure.

This diagnostics 'toolkit' is backed by our team of climate and finance experts.

Emmi believes that a low carbon economy is possible, and that properly incentivising and mobilising capital is the fastest and most cost-effective way to reach Net Zero and beyond.

Incorporating the cost of carbon into every decision will enable the finance sector, and its customers, to efficiently allocate resources towards this goal, which will accelerate decarbonisation.

To achieve this, and to meet regulatory requirements, there is a need for a broad spectrum of quality carbon emissions data and climate risk analysis. We have built Emmi to solve that problem.



EMMI

For more information

info@emmi.io