DEUTSCHE BUNDESBANK
EUROSYSTEM

A Comparative Analysis of Machine Learning and Classical Statistical Techniques for the Imputation of Corporate Green House Gas Emissions

Technical Report 2026-01

Deutsche Bundesbank, Research Data and Service Centre

Susanne Walter

Disclaimer: The views expressed in this technical report are personal views of the authors and do not necessarily reflect the views of the Deutsche Bundesbank or the Eurosystem.

Research Data and Service Centre

Abstract

The recognition that climate change has significant implications for the health of the financial system has elevated the necessity for improved access to reliable sustainability data among data producers and central banks. To capture the heterogeneity of firms within the financial system, micro-level data—specifically at the company level—are of particular importance. Despite extensive efforts to enhance corporate reporting, significant data gaps persist, particularly in the reporting of greenhouse gas (GHG) emissions. A high proportion of missing values poses a risk of sample bias in analyses. This paper addresses this issue by testing and evaluating various imputation approaches for corporate carbon emissions to mitigate these data gaps. In addition to individual balance sheet data, we utilize novel geospatial data, including characteristics of company real estate and its usage. We compare classical statistical methods with modern machine learning techniques. Approaches that account for the missing data generation process demonstrate superiority over simpler methods, yet they are outperformed by machine learning methods.

Acknowledgement: *We sincerely thank Martin Eisele for proof-reading this document and providing invaluable feedback. Furthermore, we want to thank the colleagues from the Federal Agency for Cartography and Geodesy for their significant contributions in facilitating the data exchange and the support in the Record Linkage. Your uncomplicated and friendly support made the process seamless and efficient and was a significant contribution to our project. We also want to thank the participants of Satellite Seminar on "Sustainability Data Issues and Central Banks' Experience", 4th October, 2025, in Amsterdam for their helpful feedback and discussions.*

Usage of AI: *AI assistants were used to correct grammar, for vocabulary check and code debugging.*

Keywords: data gaps, climate data, GHG emission, carbon footprint, EU ETS, missing data, imputation, machine learning, balance sheet data, Heckman correction, sample selection, LOD1, building data, digital twin

JEL classification: C38, C53, Q54

Version: 1.0

Citation: Walter,S. (2026). A Comparative Analysis of Machine Learning and Classical Statistical Techniques for the Imputation of Corporate Green House Gas Emissions, Technical Report 2026-01. Deutsche Bundesbank, Research Data and Service Centre.

Contents

1 Introduction

The awareness that the stability of the financial system is also affected by risks stemming from climate change was a fertile soil for several initiatives of central banks and other public institutions (NGFS 2019, European Commission 2024). Monitoring, evaluation and mitigation of these risks require sound reporting and reliable data (NGFS 2019, TCFD 2016). Since 2021 the IMF[1)] and since 2023 the ECB[2)] publish aggregated indicators on sustainable finance. Despite the intensive efforts to introduce harmonized and mandatory reporting frameworks, the data landscape on climate related information is still highly fragmented and inconsistent, especially in the field of corporate carbon emission data (NGFS 2024, European Commission 2024). While there exists a plethora of commercial data, reliable official data is rather scarce and characterized by low coverage (ECB 2024, NGFS 2024). Furthermore, a high proportion of commercial data is estimated based on black box of estimation methods (Kalesnik et al. 2022).

One solution to address these pressing data gaps is the imputation of missing data. Imputation of missing data is a standard technique in statistical production. Many approaches have been tested and evaluated in view of their suitability (Preising et al. 2021, Thurow et al. 2021, Schnetzer et al. 2015). Multivariate imputation with chained equations (MICE), predictive mean matching, nearest neighbor method and single regression-based techniques are among the most commonly employed and studied. More recent advances focus on the application of machine learning algorithms to fill missing data (Niederhametner at al. 2025, Jerez et al. 2010). Random forest-based methods have shown promising results (Stekhoven & Buehlmann 2012).

In the concrete case of corporate green house gas (GHG) emission data, imputation approaches can be conceptually differentiated regarding the starting point of the procedure: top down or bottom-up approaches. While top-down approaches break down sector emission aggregates to the individual company level using corporate employment shares (Hirvonen et al. 2021), bottom-up approaches predict emissions on the company level by using other company variables (for instance key financial performance figures) (see for instance Kalesnik et al. 2022, Olesiewicz et al. 2023). The ECB for instance uses the top-down approach to impute the GHG emission for constructing their carbon footprint indicators[3)]. Whilst their simplicity of implementation clearly speaks in favor of top-down approaches, they base on strong assumptions about proportionality (e.g. between emissions and employment) and neglect production and business model related heterogeneity of firms. However, microlevel analyses with company data require incorporating firm heterogeneity, especially when emission data is linked and set into relation to other firm variables (such as financial obligations and credit volumes). Therefore, more detailed bottom-up approaches can be more suitable in these cases. Machine learning methods have also demonstrated high accuracy in this particular use case of forecasting GHG emissions at the country level (Begum and Mobin 2025).

Another approach is based on using the production and consumption-oriented interrelations between sectors as depicted in Input-Output-Tables from the National Accounts (Liu&Fan 2017, Matthews et al. 2008, v. Kalckreuth 2022).

One important determinant for selecting a suitable approach is understanding the nature of the

1 https://www.imf.org/external/pubs/ft/ar/2022/in-focus/climate-change/
2 https://www.ecb.europa.eu/press/pr/date/2023/html/ecb.pr230124~c83dbef220.en.html
3 Technical Annex to the Report on Climate change-related statistical indicators, November 2025, https://www.ecb.europa.eu/stats/all-key-statistics/horizontal-indicators/sustainability-indicators/data/shared/files/Technical_annex.en.pdf

underlying data. Especially for the problem of imputation, it is crucial to understand the data generation mechanisms that resulted in the inherent missing pattern (Rubin 1976). Most of the approaches required the data to be missing at random (MAR) (Buczak et al. 2023), i.e. missing in emission data can be completely explained by observable factors. However, in the case of GHG emission data, the missing data generation process is assumed to be missing not at random (MNAR) due to the legal frameworks. In addition to observable company characteristics, the level of emissions itself determines whether a company reports its emissions. If the non-random nature of the missing data process is not considered during imputation, the estimates may be distorted due to sample selection bias (Heckman 1976, Heckman 1979).

To our knowledge, there is no study that compares several imputation approaches with respect to their accuracy when imputing individual corporate GHG emissions and also addresses the missing mechanism explicitly when predicting emission data. Thus, the aim of this study is to compare different bottom-up approaches for imputing GHG emissions on the individual company level. As training data, we use official emission data from the EU emission trading system (EU ETS), as it is deemed most reliable (exhibit the least measurement error) and has the same observational unit as the company data (single entity). Another reason for using EU ETS is that it only covers emissions originating directly from companies' operative activity (as the channel is clearer as compared to estimating scope 2 or 3 emissions)[4)].

This study represents the methodological extension of the works of Walter (2025). Walter used a Heckman correction to explicitly model the sample selection of the missing data stemming from the MNAR process in official emission data. We found that when considering a rich set of data sources and covariates, the process can be assumed to be MAR which allows for other approaches to be tested. In this study we use the same rich set of variables (including novel geospatial data) and introduce novel machine learning approaches (based on random forests) to impute missing values and compare their performance to classical imputation techniques (MICE, KNN, Heckmann correction and single regression methods) and top-down approaches.

We find that bottom-up approaches outperform simpler top-down methods. Furthermore, consistently with prior studies (Stekhoven & Buehlmann 2012), random forests outperform all other approaches. The remainder of the paper is structured as follows: in **Section 2** we provide a short overview of data sources used in this study. Subsequently, in the methodological **Section 3**, the applied imputation techniques are briefly introduced. In **Section 4**, the main results are presented. **Section 5** concludes and presents some avenues for further research.

4 Scope 2 emissions cover all indirect emissions originating from energy consumption. Scope 3 cover all indirect emissions from upstream and downstream value chain activities (GHG 2024).

2 Data

For this study we combine several data from various sources. The panel of balance sheet data is taken from the Individual financial statements of non-financial firms (JANIS)[5] from the Deutsche Bundesbank. The corresponding panel on GHG emissions is taken from the publicly available dataset on traded emissions from the EU Emission Trading System (EU ETS[6]). Companies are required to report their emissions that fall under the EU certificate trading. The obligation to report as a foundation for carbon pricing makes them the most reliable in the range of corporate emission data. Furthermore, as compared to data from corporate sustainable reports and commercial data providers, the observational units equal the one in the balance sheet data (single installation/ enterprise). The emissions on the installation level were aggregated to the operator/ company-level.

The data is enriched with information about the corporate building type and its usage from the 3D building model data from the Federal Agency for Cartography and Geodesy[7]. The combined dataset comprises 4,855 complete observations across all variables for 822 unique companies (forming an unbalanced company panel) over the period from 2012 to 2020.

To break down sector aggregates as the benchmark analysis, we also include the aggregate GHG emissions for Germany from the Air Emission Accounts on the NACE Rev2 1-digit level[8].

5 Here we use the internal, non-anonymized version of the JANIS dataset, which only contains individual financial reports from commercial data providers. DOI: 10.12757/Bbk.JANIS.9723.14.14, https://www.bundesbank.de/resource/blob/862454/016baaec31b6eefc6e47fc899e9a4b99/mL/2024-18-janis-data.pdf (Becker et al. 2024)

6 https://climate.ec.europa.eu/eu-action/carbon-markets/eu-emissions-trading-system-eu-ets_en

7 We use here Level-of-Detail 1 data because it contains the building function. https://gdz.bkg.bund.de/index.php/default/3d-gebaudemodelle-lod1-deutschland-lod1-de.html

8 https://ec.europa.eu/eurostat/databrowser/view/env_ac_aigg_q/default/table?lang=en&category=env.env_air.env_air_aa

3 Methodology

To compare several imputation approaches, from standard methods in medical research and official statistics to more modern machine learning techniques, we first split our sample randomly into 70% training and 30% test data. For the test data, the emission values were set to missing. This is necessary to calculate the accuracy metrics and to compare the prediction against the true values.

We chose the simplest approach as our benchmark approach which is the allocation of **sector specific emission aggregates** from the **Air emission accounts based on employment shares (I.)** as used by Hirvonen et al. 2021 and ECB 2024. A company's employment is divided by the total employment in the sector in which the company is active. Those employment shares are then multiplied with the total emissions of the sector to yield individual corporate emissions. Its simplicity and consistency at the aggregate level speak in favor for this approach. However, firm heterogeneity with respect to production technology and technical equipment is neglected. Furthermore, the assumption about proportionality between employees and emission output might not hold, particularly for service sectors.

As a second approach, we explicitly want to address the underlying selection mechanism that results in the missing pattern to correct for potential bias (we assume emission data to be MNAR). Furthermore, we use multivariate regression to model the relationship between a company's production process and their emission intensity. We use the widely accepted and used **Heckman correction (II.)** (Heckman 1976, Heckman 1979). The idea of this two-step approach is to explicitly model the selection (the missingness in the target variable, here emissions) and then run the imputation of emissions including a correction factor from the first step. The selection is modeled using a binary probit model, the outcome regression uses an OLS estimation. To analyse the value added of geospatial data for imputation accuracy, we first run the output model solely on the **balance sheet data and employment (IIa)** and then extend the model by also including the **building type from the geodata (IIb)** to predict emissions. Detailed descriptions and results can be found in Walter 2025.

Third, we also aim to compare a standard method used in statistics from various domains: **Multivariate Imputation by chained equations (MICE) (III.)** (van Buuren & Groothuis-Oudshoorn 2011). Usually, MICE is used to account for the uncertainty in the imputation process as it produces multiple imputed datasets (Azur et al. 2011). However, as our use case focuses on the prediction of individual emission values and not inferential analyses, we do not use this feature of multiple imputation here. We impute only one dataset (m=1) using **predictive mean matching** for 5 donors. Out of the 5 potential candidates, whose predicted values are closest to the prediction of the missing value, one is randomly chosen whose emission value replaces the missing value (Morris et al. 2014). One advantage of this method is that predicted values range within the plausible boundaries of the original values. [Morris et al. 2014] The analyses were implemented using the mice package in R.

As a fourth approach, we impute emissions by the **k-nearest neighbour algorithm (IV.)**. In this very popular approach, the mean of the k-most similar neighbours of the target variable is used to impute the missing value. We set the number of neighbours that are used to calculate the mean to 5 (k=5).

Recent evidence (e.g. Stekhoven & Buehlmann 2012, Tan and Ishwaran 2017) has shown that modern machine learning methods might outperform these widely applied, classical imputation approaches. For this reason, we also include modern imputation techniques that are based on **random forests (V.)** These Machine Learning approaches have the advantages that they require less assumptions about distributions and parameters as compared to single regression-based methods (Stekhoven & Buehlmann 2012). Random forests are estimated to predict missing emission values. We test the most common implementations of the random forest-based approaches in R: the **MissForest (Va)** packages by Stekhoven & Buehlmann (2012). This package is very flexible with regards of data types and distributional assumptions (Stekhoven & Buehlmann 2012). We then compare it to the **Ranger Random Forest (Vb)** and the **XGBoost (Gradient Boosting) method (Vc)** from the VIM package (Kowarik & Templ 2016), developed by Statistics Austria. Like MissForest, these two imputation methods focus on random forests, while XGBoost is based on Gradient Boosting.

To compare the accuracy of the predictions the deviation between the predicted values from the true values was calculated. Here we use the standard metric from the machine learning literature as a diagnostic tool (Hodson 2022): the root mean squared error (RMSE). It calculates the average deviation of the actual values from the predicted ones. As we want to come closest to the real values, when imputing them, RMSE is deemed an appropriate measure.

To compare all the approaches, only one column was set to missing (emissions as the target variable to impute) while all other variables did not contain any missings. Table 1 provides an overview of the covariates used in each approach to impute the missing emissions. The continuous variables were transformed to correct for outliers and skewed distributions (censoring to 99% percentile, normalized, logarithmic values).

Table 1: Variables used in the models

Variables	Approaches					
Dependent: GHG-Emissions	I. Employment shares	IIa. Heckman balance sheet data	IIb. Heckman balance sheet data and building data	III. MICE	IV. KNN	V. Random Forest
Employment	X	X				
Intangibles		X	X	X	X	X
Land and buildings		X	X	X	X	X
Technical equipment and machinery		X	X	X	X	X
Raw materials and consumables		X	X	X	X	X
Building type			X	X	X	X
NACE (2-digit)	X (1-digit)	X	X	X	X	X

4 Results

Table 2 provides an overview over the imputation accuracy of the implemented approaches. The simple top-down approach allocating sector aggregates by employment shares (I.) is outperformed by all bottom-up approaches, as firm heterogeneity seems to drive companies' emission intensities. The explicit consideration of the selection process with the Heckman correction yields more accurate estimates (IIa&IIb.). Including building function into the model besides sector dummies and balance sheet items improve accuracy even more (IIb). Walter 2025 has additionally shown that the missingness in emissions can be treated as MAR process when a rich set of firm variables is included and the selection and outcome can sufficiently be explained the relevant covariates. This is the foundation for the usage of further imputation methods as most approaches require the process to be MAR to yield unbiased estimates (Tan & Ishwaran 2017, Buczak et al. 2023). The single value imputation with the Heckman correction even outperforms the predictive mean matching in mice and the KNN-method. However, all approaches were outperformed by random forest-based methods (especially MissForest Va). This is in line with other findings from other domains (see for instance Jerez et al. 2010 and Tan & Ishwaran 2017 for the medical domain, Stekhoven & Buehlmann 2012 for bioinformatics and Niederhametner at al. 2025).

Predictive mean matching in MICE, when accounting for firm heterogeneity, yields values that are almost as accurate as those produced by the very simple approach based on employment shares.

Table 2: Accuracy of imputation approaches

Approach	Top down	Bottom up approaches						
	(I) Traditional: Using employment shares per industry	(IIa) Heckman correction Including balance sheets	(IIb) Heckman correction Including balance sheets Including building information	(III) MICE (m=1, pmm)	(IV) KNN (k=5)	(Va) MissForest	(Vb) Ranger Random Forest (VIM)	(Vc) XGBoost (VIM)
Root Mean Squared Error (Deviation imputed from actual values)	2.5252	1.8865	1.7070	2.3998	2.1356	1.0009	1.1167	1.1371

5 Conclusion

Our analyses for the German case have shown, that modern machine learning technique (especially random forest imputations) can benefit and improve the accuracy of imputing corporate GHG emissions. Even simpler bottom-up approaches, that account for firm heterogeneity at the micro-level, outperform top-down approaches based on employment weights. The higher accuracy at the microlevel emphasizes the superiority of bottom-up approaches when conducting analyses with microdata. Moreover, we have seen that novel data sources such as building data can help to overcome potential selection bias, add more explanatory power and increase the accuracy of the imputations.

Given that machine learning methods require bigger training data sets, our analyses could be further improved by multiple dimensions. First with respect to coverage, as individual emission data from the federal statistical offices might be integrated. In doing so, the proportions of missing in the data could be reduced and the robustness of our results could be evaluated. One approach could be to run the analysis for all countries covered by EU ETS and proof the generalizability of our results.

Second, as we only assume the data to be MAR, a combination of explicitly addressing an MNAR process and machine learning techniques could shed more light on the latent selection bias. Furthermore, other novel data sources (such as satellite or overflight images from the company property) can be combined to add more relevant information for the variation in corporate emission intensities.

Furthermore, our analyses could be complemented by the equally popular approach using input-output-tables (Liu&Fan 2017, Matthews et al. 2008, v. Kalckreuth 2022).

References

1. Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple imputation by chained equations: what is it and how does it work? Int J Methods Psychiatr Res. 2011 Mar;20(1):40-9. doi: 10.1002/mpr.329. PMID: 21499542; PMCID: PMC3074241.
2. Becker, T., Biewen, E., Hüwel, D., and Schultz, S. (2024). Individual financial statements of non-financial firms (JANIS), Data Report 2024-18 – Metadata Version JANIS Data-Docv14. Deutsche Bundesbank, Research Data and Service Centre.
3. Begum, A.M., Mobin, M.A. A machine learning approach to carbon emissions prediction of the top eleven emitters by 2030 and their prospects for meeting Paris agreement targets. Sci Rep 15, 19469 (2025). https://doi.org/10.1038/s41598-025-04236-5
4. Buczak P, Chen JJ, Pauly M. Analyzing the Effect of Imputation on Classification Performance under MCAR and MAR Missing Mechanisms. Entropy (Basel). 2023 Mar 17;25(3):521. doi: 10.3390/e25030521. PMID: 36981409; PMCID: PMC10048089.
5. ECB. Climate change-related statistical indicators. ECB Statistics Paper Series No 48. 2024
6. European Commission 2024 Report from the commission on the monitoring of climate-related risk to financial stability. Brussels, 28.6.2024 C(2024) 4372 final. https://finance.ec.europa.eu/document/download/9e2c0695-9da6-4b09-ae43-78729fc7609e_en?filename=240701-climate-risks-report_en.pdf
7. Heckman JJ. Sample selection bias as a specification error. Econometrica 1979; 47: 153–161.
8. Heckman, J.J., 1976. The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models, in: Annals of economic and social measurement, Volume 5, number 4. NBER, pp. 475-492.
9. Hirvonen A, Karhu A and Tolkki V. Measuring the carbon footprint of loans to domestic non-financial corporations (NFC) by banks. Paper presentation at the 19th Statistical Forum of the IMF, 17–18 November 2021, Washington D.C. https://www.imf.org/-/media/Files/Conferences/2021/9th-stats-forum/33FinalPaperVilleTolkki.ashx
10. Hodson, T. O.: Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not, Geosci. Model Dev., 15, 5481–5487, https://doi.org/10.5194/gmd-15-5481-2022, 2022.
11. Jerez JM, Molina I, García-Laencina PJ, Alba E, Ribelles N, Martín M, Franco L. Missing data imputation using statistical and machine learning methods in a real breast cancer problem. Artif Intell Med. 2010 Oct;50(2):105-15. doi: 10.1016/j.artmed.2010.05.002. Epub 2010 Jul 16. PMID: 20638252.
12. Kalesnik V, Wilkens M and Zink J. Green data or greenwashing? Do corporate carbon emissions data enable investors to mitigate climate change? The Journal of Portfolio Management 2022; 48(10): 119–147
13. Kowarik A, Templ M (2016). "Imputation with the R Package VIM." Journal of Statistical Software, 74(7), 1–16. doi:10.18637/jss.v074.i07.
14. Liu, H., Fan, X., 2017. Value-added-based accounting of co2 emissions: a multi-regional input-output approach. Sustainability 9, pp. 1-18.
15. Matthews, H. S., Weber, C., Hendrickson, C.T.(2008). Estimating Carbon Footprints with Input-Output Models. International Input Output Meeting on Managing the Environment. 9-11 July 2008, Seville (Spain).
16. Morris, T.P., White, I.R. & Royston, P. Tuning multiple imputation by predictive mean matching and local residual draws. BMC Med Res Methodol 14, 75 (2014).
17. NGFS 2019, A call for action - Climate Change as a source of financial risk, https://www.ngfs.

net/system/files/import/ngfs/medias/documents/ngfs_first_comprehensive_report_-_17042019_0.pdf.

18. NGFS 2024, Improving Greenhouse Gas Emissions Data, https://www.ngfs.net/system/files/import/ngfs/medias/documents/ngfs_information_note_on_improving_ghg_emission_data.pdf.
19. Niederhametner N, Gussenbauer J, Kowarik A. Performance evaluation of machine learning-based imputation for missing data analysis in the R package VIM. Statistical Journal of the IAOS. 2025;41(2):254-264. doi:10.1177/18747655251339401
20. Olesiewicz, M.P., Kooroshy, J., Greven, S. 2023. Navigating the corporate disclosure gap: Modelling of Missing Not at Random Carbon Data. The Journal of Impact and ESG Investing. Volume 4, Issue 4.
21. Preising, Marcel & Lange, Kerstin & Dumpert, Florian, 2021. Imputation zur maschinellen Behandlung fehlender und unplausibler Werte in der amtlichen Statistik, WISTA – Wirtschaft und Statistik, Statistisches Bundesamt (Destatis), Wiesbaden, vol. 73(5), pages 40-52.
22. Rubin, D.B., 1976. Inference and missing data. Biometrika 63, pp.581-592.
23. Schnetzer, M., Astleithner, F., Cetkovic, P., Humer, S., Lenk, M., & Moser, M. (2015). Quality Assessment of Imputations in Administrative Data. Journal of Official Statistics, 31(2), 231-247. https://doi.org/10.1515/jos-2015-0015 (Original work published 2015)
24. Stekhoven DJ, Buehlmann P (2012). "MissForest - non-parametric missing value imputation for mixed-type data." Bioinformatics, 28(1), 112–118.
25. Tang F, Ishwaran H. Random Forest Missing Data Algorithms. Stat Anal Data Min. 2017 Dec;10(6):363-377. doi: 10.1002/sam.11348. Epub 2017 Jun 13. PMID: 29403567; PMCID: PMC5796790.
26. TCFD 2016, Phase 1 Report of the Task Force on Climate-Related Financial Disclosures. https://www.fsb.org/uploads/TCFD-Phase-1-report.pdf
27. The Greenhouse Gas (GHG) Protocol. A Corporate Accounting and Reporting Standard 2024. https://ghgprotocol.org/sites/default/files/standards/ghg-protocol-revised.pdf
28. Thurow, Maria & Dumpert, Florian & Ramosaj, Burim & Pauly, Markus. (2021). Imputing missings in official statistics for general tasks – our vote for distributional accuracy. Statistical Journal of the IAOS. 37. 1-12. 10.3233/SJI-210798.
29. van Buuren S, Groothuis-Oudshoorn K (2011). "mice: Multivariate Imputation by Chained Equations in R." Journal of Statistical Software, 45(3), 1-67. doi:10.18637/jss.v045.i03.
30. von Kalckreuth U. Pulling Ourselves Up by Our Bootstraps: The Greenhouse Gas Value of Products, Enterprises and Industries 2022. Deutsche Bundesbank Discussion Paper No. 23/2022. http://dx.doi.org/10.2139/ssrn.4201913.
31. Walter S. Estimating companies' GHG emissions using information from individual financial accounts and building data. Statistical Journal of the IAOS. 2025;41(2):305-316. doi:10.1177/18747655251341456