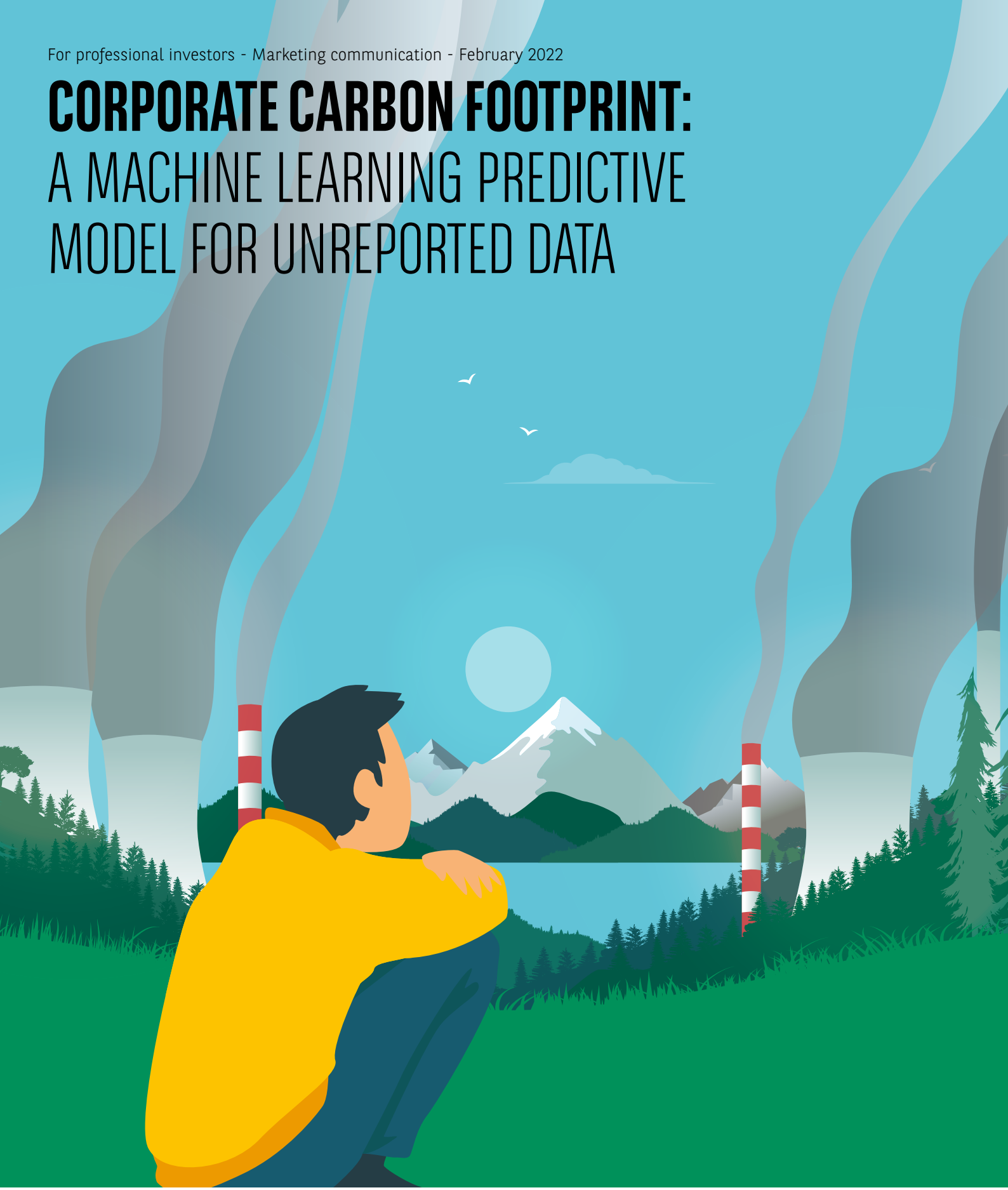
For professional investors - Marketing communication - February 2022

CORPORATE CARBON FOOTPRINT: A MACHINE LEARNING PREDICTIVE MODEL FOR UNREPORTED DATA



BNP PARIBAS ASSET MANAGEMENT

The sustainable investor for a changing world

Thibaut Heurtebize

is a senior quantitative analyst in the Quant Research Group at BNP Paribas Asset Management
14 rue Bergère, 75009 Paris, France
thibaut.heurtebize@bnpparibas.com,
+33 (0) 1 58 97 78 71

Frederic Chen

is a data scientist at BNP Paribas Corporate and Institutional Banking
20 Boulevard des Italiens,
75009 Paris, France
frederic.chen@bnpparibas.com,
+33 (0) 1 42 98 76 63

François Soupé

is co-head of the Quant Research Group at BNP Paribas Asset Management
14 rue Bergère, 75009 Paris, France
francois.soupe@bnpparibas.com,
+33 (0)1 58 97 21 96

Raul Leote de Carvalho

is co-head of the Quant Research Group at BNP Paribas Asset Management
14 rue Bergère, 75009 Paris, France
raul.leotedecarvalho@bnpparibas.com,
+33 (0)1 58 97 21 83

ABSTRACT

We propose a model based on statistical learning techniques to predict unreported corporate greenhouse gas emissions, which generates considerably better results than existing approaches. The model uses one linear and one non-linear learners only, which reduces its complexity to the minimum required. An iterative approach to detecting and correcting data significantly improves the model predictions. Unlike mainstream approaches, which tend to construct one model for each industry, we propose to construct one single global model that uses industries as a factor. This addresses the problem of lack of breadth or lack of reported data in some sectors and generates practical results even for industries where other approaches have failed. We show results for scope 1 and scope 2. Adapting the framework to scope 3 corporate emissions will be the focus of a future publication.

KEY FINDINGS

- **We address the question of how to predict carbon emissions for companies that have yet to report theirs, by proposing a framework based on statistical learning techniques which predicts scope 1 and 2 corporate carbon emissions more accurately than others in the literature**
- **The model shows high in-sample accuracy for all industries at GICS 2 level, which is explained by us choosing to model corporate emissions across the entire global universe while using industries as a variable**
- **The use of an iterative approach to detecting and correcting errors in the data consumed by the model also plays an important role in explaining its greater accuracy**

KEYBORDS

Climate change, corporate carbon footprints, machine learning, carbon emissions, greenhouse gas emissions, Scope 1, Scope 2

JEL codes: G17, G18, Q51, Q52, Q54

The term 'footprint' was introduced in the field of ecology by Wackernagel & Rees (1996) to measure how fast humans consume resources and generate waste compared to how fast nature can absorb our waste and generate resources. In turn, Wiedmann and Minx (2007) defined 'carbon footprint' as a measure of the exclusive total amount of carbon dioxide emissions (CO_2) that is directly and indirectly caused by an activity or is accumulated over the life stages of a product. Today, the carbon footprint tends to account for all greenhouse gas (GHG) emissions caused by an individual, event, organization, service, place or product, and is expressed in units of carbon dioxide equivalent (CO_2e).

The global warming potential (GWP) is the heat absorbed by any GHG in the atmosphere and is defined as a multiple of the heat that would be absorbed by the same mass of CO_2. GWP is 1 for CO_2 while for other gases it depends on the gas and the time frame. For any gas, CO_2e is calculated from GWP as the mass of CO_2 which would warm the earth as much as the mass of that gas. Thus, it provides a common scale for measuring the climate effects of different gases. It is calculated as GWP times the mass of the other gas. In this way, for a gas with a GWP of 10, two tons of it would have CO_2e of 20 tons. Additional GHG include methane (CH_4), nitrous oxide (N_2O), and many fluorinated gases.

GHG footprint assessments are being used in an increasing number of countries, typically within the framework of the United Nations Framework Convention on Climate Change (UNFCCC) whose supreme body, the UN Climate Change Conference, meets annually at the World Conferences of the Parties (COP). These conferences provide the opportunity to review the objectives of the concerted effort to fight climate change. In the context of the regulation of GHG emissions and their reduction, listed and unlisted companies are increasingly reporting their emissions in their extra financial communication. For companies, the assessment of their GHG footprint can be useful not only for disclosure, but also to implement strategies designed to mitigate and reduce their emissions. In addition, GHG emissions assessments can be used as a predictor of a company's vulnerability to transition risk, both in absolute terms and on a relative basis when comparing the emissions of a given company with those of their industry peers.

GHG footprint assessments maybe be voluntary or mandatory depending on location and according to the defined nomenclatures, for example the type of activity and the size of companies. In countries where the reporting of GHG emissions is mandatory at least for certain companies, it is also likely that the calculation methodology is defined along with the regulation. In practice, the heterogeneity of these obligations can sometimes make comparisons among companies in different countries difficult. It can also create geographical biases. Moreover, not only may calculation methodologies vary, they are often accompanied with little explanation, which makes comparisons less straightforward.

In practice, conducting a GHG assessment requires the definition of emission factors, which relate the quantity of a GHG released to the atmosphere with an activity associated with their release. This allows companies to sum their GHG emissions using a single approach. Such emission factors need to be updated periodically. The most widely used are defined in the GHG Protocol, first published in 2001 (Ranganathan et al. (2015)), or the carbon-balance tool used in France. Large companies now report their GHG emissions according to the GHG Protocol by the World Business Council for Sustainable Development (WBCSD) and the World Resources Institute (WRI), or, in some case, according to ISO 14064 standards. The GHG Protocol has become the most widely used methodology in the world when it comes to assessing GHG emissions. The carbon inventory is divided into three scopes corresponding to direct and indirect emissions:

- **Scope 1 (mandatory):** Sum of direct GHG emissions from sources that are owned or controlled by the company, which include stationary combustion, e.g. burning oil, gas, coal and others in boilers or furnaces; mobile combustion, e.g. from fuel-burning cars, vans or trucks owned or controlled by the firm; process emissions, e.g. from chemical production in owned or controlled process equipment such as the emissions of CO_2 during cement manufacturing; and fugitive emissions from leaks of GHG gases, e.g. from refrigeration or air conditioning units

- **Scope 2 (mandatory):** Sum of indirect GHG emissions associated with the generation of purchased electricity, steam, heat or cooling consumed by the company

- **Scope 3 (optional):** Sum of all other indirect emissions that occur in the value chain of the company, including upstream emissions from purchased goods and services, capital goods, fuel and energy-related activities, transportation and distribution, waste generated in operations, business travel or employee commuting; and downstream emissions from leased assets, processing of sold products, use of sold products, end of life treatment of sold products, franchises or investments.

According to the GHG Protocol, reporting on scope 1 and scope 2 is mandatory while reporting on scope 3 is optional. Unfortunately, scope 3, also referred to as value chain emissions, is often the largest component of companies' total GHG emissions.

In practice, if the methods for calculating emissions in a given industry converge then it becomes easier not only to model but also to compare the emissions of each company with those of its peers. Having emissions disclosed by independent bodies such as the Carbon Disclosure Project (CDP), a not-for-profit charity that runs the global disclosure system, or audited by external auditors in extra financial reports is a way of increasing convergence.

Overall GHG emissions from large firms in developed countries follow a common methodology for calculating of scope 1 and 2 emissions: results are either published, or validated, or both, by independent bodies such as external auditors, the CDP, or both. In 2018, this was the case for about 2,800 companies worldwide. Scope 3 was reported by about 1,900 companies with different levels of detail in the methodology employed. This means that about 12,000 companies were not reporting their GHG emissions. This breadth of reporting is insufficient for the increasing number of investors who either want or are required (for example, by regulators), to take into account the GHG emissions of companies in their investment decisions. That is why it is so important to develop methods to predict the unreported emissions of companies as accurately as possible. The current demand for such data is huge.

To date, most academic studies focus on scopes 1 and 2 only, for which the calculation methodologies are more standardized and for which more reported data is available. Nevertheless, even for scopes 1 and 2, available reported data is typically based on voluntary reporting based on the CDP or on extra financial reports from companies. With a few exceptions, e.g. France (Art 173 FIR 2016), GHG emissions reporting is not yet mandatory in most parts of the world despite the efforts from the Task Force on Climate related Financial Reporting (TCFD).

The most accurate models for the estimation of GHG emissions link the industrial processes of each business model with the carbon emissions associated with each stage of those processes. The Environmental Input Output Analysis (EIO), Process Analysis (PA) models give precise results for a given industrial process (Wiedmann, 2009). However, neither the information required to quantify companies' use of those processes, nor their intensity in the overall annual production chain, is publicly available. This makes it difficult to apply such models for calculating company emissions at a global level.

In turn, specialized data vendors rely on relatively simple models to predict the likely GHG emissions of at least some of the companies that do not currently report. These predictions are usually sector level extrapolations based on variables such as the number of employees and income generated, or both. The models used can be as simple as taking sector averages, or using regression models constructed from the existing reported GHG emissions data from peer companies. The number of regressors is usually limited, as are the sample sizes. Model validation tends to rely on the quality of the regression in samples where data is available. Scope 3 data is significantly more difficult to extrapolate, so attempting to use such simplistic approaches tends to lead to poor in-sample results.

Data providers such as MSCI ESG CarbonMetrics (Shakdwipee and Lee (2016); Andersson, Bolton and Samama (2016); de Jong and Nguyen (2016)), Refinitiv ESG Carbon Data – previously known as Thomson Reuters ESG – (Refinitiv (2017), BNP Paribas (2016); Boermans and Galema (2017)) and S&P Global Trucost/CDP use models to predict the GHG emissions of companies that fail to publish emissions data. Such models rely on rules of proportionality between emissions and the size of the company's operations. They tend to use historical data available for the industry as a basis for the calculation, and focus on predicting the logarithm of GHG emissions. Occasionally, they also use the company's energy consumption and production. When applied to the reported data, these models produce results that can be compared with the actual reported data for scope 1 and scope 2, with the R^2 reaching 60% for some samples. However, these are not the most effective modelling approaches.

Goldhammer, Busse and Busch (2017), Griffin, Lont and Sun (2017) and the CDP Statistical Framework (2020) proposed the use of Ordinary Least Squares (OLS) and Gamma Generalized Linear Regression (GGLR) with a broader dataset of publicly available company data for the construction of models for company GHG emissions. Such models go beyond using just simple factors such as revenues and number of employees and tend to be applied to restricted universes of companies. They tend to generate a better fit when applied to the sample of reported data used to construct them, with their R^2 exceeding 80% for some samples.

More recently, Nguyen, Diaz-Rainez and Kuruppuarachchi (2021) proposed the use of statistical learning techniques to develop models for predicting corporate GHG emissions from publicly available data. The machine learning approach proposed is a meta-learner that relies on the optimal set of predictors combining a collection of eight base-learners, namely: OLS, ridge, LASSO, elastic net, multilayer perceptron neural net, K-nearest neighbors, random forest, extreme gradient boosting, as base learners. Their approach generates considerably more accurate predictions than previous models even in out-of-sample situations, i.e., when used to predict reported emissions that were not used to construct the model. Nevertheless, the strongest predictive efficacy of the model was found for predicting aggregate direct and indirect emission scopes as opposed to predicting each separately. Furthermore, despite the improvement over existing approaches, the authors also noted that relatively high prediction errors were still found, even in their best model. Indeed, if we consider the five dirtiest industries with about 90% of total scope 1 emissions (Utilities, Materials, Energy, Transportation, Capital Goods), their average R^2 for these industries is only 51%. If we consider the five dirtiest industries with about 70% of the total emissions in terms of scope 2 (Materials, Energy, Utilities, Capital Goods, Automobiles & Components), their average R^2 for these is only 52%. Moreover, their model fails for Insurance, both for scope 1 and scope 2, with R^2 of -378% and -151%, respectively.

Understanding the risks and opportunities arising from the GHG emissions of companies requires good data. Similarly, if regulators want regulations designed to help with the fight against climate change to have an impact, then it is clear that good emissions data is needed. For as long as the reporting and auditing of company emissions is not compulsory, the only viable alternative is to predict non-reported company emissions relying on models capable of generating the most accurate predictions possible. From the above, we believe that the current state-of-the-art does not yet provide good enough models for the task at hand. In our view, the quality of data made available by the specialized data vendors is not yet sufficient. Understanding the reasons behind this problem and being able to propose alternative approaches that can lead to better models and more accurate predictions of unreported data is thus of great importance.

The recently proposed approach by Nguyen, Diaz-Rainez and Kuruppuarachchi (2021), based on statistical learning offers a promising starting point. However, the central challenge with such statistical learning approaches is to strike the right balance between increasing the model's complexity and accuracy while limiting the risk of overfitting.

In this paper, we propose a statistical learning model to predict unreported scope 1 and 2 company emissions in an investment universe of about 15,000 companies of which only about 3,500 companies actually report. This model is inspired by the work of Nguyen, Diaz-Rainez and Kuruppuarachchi (2021), but we have adapted it to

i) increase its robustness, aiming at higher accuracy
ii) increase its transparency, making sure that replication of results is possible, and
iii) increase its flexibility, making sure the model can be adapted to account for future changes in regulations related to the reporting of GHG emissions.

Additionally, the model should be systematic, at least when applied to corrected data. To succeed in this, we made some significant choices in departing from existing approaches. One key decision was to create one single model for the entire universe while using each industry as a factor instead of creating one model per industry. This helps us address the problem of lack of breadth of companies or reported data in some industries. The second important decision was the choice to keep model complexity to a minimum by relying only on two machine learning approaches, one linear and one non-linear. The final important decision was to correct reported emissions data when required. Indeed, reported data is not always correct. The use of incorrect emissions data can reduce significantly the accuracy of models. It is important to correct or remove data that is clearly mis-reported. Unfortunately, while this step is of extreme importance, it is also time consuming. We chose to investigate the output of the model iteratively and investigate the reported data that failed to fit the model at each iteration, correcting data if it was necessary to do so after consulting the respective company reports.

In the reminder of the paper, we shall describe the framework and the data retained for the model in the next section. This will be followed by a section in which we present the results of the model when applied to predicting GHG scope 1 and 2 emissions. We shall also show how the proposed framework increases significantly the ability to predict reported data when compared to the approach proposed by Nguyen, Diaz-Rainez and Kuruppuarachchi (2021), in particular for industries for which their approach fails, e.g., Insurance.

FRAMEWORK

GHG EMISSIONS DATA

At present, most of the company GHG emissions data provided by data vendors is based on their approaches to predicting unreported data. We considered a broad global investment universe of 15,726 companies, which includes small, mid and large capitalization public companies and private companies. This is the typical investment universe considered by asset managers when investing for their clients. In 2018, less than 20% of the companies in this universe reported their scope 1 and scope 2 emissions. If we consider the data from Bloomberg, a data vendor, then only about 22% of the 11,700 companies in their database actually reported scope 1 and 2 emissions data; all other available data was predicted. Even constraining the universe to the 3,000 largest capitalization companies, we still find only 46% of reported data at best.

Exhibit 1: Companies reporting GHG emissions (Dec-2018)

	Broad Global Universe		Global Universe (Bloomberg)		Large Market Cap Universe	
COMPANIES	NUMBER	PERCENTAGE	NUMBER	PERCENTAGE	NUMBER	PERCENTAGE
Total	15.726	100%	11.700	100%	3.000	100%
Reporting scope 1	2.836	18%	2.622	22%	1.385	46%
Reporting scope 2	2.737	17%	2.555	22%	1.343	45%
Reporting scope 3	1.910	12%	1.661	14%	996	33%

In all these cases, scope 3 was clearly the least reported, and by a large margin. Predictions of scope 3 are also often not available. This is because of the difficulty of estimating and predicting scope 3, which is itself a sum of emissions in the upstream and downstream supply chain of the company, as mentioned previously. It is also the reason why we shall address the reporting and prediction of scope 3 in a future publication.

MODEL

To predict unreported scope 1 and scope 2 company emissions, we propose a framework designed to be useful for investors, for the asset management industry and for carbon offsetting. For this reason, we required the framework to be:

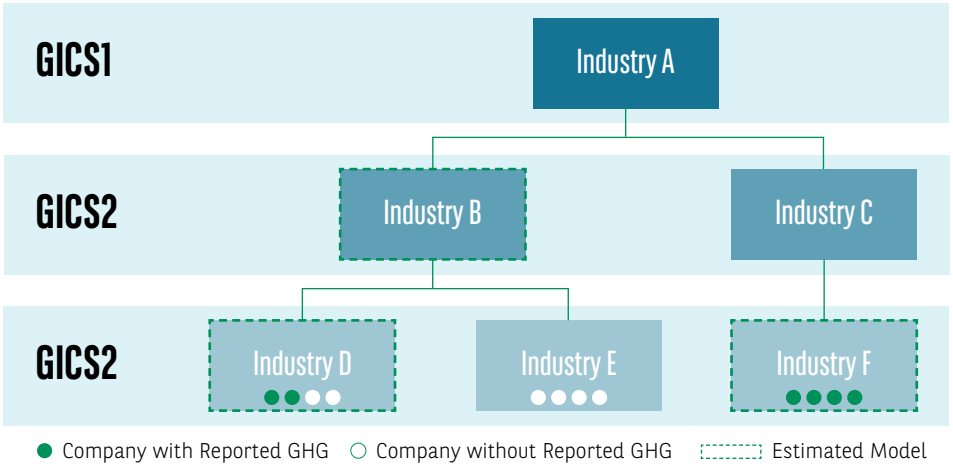
- **Robust:** The model predictions for the reported data should be as accurate as possible, or at least conservative
- **Transparent:** The model should use only publicly available data and the model complexity should be limited to what is strictly necessary by using a limited number of underlying learners
- **Flexible:** The model should be easy to adapt to future changes in regulations covering the reporting of company GHG emissions.

These constraints are important when it comes to producing reliable predictions of company GHG emissions and to industrializing the calculation of emissions for a large universe of companies while making sure that predictions are accurate enough for the purpose of taking informed investment decisions or for carbon offsetting. When it comes to flexibility, it is important to note that under certain regulatory environments and for certain applications, e.g., under the AMF (Autorité des Marchés Financiers) in France when it comes to carbon offsetting (AMF (2020)), such a model should either be accurate enough or at least conservative when it comes to its predictions.

MODEL ARCHITECTURE

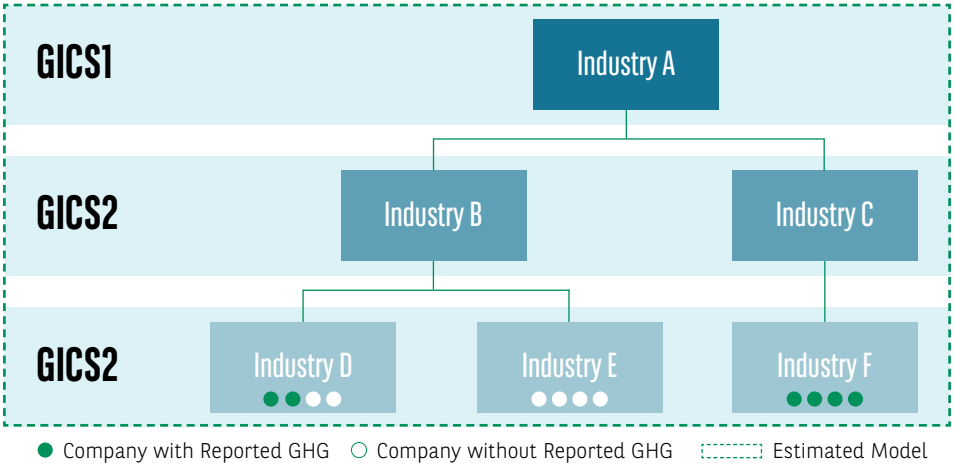
There are two main approaches to modelling company GHG emissions. In the most common approach, companies are grouped into industries reflecting similarities in their business exposures. Individual models potentially relying on different predictors are then created for each industry. This approach relies on the idea that the predictors of emissions for companies in different industries can be different, and that it thus makes sense to have independent models. The model created for each industry is then used to predict the emissions of non-reporting companies in that same industry. However, this approach reaches its limits for industries with a relatively small number of companies or with too few companies reporting emissions.

Exhibit 2: Modelling each industry separately



A second approach focusses on creating one single model that learns the patterns of emissions data using a much larger set, e.g., using all companies at a global level, for all countries and industries. Such models can use industries as a predictor and include non-linear components that may allow for different predictors in different industries when that increases efficacy. We opted for this approach, which stands a better chance of predicting emissions in industries with a smaller number of companies or with not many companies reporting emissions.

Exhibit 3: Global model relying on all industries



MODEL PREDICTORS

During the preliminary phase of the model construction, we tested a large number of possible predictors. However, we found that models with a relatively small number of predictors are accurate enough, in particular as errors in data are corrected. We thus opted for parsimonious models.

As shown in Exhibit 4, the final model for scope 1 and 2 relies on regional (or country) predictors, where regional information provides insights into the local operating environment of the company. The model also relies on company-specific predictors such as company financial data, in particular for getting a sense of its size, its assets size and age and its profitability, as reported in financial statements.

Exhibit 4: Data categories and sources

CATEGORIES	Sources	Description
Regional information	World Bank and International Energy Agency	Country information: Region, revenue group, CO_2 tax regulations, CO_2 emissions, carbon intensity of energy mix, CO_2 emissions per GDP, CO_2 emissions per GDP (purchasing power parity)
Financial metrics	FactSet, Refinitiv and Worldscope	Financial metrics of each company: Revenues, number of employees, total assets, gross property plant and equipment, capex, age of assets
Industry Classification	Bloomberg	GICS 1-2-3-4 sector and industry definitions and minor adjustments based on a more sustainability-oriented classification, in particular for the utilities and energy sectors
Energy indicators	Bloomberg	Production and energy consumption of the company

MODEL TRAINING

To train the model, we use only reported emissions data. The universe of companies was split into an in-sample group, with companies that reported GHG emissions data; an out-of-sample group, with companies for which we have predicted emissions provided by data vendors; and a group of companies that do not report emissions and for which there are no predictions.

The in-sample group with companies reporting emissions was then split into different groups so that the model could be trained using a group of companies and the model performance could be evaluated on another group not used for training.

Finally, companies reporting an increase in emissions of more than 50% from one year to the next without any clear explanation in terms of corporate actions, e.g., mergers and acquisitions, were excluded from the in-sample group and included in the out-of-sample group instead. Between 5% for automobiles & components and 27% for software & services of the companies reporting emissions in each industry were moved to the out-of-sample group for this reason.

MODEL EVALUATION

To assess the efficacy of the model, we need a measure of the performance. We chose to use the most common measure in GHG emissions modelling research, which is the R^2 of the log-transformed emissions defined for each industry as:

y_i = log-transformed of reported emissions of company i in the industry

$\widehat{y_i}$ = log-transformed of predicted emissions of company i in the industry

$\bar{y}$ = average of log-transformed reported emissions of all reporting companies in the industry

$$R^2 = 1 - \frac{\sum(y_i - \widehat{y_i})^2}{\sum(y_i - \bar{y})^2}$$

A k-fold cross-validation method was used to validate the model, evaluating its performance when it comes to predicting correctly the company log-transformed carbon emissions. This is a resampling method commonly used to evaluate predictive machine learning models; it uses different portions of the data to train and test a model in each iteration. Each round of cross-validation involves partitioning the sample of data into k complementary smaller subsets (the data folds), training the model on k-1 of those subsets (the training set), and validating the analysis on the last k subset, used as the testing set. To reduce variability, multiple rounds of cross-validation were performed using different partitions, with the validation results averaged over the rounds to give an estimate of the model's predictive performance

MACHINE LEARNING MODEL

In our framework, we chose to use both linear and non-linear learners. We chose to combine just one linear with one non-linear learner to minimize the model's complexity. The choice of the linear and the non-linear learner was based on their efficacy when compared to alternative approaches. For this assessment, we considered the predictions in-sample and out-of-sample, investigating the efficacy of the approaches in different industries. Our focus was on the best prediction for each scope of emissions for each company rather than at aggregate level, e.g., aggregating scopes, where offsetting of errors is more likely. We also looked at the overall

efficacy of the model in terms of R^2 of the log-transformed predictions against reported emissions in different industries. Finally, we also investigated the extent to which the model is conservative by comparing its predictions with those from data vendors.

In the same way that the model aims at predicting the log-transformed emissions of companies, a number of predictors are also log-transformed, namely revenues, number of employees, total assets, gross property plant and equipment, capex and production and energy consumption of the company. Much like carbon emissions, these variables tend to show a more normal distribution in the cross-section of companies after log-transforming. In turn, industry predictors are assigned 1 for the industry where the company is classified and 0 otherwise.

The Linear Regression Model component could have relied on learners such as OLS, Ridge Regression, Lasso Regression, Elastic Net and others. We chose to use Elastic Net, which is a linear regression model that reduces the number of used predictors in a way similar to Lasso Regressions, while also reducing the size of the less relevant regression coefficients to reduce the risk of overfitting in a way similar to Ridge Regressions. The advantage of this linear component relative to more sophisticated non-linear approaches is that is makes it easier to explain predictions from predictor values. However, this comes at the cost of accuracy, which is limited by imposing a linear dependence on predictors.

The Non-linear Regression Model component could have relied on learners such as Random Forests, Extremely Randomized Trees, Gradient Boosting and others. We chose to use Extremely Randomized Trees (Geurts, Ernst and Wehenkel (2006)), a version of Random Forests that includes an additional layer of randomness when building the underlying Decision Trees: Instead of building a Decision Tree that splits the observations optimally, this approach generates many random splits and takes the best performing split created. The advantage of using a non-linear component is the added flexibility from allowing for complex dependencies between reported emissions and predictors. Overall, the non-linear component tends to produce more accurate predictions than the linear component, although this comes at the cost of making it harder to explain how predictors generate the prediction.

The Basic Combinations of the learner components mentioned above could have relied on a Mean Combination, a Median Combination, a Maximum Prediction Combination or others. We chose to use the Maximum Prediction Combination between the two components described above which, for each company, consists on picking the component that produces the most conservative (higher emissions) prediction.

ITERATIONS FOR DATA CORRECTION

The predictive efficacy of the model can suffer significantly when the data used to construct the model contains errors. For this reason, we opted to implement an iterative approach to detect and correct errors in data, either with reported emissions or with predictors, and whenever errors were confirmed and corrections were found. This can be done by investigating the predictions that fall further from reported data at each iteration. The model can then be re-trained with the corrected data, which tends to significantly improve the model efficacy. Often, data errors result from corporate actions such as mergers and acquisitions that were not properly taken into account. It is also sometimes the case that the data supplied by data vendors is not correct. Corrected data can be found in the company annual reports. We noted that this a more frequent problem for Chinese companies whose reports can be mistranslated. In general, our experience shows that the efficacy of the model and its R^2 can be significantly improved after some iterations to correct data using this procedure.

RESULTS

We now present the results obtained using the in-sample group of companies that report emissions, comparing the model predictions with their reported data. We also present out-of-sample results, comparing the model predictions for companies that are not reporting emissions (or that were excluded from the in-sample group because of large unexplained changes in emissions from one year to the next) with predictions for those same companies provided by two independent data vendors.

In exhibit 5 and 6 we show the R^2 of the model for scope 1 and the model for scope 2 based on reported emissions. In the tables we show separately the R^2 for the elastic net model, for the extreme randomized trees model and for the maximum prediction combination of both. In both cases, scope 1 and scope 2, the R^2 from the extreme randomized trees model tends to be higher than that from the elastic net. The conservative choice of using the one that produces the most conservative prediction of emissions, i.e. the higher prediction, explains why the R^2 tends to follow between the R^2 for the two components of the model.

As a matter of fact, the R^2 measures the model ability to correctly predict the available reported data in sample, but it does not necessarily reflect lower model risk (Cont (2006)). Thus, when using the same model to predict non-reported data, searching for the highest R^2 may not be the most adequate choice for all applications. That is why here we chose to use the most conservative model output (between linear and non-linear components), which results in more conservative predictions when the uncertainty observed in the other competing models from data vendors is relatively high.

ANALYSIS OF PERFORMANCE OF IN-SAMPLE PREDICTIONS

The in-sample analysis of performance for scope 1 model predictions is based on reported emissions for a total of 16,800 data points from 2011 through 2018. For scope 1, more than 95% of emissions come from just five industries (GICS 2): Utilities, Materials, Energy, Transportation, Capital Goods. For companies in those sectors, the model predicts reported emissions with an R^2 of 81% on average (a minimum of 74% for Capital Goods and a maximum of 86% for Transportation). The average R^2 found by Nguyen, Diaz-Rainez and Kuruppuarachchi (2021) for these industries using their modelling approach was only 51%. In fact, for most industries, the R^2 of our proposed model is higher than the R^2 of their model. Moreover, the lowest R^2 from our proposed model is still a strong 57%, for Consumer Durables & Apparel. Finally, unlike their model, which fails for Insurance, our proposed model generates predictions for this industry with an R^2 of 68%.

Exhibit 5: R^2 of the machine learning model for reported scope 1 emissions

GICS 2	Number of unique issuers	R^2			Average scope 1 emissions (tCO_2e)	Share of total scope 1 emission	Cumulative sum of share
		Maximum Prediction Combination	Extremely Randomized Trees	Elastic Net			
Utilities	250	80%	89%	70%	21 122 718	40%	40%
Materials	547	84%	92%	78%	6 771 192	28%	67%
Energy	281	81%	88%	77%	9 046 819	19%	86%
Transportation	168	86%	94%	81%	5 230 486	7%	93%
Capital Goods	498	74%	85%	66%	689 198	3%	95%
Food Beverage & Tobacco	200	79%	89%	71%	727 153	1%	96%
Commercial & Professional Services	99	82%	85%	74%	991 566	1%	97%
Automobiles & Components	134	77%	85%	72%	451 435	0%	98%
Consumer Services	99	86%	93%	83%	523 669	0%	98%
Real Estate	256	65%	82%	55%	178 692	0%	98%
Food & Staples Retailing	57	76%	86%	64%	594 411	0%	99%
Diversified Financials	146	69%	79%	63%	202 358	0%	99%
Pharmaceuticals Biotechnology & Life Sciences	114	83%	90%	80%	215 333	0%	99%
Technology Hardware & Equipment	197	67%	78%	55%	114 656	0%	99%
Semiconductors & Semiconductor Equipment	95	68%	84%	60%	210 589	0%	99%
Household & Personal Products	42	74%	81%	66%	414 386	0%	100%
Consumer Durables & Apparel	143	59%	75%	49%	111 742	0%	100%
Retailing	120	72%	82%	68%	102 160	0%	100%
Telecommunication Services	93	77%	90%	66%	124 939	0%	100%
Banks	207	69%	85%	58%	51 071	0%	100%
Health Care Equipment & Services	89	69%	80%	68%	78 834	0%	100%
Media & Entertainment	74	75%	83%	72%	40 836	0%	100%
Software & Services	93	69%	80%	60%	25 935	0%	100%
Insurance	106	68%	85%	55%	18 320	0%	100%

Source: BNP Paribas Asset Management, World Bank, IEA, FactSet, Refinitiv, Worldscope, Bloomberg. Based on annual data from 2011 through 2018. For illustration purposes only.

The in-sample analysis of performance of scope 2 model predictions is based on reported emissions for a total of 16,497 data points from 2011 through 2018. For scope 2, the five industries (GICS 2) emitting the most - Materials, Energy, Utilities, Capital Goods, Automobiles & Components - contribute about 70% of total scope 2 emissions, with the largest scope 2 emissions coming from the Materials industry. 95% of the total scope 2 emissions are created by 17 industries (GICS 2), which are more evenly spread out across industries than scope 1 emissions.

For companies in these 17 industries, the model predicts reported emissions with an R^2 of 80% on average (a minimum of 65% for Real Estate and a maximum of 89% for Pharmaceuticals Biotechnology & Life Sciences). The average R^2 found by Nguyen, Diaz-Rainez and Kuruppuarachchi (2021) based on their approach applied to these 17 industries was only 51%. Overall, the R^2 are higher than those found with their approach for most industries, and the lowest R^2 from our proposed model is a strong 65% for Real Estate. Moreover, unlike their model, which fails again for Insurance, the R^2 for this industry with our proposed approach is high at 79%.

Exhibit 6: R^2 of the machine learning model for reported scope 2 emissions

GICS 2	Number of unique issuers	R^2			Average scope 2 emissions (tCO_2e)	Share of total scope 2 emission	Cumulative sum of share
		Maximum Prediction Combination	Extremely Randomized Trees	Elastic Net			
Materials	531	76%	86%	61%	1 555 227	35%	35%
Energy	273	78%	86%	62%	1 081 211	13%	48%
Utilities	223	69%	69%	34%	997 460	9%	57%
Capital Goods	497	83%	87%	70%	308 777	7%	64%
Automobiles & Components	131	86%	88%	79%	883 613	5%	69%
Telecommunication Services	92	78%	89%	70%	1 046 732	4%	73%
Technology Hardware & Equipment	201	84%	93%	79%	443 214	4%	77%
Transportation	164	76%	94%	84%	483 537	3%	80%
Food Beverage & Tobacco	194	82%	88%	67%	390 322	3%	83%
Food & Staples Retailing	55	76%	90%	76%	1 227 244	3%	86%
Semiconductors & Semiconductor Equipment	96	77%	86%	67%	481 999	2%	88%
Consumer Services	99	86%	91%	67%	393 041	2%	90%
Retailing	122	88%	93%	78%	303 500	2%	91%
Real Estate	276	65%	92%	85%	120 302	1%	93%
Consumer Durables & Apparel	146	86%	87%	72%	220 952	1%	94%
Pharmaceuticals Biotechnology & Life Sciences	117	89%	84%	55%	265 355	1%	95%
Banks	222	78%	91%	81%	120 058	1%	96%
Household & Personal Products	44	86%	88%	81%	389 151	1%	97%
Software & Services	92	88%	95%	85%	162 513	1%	98%
Health Care Equipment & Services	89	85%	90%	85%	158 188	1%	98%
Media & Entertainment	79	83%	90%	80%	159 570	1%	99%
Diversified Financials	152	80%	83%	70%	66 254	0%	99%
Commercial & Professional Services	98	75%	83%	67%	87 704	0%	100%
Insurance	106	79%	90%	74%	54 924	0%	100%

Source: BNP Paribas Asset Management, World Bank, IEA, FactSet, Refinitiv, Worldscope, Bloomberg. Based on annual data from 2011 through 2018. For illustration purposes only.

COMPARISON OF OUT-OF-SAMPLE PREDICTIONS

The out-sample group is made of companies that do not report emissions and a smaller number of companies that report emissions which appear incorrect. For the analysis of out-of-sample performance, we compared the predictions for 2018 corporate emissions from three different providers: i) from the proposed model (as of end of 2019), ii) from Bloomberg (as of end of 2020) and iii) from S&P Global Trucost (as of end of 2020). In other words, in this out-of-sample analysis, we compared different independent predictions for companies for which the emissions are either not known, or for which the quality of the data reported is poor.

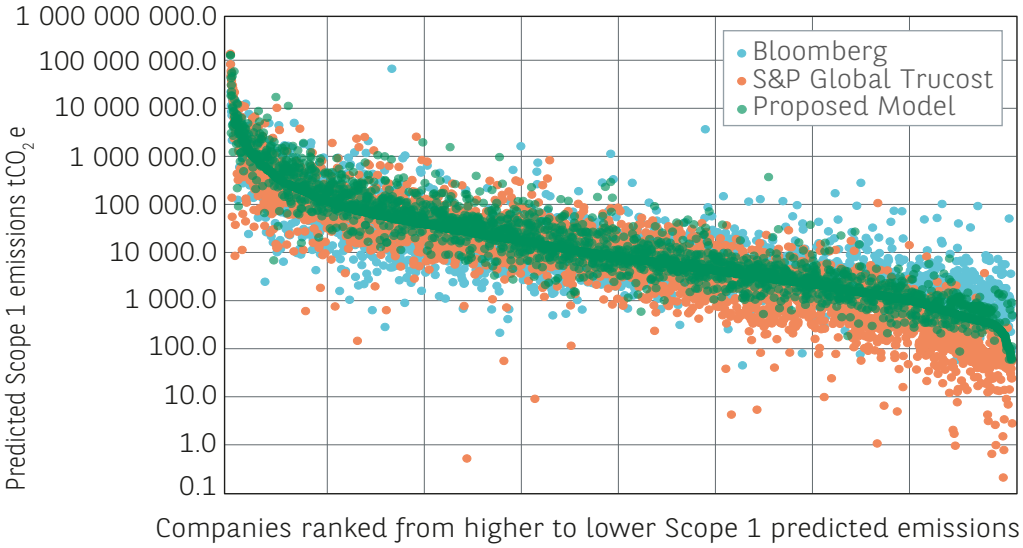
As proposed by Cont (2006), we measure our model uncertainty by comparing it with other models with the same level of uncertainty in their calibration, at the same date. For each company we define best estimate, BE, as the median of the three predictions: That of our model and those from the two data vendors. For each prediction provider and for each company, we define the error of the prediction, $Error_{provider}$, as the absolute difference between each prediction and the best estimate:

$$Error_{provider} = \left| CF_{provider} - BE \right|$$

For each company in the out-of-sample group, the relative error of each provider, $RE_{provider}$, is defined by dividing the error of that provider by the best estimate:

$$RE_{provider} = \frac{Error_{provider}}{BE}$$

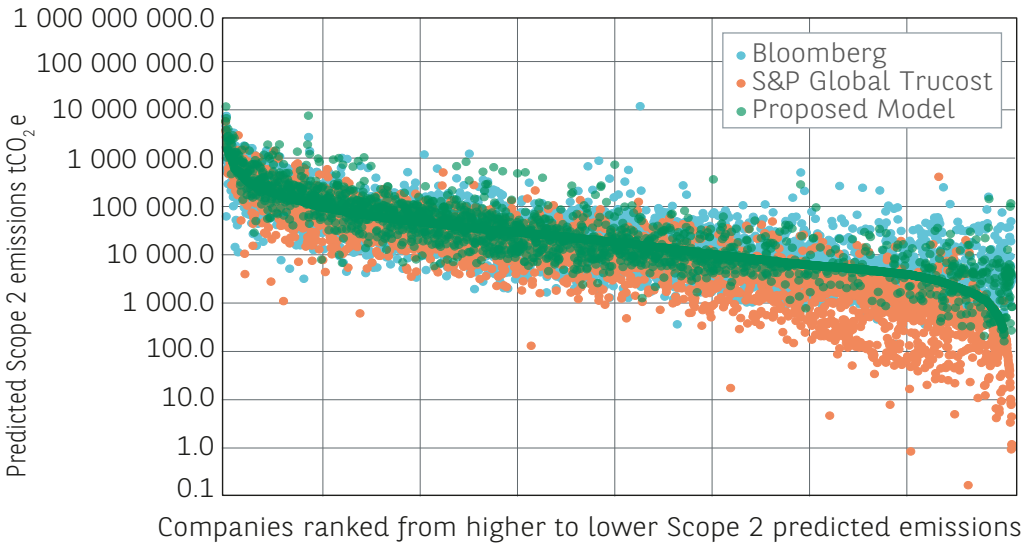
Exhibit 7: Predictions of unreported Scope 1 corporate emissions for 2018



Source: BNP Paribas Asset Management, Bloomberg, S&P Global Trucost. For illustration purposes only.

For scope 1, the out-of-sample universe comprises 7,761 companies for which we had predictions from both Bloomberg and S&P Global Trucost for 2018 (as of end of 2020). In Exhibit 7, we compare the predicted 2018 scope 1 emissions for each of these companies based on our model with those from Bloomberg and S&P Global Trucost. Our model produces the best estimate for 48% of the observations, i.e., the median value of the three. In Exhibit 9, we show that when the proposed model does not produce the best estimate, then the average median error is 63% versus 388% and 125% for the other two models, respectively, when their models do not produce the best estimate.

Exhibit 8: Predictions of unreported Scope 2 corporate emissions for 2018



Source: BNP Paribas Asset Management, Bloomberg, S&P Global Trucost. For illustration purposes only.

For scope 2, the out-of-sample universe comprises 7,806 companies for which we had predictions from Bloomberg and S&P Global Trucost for 2018 (as of end of 2020). In Exhibit 8, we compare the predicted 2018 scope 1 emissions for each of these companies based on our model with those from Bloomberg and S&P Global Trucost. Our model produces the best estimate for 47% of the observations, i.e., the median value of the three. As show in Exhibit 9, when the proposed model does not produce the best estimate, then the average median error is 57% versus 213% and 91% for the other two models, respectively, when their models do not produce the best estimate.

Exhibit 9: comparison of predictions of corporate GHG emissions

	Scope 1			Scope 2		
Prediction providers	Number of estimates	Average error ($ktCO_2e$)	Average relative error	Number of estimates	Average error ($ktCO_2e$)	Average relative error
Proposed Model	3709	126.5	63%	3693	26.6	57%
Bloomberg	2085	183.4	388%	2499	38.7	213%
S&P Global Trucost	1968	217.4	125%	1614	41.5	91%
Total	**7762**			**7806**		

Source: BNP Paribas Asset Management, Bloomberg, S&P Global Trucost. Based on model predictions for corporate emissions in 2018. For illustration purposes only.

CONCLUSION

Corporate carbon footprints are now an unavoidable factor for investment decisions, which needs to be taken into consideration when assessing not only the risk facing a company but also the opportunity they offer as an investment. However, the vast majority of companies do not yet report their GHG emissions. Accurate predictions of their emissions are thus the only option to be able to take each company's carbon footprint into account.

In this paper, we propose a framework based on statistical learning techniques, which generates significantly more accurate predictions of scope 1 and scope 2 corporate emissions. The model proposed uses a single linear method and a single non-linear method, each relying on a relatively small number of pertinent predictors. This parsimonious approach to modelling reduces the complexity of the model to the minimum required and reduces the risk of overfitting. Using an iterative approach to correct data at each iteration also significantly improves the results. We are the first to propose the use of such a process as a way of improving the model accuracy.

Departure from more traditional approaches, which focus on modelling industries independently, by proposing the construction of a single global model also plays an important role in increasing the accuracy of the model. This is particularly the case for industries with just a few companies, or industries with many companies not reporting their emissions. This allows us to generate significantly better predictions for all industries, thanks to the increase in the sample used to construct the model. Our iterative data correction, mentioned earlier, also plays an important role in facilitating the construction of a single model.

The proposed model better replicates the reported data from companies that publish their carbon emissions across all industries (GICS 2). We have gained in robustness as measure by the R^2 of log-transformed emissions for both scope 1 and 2.

Moreover, it is the first study that individually compares the prediction of carbon emissions with three different models for more than 7,800 companies (scopes 1 and 2). We show that our model gives more than 47% of the median estimates between the two competing models. We believe this paper makes an important contribution to the question of how to predict carbon emissions for companies that are not yet reporting them. The approach proposed here leads to more accurate predictions of corporate emissions, despite the fact we chose a conservative approach with preference for the highest predictions. It should thus lead both to better investment decisions and to more relevant decisions when it comes to fighting climate change. In a future publication we shall address the prediction of scope 3 emissions using a model derived from the one presented here.

ACKNOWLEDGEMENTS

We are grateful Quyen Nguyen and Prof. Ivan Diaz-Rainey for the valuable discussion and suggestions as well as Alexander Bernhardt, Jane Ambachtsheer, Pierre Moulin and Thibaud Clisson for their useful comments.

REFERENCES

AMF. 2020.
"Les approches extra-financières dans la gestion collective.
Troisième rapport. " amf-france.org

Andersson, Mats, Patrick Bolton, and Frédéric Samama. 2016.
"Hedging Climate Risk." Financial Analysts Journal Vol. 72, No. 3, pp. 13-32.

BNP Paribas. 2016.
"Stress-Testing Equity Portfolios for Climate Change Factors: The Carbon Factor."
BNP Paribas Securities Services.

Boermans, Martijn, and Rients Galema. 2017.
"Pension Funds' Carbon Footprint and Investment Trade-offs."
DNB Working Papers 554, De Nederlandsche Bank.

CDP. 2020.
"CDP Full GHG Emissions Dataset – Technical Annex III: Statistical Framework."
CDP Disclosure Insight Action.

Cont, Rama. 2006.
"Model Uncertainty and its Impact on the Pricing of Derivative Instruments."
Mathematical Finance Vol. 16, No. 3, pp. 519–547

de Jong, Marielle, and Anne Nguyen. 2016.
"Weathered for Climate Risk: a Bond Investment Proposition." Financial Analysts Journal Vol. 72, No. 3, pp. 34-39.

Goldhammer, Bernhard, Christian Busse, and Timo Busch. 2017.
"Estimating Corporate Carbon Footprints with Externally Available Data."
Journal of Industrial Ecology Vol. 21, No. 5, pp. 1165-1179.

Griffin, Paul A., David Lont, and Estelle Sun. 2017.
"The Relevance to Investors of Greenhouse Gas Emission Disclosures."
Contemporary Accounting Research Vol. 34, No. 2, pp. 1265-1297.

Geurts, Pierre, Damien Ernst, and Louis Wehenkel. 2006.
"Extremely Randomized Trees." Machine Learning Vol. 63, pp. 3–42.

Nguyen, Quyen, Ivan Diaz-Rainey, and Duminda Kuruppuarachchi. 2021.
"Predicting Corporate Carbon Footprints for Climate Finance Risk Analyses: A Machine Learning Approach." Energy Economics Vol. 95, issue 3, 105129.

Ranganathan, Janet, Laurent Corbier, Pankaj Bhatia, Simon Schmitz, Peter Gage, and Kjell Oren.
2015. "The Greenhouse Gas Protocol: a Corporate Accounting and Reporting Standard, Revised Edition." World Business Council for Sustainable Development and World Resources Institute.

Shakdwipee, Manish, and Linda-Eling Lee. 2016.
"Filling The Blanks: Comparing Carbon Estimates Again Disclosures." MSCI ESG Research Issue Brief.

Refinitiv. (2017).
"Refinitiv ESG Carbon Data an Estimate Models." Refinitiv.com/ESG.

Wackernagel, Mathis, and William Rees. 1996.
"Our Ecological Footprint: Reducing Human Impact on The Earth." New Society Publishers.

Wiedmann, Thomas. 2009.
"Editorial: Carbon Footprint and Input-Output - An Introduction." Economic Systems Research, Vol. 21, No. 3, pp. 175-186.

Wiedmann, Thomas, and Jan Minx. 2007.
"A Definition of 'Carbon Footprint'." In Ecological Economics Research Trends, Chapter 1, C. C. Pertsova (ed.). Nova Science Publishers, Inc.

DISCLOSURE STATEMENT

The authors report no conflicts of interest. The authors alone are responsible for the content and writing of the paper, not BNP Paribas Asset Management.

BNP Paribas Asset Management France, "the investment management company," is a simplified joint stock company with its registered office at 1 boulevard Haussmann 75009 Paris, France, RCS Paris 319 378 832, registered with the "Autorité des marchés financiers" under number GP 96002. This material is issued and has been prepared by the investment management company.

This material is produced for information purposes only and does not constitute:

1. an offer to buy nor a solicitation to sell, nor shall it form the basis of or be relied upon in connection with any contract or commitment whatsoever or
2. investment advice.

This material makes reference to certain financial instruments authorised and regulated in their jurisdiction(s) of incorporation. No action has been taken which would permit the public offering of the financial instrument(s) in any other jurisdiction, except as indicated in the most recent prospectus and the Key Investor Information Document (KIID) of the relevant financial instrument(s) where such action would be required, in particular, in the United States, to US persons (as such term is defined in Regulation S of the United States Securities Act of 1933). Prior to any subscription in a country in which such financial instrument(s) is/are registered, investors should verify any legal constraints or restrictions there may be in connection with the subscription, purchase, possession or sale of the financial instrument(s). Investors considering subscribing to the financial instrument(s) should read carefully the most recent prospectus and Key Investor Information Document (KIID) and consult the financial instrument(s') most recent financial reports. These documents are available on the website. Opinions included in this material constitute the judgement of the investment management company at the time specified and may be subject to change without notice. The investment management company is not obliged to update or alter the information or opinions contained within this material. Investors should consult their own legal and tax advisors in respect of legal, accounting, domicile and tax advice prior to investing in the financial instrument(s) in order to make an independent determination of the suitability and consequences of an investment therein, if permitted. Please note that different types of investments, if contained within this material, involve varying degrees of risk and there can be no assurance that any specific investment may either be suitable, appropriate or profitable for an investor's investment portfolio. Given the economic and market risks, there can be no assurance that the financial instrument(s) will achieve its/ their investment objectives. Returns may be affected by, amongst other things, investment strategies or objectives of the financial instrument(s) and material market and economic conditions, including interest rates, market terms and general market conditions. The different strategies applied to financial instruments may have a significant effect on the results presented in this material. Past performance is not a guide to future performance and the value of the investments in financial instrument(s) may go down as well as up. Investors may not get back the amount they originally invested. The performance data, as applicable, reflected in this material, do not take into account the commissions, costs incurred on the issue and redemption and taxes.

All information referred to in the present document is available on www.bnpparibas-am.com

INVESTORS' corner

February 2022 - Design: Creative Services BNPP AM - P2022038

BNP PARIBAS
ASSET MANAGEMENT

The sustainable investor for a changing world