Estimating corporate carbon emissions using artificial intelligence

By Maxime Barthe, Thomas Choquet, Tristan Jourde

Where firms fail to disclose their carbon emissions, the data can be estimated using machine learning models, which yield better predictive performances than standard methods. When combined with human expertise, these models can fill in gaps in the data and refine the assessment of transition risk.

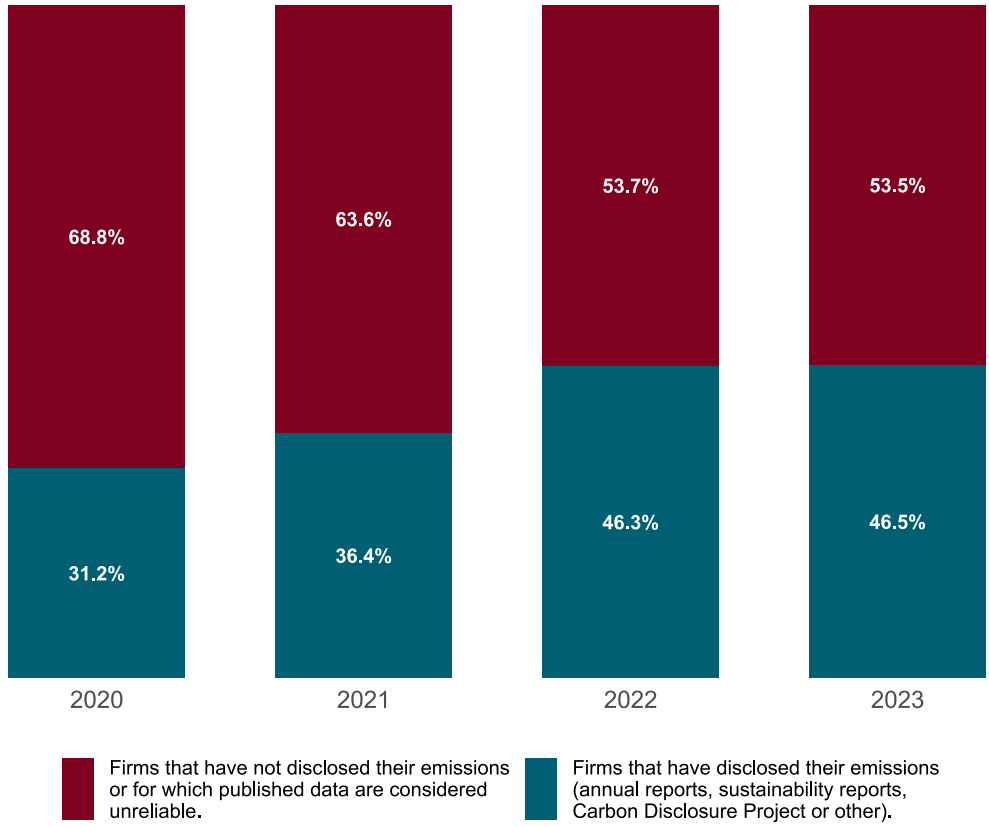

Chart 1:
Proportion of firms that disclose their carbon emissions
Sources: ISS Stoxx, Banque de France calculations.
Note: Sample of firms covered by ISS, all geographical areas combined.

For central banks, data on corporate carbon emissions have become essential for performing their core missions – i.e. monetary policy, financial stability and the provision of services to the economy and society. The information is needed to assess the exposure of the

financial system and economy to transition risk, which is the risk stemming from the emergence of new climate policies, disruptive technology and changes in consumer preferences. Carbon emissions data are also required to steer the greening of financial portfolios, including those held by central banks. For example, the Banque de France recently announced that it would include environmental criteria – and notably corporate carbon emissions – in its monetary policy operational framework to protect against a "potential decline in the value of collateral in the event of adverse climate-related transition shocks" (see press release of 29 July 2025). It has also adopted a responsible investment strategy for its proprietary portfolios (see 2024 Report on the Banque de France and ACPR's climate sustainable action).

The coverage of corporate carbon emissions data is improving but remains incomplete

Despite their growing importance, the data available on corporate carbon emissions remain partial (Grisey, 2022). While disclosures of emissions increased between 2020 and 2023, the trend is still insufficient, uncertain and unevenly distributed. In 2023, out of a sample of several tens of thousands of firms, only 47% disclosed their emissions (*Chart 1*). In reality, the proportion is even lower as the sample excludes small firms, most of which are not subject to reporting obligations. Moreover, with the introduction of European regulatory simplification initiatives such as the Omnibus Directive, many European companies may continue to not publish their emissions.

In this context, and for the use cases cited above, it is vital to be able to estimate carbon emissions to measure transition risk. This can be done through expert analysis of companies, or by using commercial data suppliers. Although expert assessments generally provide reliable results, they are costly in terms of time and resources. Meanwhile, commercial suppliers are often insufficiently transparent over the methodology and predictive performance of their models.

This blog post proposes a way to fill in the data gaps for firms that do not disclose their emissions. Using a set of machine learning tools, we estimate emissions automatically, transparently and on a large scale. We then compare the predictive performance of these algorithms with the estimates produced using standard econometric methods. Using this comparison, we identify the most accurate and robust model for estimating undisclosed corporate emissions. Thanks to their flexibility, machine learning models can analyse and incorporate more complex relationships between predictive variables and carbon emissions than standard econometric models.

This blog post is part of a growing academic literature on the modelling of corporate carbon emissions. Goldhammer et al. (2017), for example, were among the first to propose a quantitative approach using linear regression models. Nguyen et al. (2021) then took this a step further by incorporating machine learning models. They demonstrated that these methods considerably increase the accuracy of emissions estimates.

Machine learning provides accurate and large-scale emissions estimates

To fill in the gaps in existing data, the first stage consists in calibrating several models using data from more than 7,000 listed firms from around 100 countries. The models aim to predict as accurately as possible the firms' carbon intensities– i.e. the quantity of greenhouse gases they emit, both directly (Scope 1) and indirectly through their use of electricity (Scope 2), per million euro of turnover. This indicator measures each firm's ratio of emissions to output (in monetary value). An alternative method would be to predict the absolute level of emissions (tCO2e).

The variables used to calculate the predictions include firm-level environmental criteria (carbon reduction targets, sector, etc.) and financial characteristics (investment ratio, market risk, etc.). In a second step, we select the model that produces the best predictions in the training sample, and use it to estimate the carbon intensity of firms that do not disclose their emissions.

The best model is the random forest model developed by Breiman (2001). It manages to predict nearly 70% of the carbon intensities of firms out of sample with an error of less than 100 tCO2e/million euro of turnover (*Chart 2*), or around a quarter of the average value of published emissions (370 tCO2e/million euro of turnover). The high correlation (0.78) between estimated values and published values confirms the model's reliability (Chart 3). These performances are consistent with the results of other studies, such as those of the Institut Louis Bachelier (Barreau et al., 2024), and much higher than those obtained using a standard linear regression model.

According to our model, the most relevant variables for estimating carbon intensities are sectoral classification, geographical location and, to a lesser extent, stock market capitalisation and share of fixed assets. The carbon reduction targets published by firms play a limited role in estimating their carbon intensity.

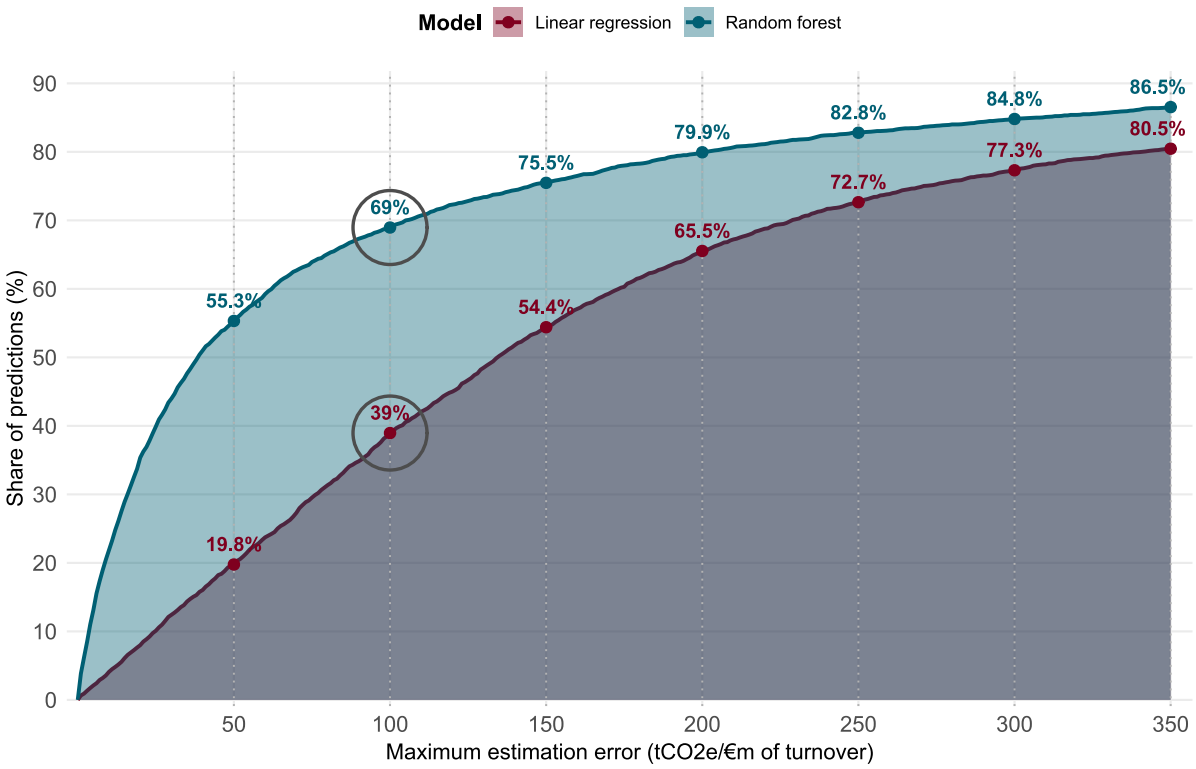

Chart 2: Comparison of the performances of two models in terms of estimate accuracy

Sources: ISS Stoxx, Eikon, Banque de France calculations.

Interpretation: The random forest model produces predictions with an error of less than 100 tCO2e/€m of turnover in 69% of cases, compared with 39% for the linear regression model.

AI models need to be supplemented with human expertise

One limitation of the model is its tendency to overestimate carbon intensities, especially for low-polluting firms (*Chart 3*), and, conversely, to underestimate carbon intensities for high-polluting firms. In these cases, expert analysis is needed to assess emissions in light of each firm's individual and sectoral specific features. The development of sub-models specific to groups of firms and incorporating new variables such as fossil fuel use could also improve overall predictive capacity.

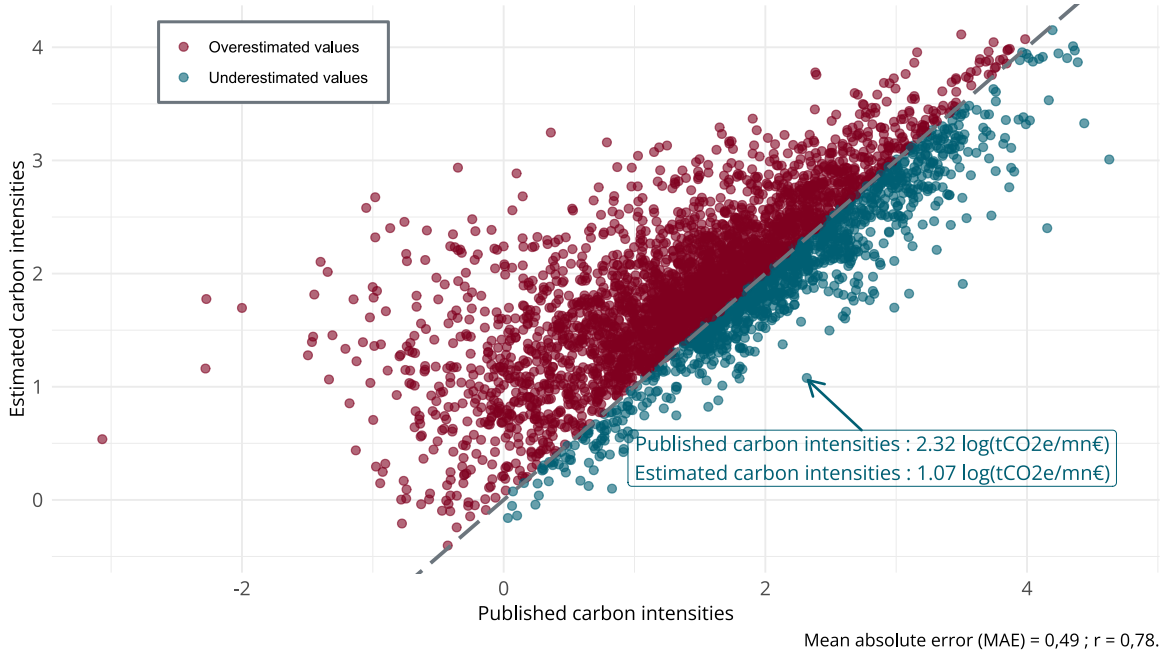

Chart 3: Carbon intensities estimated by AI versus those disclosed by firms

Sources: ISS Stoxx, Eikon, Banque de France calculations.

Note: Carbon intensities are direct and indirect emissions divided by turnover. Estimates are obtained using a random forest-type machine learning model. Values are expressed as decimal logs.

In conclusion, the quality and availability of data on carbon emissions is a key issue for central banks and financial institutions. Given the gaps in published emissions data, machine learning models are an effective solution for producing estimates on a large scale. These estimates deepen our understanding of transition risk, its impact on monetary policy and financial stability, and the extent to which the Eurosystem is exposed. While not a substitute for human expertise, machine learning models are a useful complementary tool. Moreover, while this blog post aims to estimate missing data, further work could be carried out to predict firms' future emissions, which is vital for anticipating their carbon reduction trajectories.