

Estimating companies' GHG emissions using information from individual financial accounts and building data

Dr. Susanne Walter, Data Service Centre of the Deutsche Bundesbank

Outline

1

Motivation

2

Challenges

3

Data Overview

4

Results


5

Conclusion and extensions

Motivation

Data gaps in the sustainable finance landscape

- Measuring potential **risk exposure** imposed by **climate change** for financial systems
- **Publication** of climate related **indicators** to substantiate policy making:
 - Central banks (ECB and national banks) started to publish climate related indicators to depict physical and transitory risks (such as emerging from CO2 emissions)
 - An international working has developed indicators for carbon footprint
- **Data gaps** in GHG emission data
 - Low **coverage**: Plethora of commercial data on GHG but reliable official data is rather scarce and low coverage (due to existing framework)
 - Lack of **transparency**: black box of estimation methods
 - Reported data by companies voluntary, not properly audited and not harmonised




EUROPEAN CENTRAL BANK | EUROSISTEM

Monetary policy & markets | Payments & financial stability | Statistics | The euro | Research

Home > Statistics > All key statistics

Analytical indicators on carbon emissions

Our carbon emissions indicators provide information on the carbon intensity of the securities and loan portfolios of financial institutions, and on the financial sector's exposure to counterparties with carbon emissions indicators help users in financing carbon-related activities, ar



© Adobe Stock / Miha Creative

Climate-related disclosures by the Deutsche Bundesbank 2023

Part of the Eurosystem-wide climate-related disclosures on the non-monetary policy portfolios (NMPPs)

23.03.2023 DE

Can we use other official data to close the data gaps

- Propose **bottom-up approach** to close data gaps (beyond employment and sector information)
- Combine financial data and novel geospatial data with emission data
- Model the **data generation process** explicitly for missing emission data
- Evaluate **official data** on company GHG emissions
- Compare **performance** of different approaches

Challenges

Data characteristics

1	Motivation
2	Challenges
3	Data Overview
4	Results
5	Conclusion and Extensions

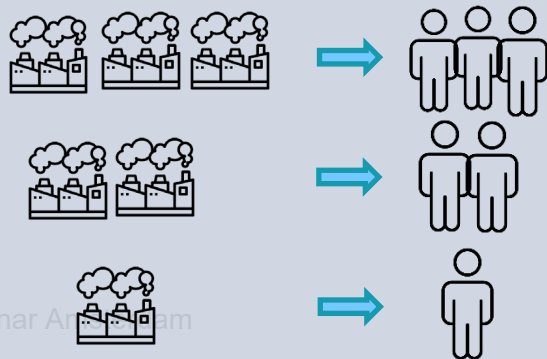
- **Missing data generation** (Rubin classification 1976)
 - Missing completely at random (MCAR)
 - Missing at random (MAR)
 - Missing not at random (MNAR)
- Emission data are **MNAR** because the amount of emissions would determine whether a company reports emissions or not
- **Only 1,470 German companies** report emission data in the official registry
- Commercial emission data **mostly estimated** (models are blackbox)
- Reported emissions in financial reports comprise the GHG emissions of the **whole enterprise group**

Challenges

Solution approaches

Top down

- **Breakdown** of country/ sector aggregates to company level proportional to employment
- Easy to implement
- Proportionality assumption

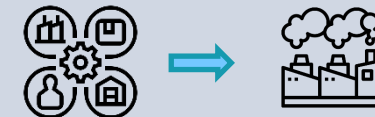


Input Output Tables

- Company emissions are estimated based on **Input-Output-Tables**
- e.g. Liu & Fan (2017) or Matthews et al. (2008)

Bottom up

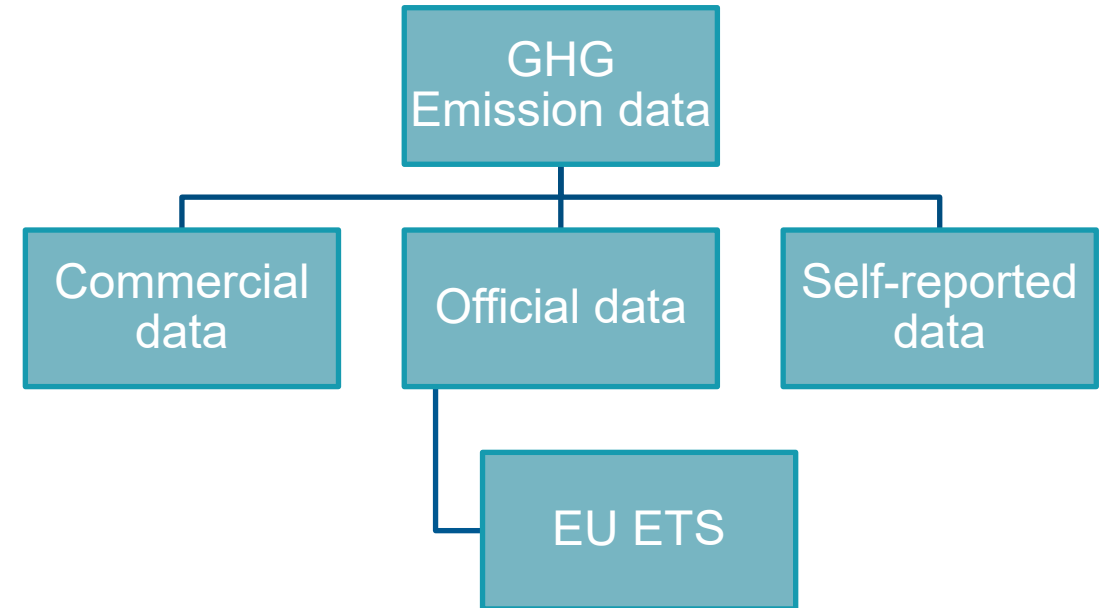
- Impute/ Predict emissions by means of **other company variables**
- Kalesnik et al. (2020): evaluate voluntary reporting and commercial data for listed companies in UK
- Olesiewicz et al. (2023) advanced imputation approach, but restricted set of variables



Data Overview

Official Emission data

- Emission data from the EU Emission Trading Systems (EU ETS) starting from 2008
 - ✓ **Reliable**
 - ✓ **Standardised:** multiple countries engaged in emission trading
 - ✓ **Granular:** Information on plant level (not like commercial data or data from financial reports)
 - ✓ **Linkable:** to financial data or other company data (identifiers, entity concepts)
- Cover **Scope 1** emissions (directly produced during production process)



Data Overview

Bundesbank data

1	Motivation
2	Challenges
3	Data Overview
4	Results
5	Conclusion and Extensions

- Individual **financial statements** of German non-financial firms from public sources (provided by commercial providers) (JANIKa)
- Financial figures from financial reporting (balance sheets and income statements) of German companies from 1997-2023
- Anonymous version is **accessible** for researchers ([Annual financial statements of non-financial firms \(JANIS and Ustan\) | Deutsche Bundesbank](#))
- Balance sheet items as **depictions of production factors** and value added → approximation of production process
- Gather information on **nexus** between emissions and production factors
- More **exact distribution** of emissions

Data Overview

Geospatial Data

- Digital Twin Data from the Federal Agency for Cartography and Geodesy
- LoD1 Data: Level of Detail ([3D-Gebäudemodelle LoD2 Deutschland \(bund.de\)](https://www.bund.de/3D-Gebäudemodelle-LoD2-Deutschland))
- Detailed information on buildings in the German jurisdiction, including the building type and function as well as 3D model with geocoordinates (CityGML) and other characteristics (# floors, volume, height, roof types)
- Can the function of a company building predict emissions?



F. Biljecki et al. / Computers, Environment and Urban Systems 59 (2016) 25–37



Fig. 1. The five LODs of CityGML 2.0. The geometric detail and the semantic complexity increase, ending with the LOD4 containing indoor features

Data Overview

Model to account for sample selection

1	Motivation
2	Challenges
3	Data Overview
4	Results
5	Conclusion and Extensions

Heckman's two step procedure (Heckman 1976) :

I. Model the **Selection** (the „missingness“)

$$M_{it} = Size_{it} + NACE_{it} \rightarrow IMR_{it}$$

II. Model the **Outcome** (GHG emissions values)

$$E_{it} = \beta_0 + \beta_1 build_{it} + \beta_2 NACE_{it} + \beta_3 Intang_{it} + \beta_4 Land_{it} + \beta_5 Tech_{it} + \beta_6 Raw_{it} + \beta_7 IMR_{it} + \sum_{j=8}^n \beta_j X_{itj}$$

Inverse Mills Ratio as
correction factor for
selection

Data cleaning

- Censoring (outlier correction)
- Normalisation (in relation to total assets, emissions relative and log transformation)
- Clustered standard errors (company level)
- Industry fixed effects (unobserved heterogeneity)

Variables

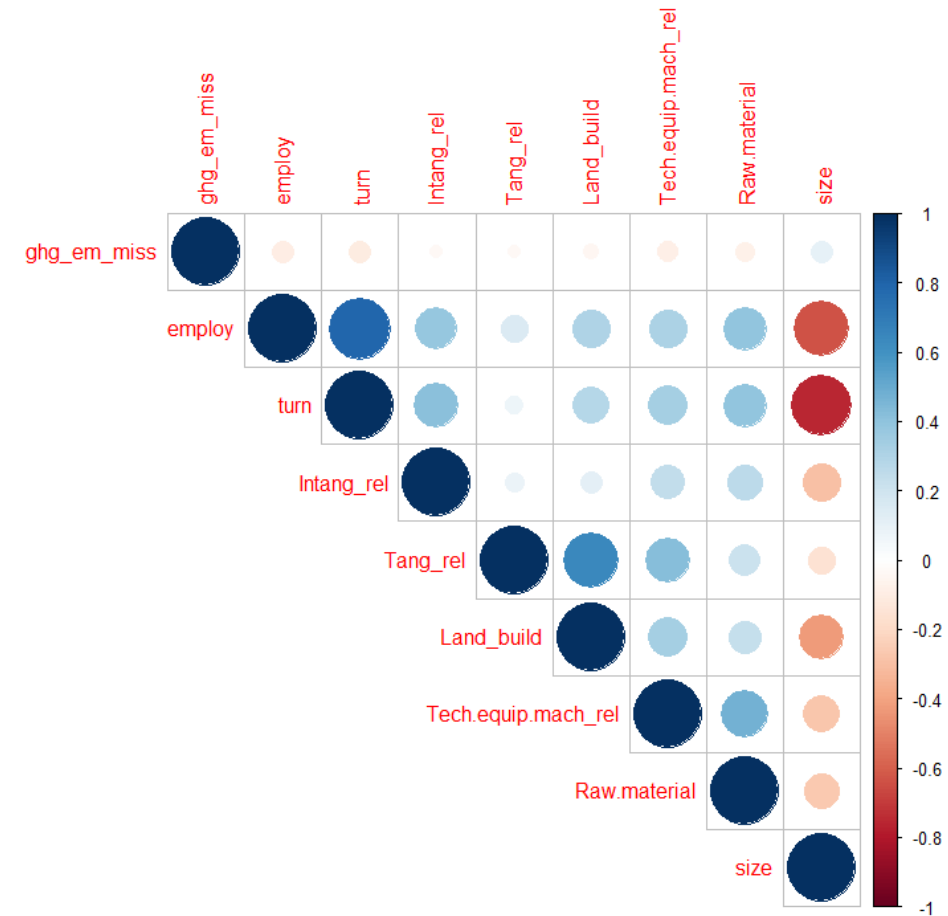
- *M* – Missing in emissions (yes=0, no=1)
- *E* - Emissions
- *Build* – Building function
- *Intang* –Intangible Assets
- *Size* – Company size class (XS-L), employment
- *Land* – Land and buildings
- *IMR* – Inverse Mills Ratio
- *Tech* – Technical equipment and machinery
- *Raw* – Raw materials and consumables
- *NACE* – Industry Fixed Effects

Results

Selection model

- Size mainly drives Reporting of Emissions
- Due to nature of data opposite to Olesiewicz et al. (2023)
- Larger companies are more likely to report emissions (largest companies as reference category here)

Probit					
	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	-5.53	130.20	-0.042	0.966	
sizeM	-0.75	0.02	-43.788	<2e-16	***
sizeS	-1.25	0.03	-42.486	<2e-16	***
sizeXS	-1.89	0.08	-24.614	<2e-16	***
Industry FE	Yes				
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					
(Dispersion parameter for binomial family taken to be 1)					
Null deviance: 57,619 on 573,663 degrees of freedom					
Residual deviance: 35,583 on 573,575 degrees of freedom					



Results

Outcome model

1	Motivation
2	Challenges
3	Data Overview
4	Results
5	Conclusion and Extensions

Bottom up approach: Step Two Estimation of Outcome Model (Estimating GHG based on balance sheet items)

Dependent Variable: GHG emissions (rel_log)			
Independent Variables	(1)	(2)	(3)
Constant	4.79*** (0.425)		
log(employ)	-0.121* (0.052)	-0.075 (0.254)	-0.243 (0.241)
IMR	-0.743*** (0.135)	0.244 (0.269)	0.101 (0.279)
log(empl)^2		0.012 (0.024)	0.029 (0.023)
Intangibles_rel			-0.420*** (0.092)
Land_rel			0.155* (0.069)
Tech_and_mach_rel			0.025 (0.069)
Raw_materials_and_consum_rel			0.439*** (0.094)
Fixed-Effects:			
Industry	No	Yes	Yes
S.E.: Clustered	by: Company ID	by: Company ID	by: Company ID
Observations	5,025	5,025	5,025
R ²	0.040	0.156	0.198
Within R ²	--	0.002	0.051

Results

Outcome model

1	Motivation
2	Challenges
3	Data Overview
4	Results
5	Conclusion and Extensions

Bottom up approach: Outcome model incorporating building data

Dependent Variable: GHG emissions (rel_log)		
Independent Variables	(4)	(5)
Constant	4.79*** (0.425)	
log(employ)	-0.340 (0.243)	
IMR	0.160 (0.242)	0.156 (0.209)
log(empl)^2	0.033 (0.023)	
Intangibles_rel	-0.386*** (0.088)	-0.387*** (0.088)
Land_rel	0.178* (0.072)	0.169* (0.071)
Tech_and_mach_rel	0.002 (0.065)	-0.007 (0.065)
Raw_materials_and_consum_rel	0.333*** (0.080)	0.321*** (0.079)
Controls: Building type	Yes	Yes
Fixed-Effects:		
Industry	Yes	Yes
S.E.: Clustered	by: Company ID	by: Company ID
Observations	4,855	4,855
R ²	0.378	0.376
Within R ²	0.260	0.258

Performance of different approaches (Top down, bottom up)

Approach	Top down	Bottom up				
	(I)	(II)	(III)	(IV)	(V)	(VI)
	Using employment shares per industry	Heckman correction Including balance sheets	Heckman correction Including balance sheets Including building information	KNN (k=5)	MissForest	Ranger Random Forest (VIM)
Root Mean Squared Error (Deviation imputed from actual values)	2.4624	1.8865	1.7070	2.1356	1.0009	1.1167

Conclusion

Potential extensions and further research avenues

1	Motivation
2	Challenges
3	Data Overview
4	Results
5	Conclusion and Extensions

Summary

- **Bottom up** approaches that incorporate further data **outperform** top down approach based on employment weights alone
- Machine Learning approaches (especially **random forests imputations**) **outperform** classical statistical approaches
- **Building** data adds significant **explanatory power**
- Emissions are more closely related to production processes and input intensities—such as **tangible fixed assets, land and buildings, and raw materials**—than to firm size, even after controlling for industry-specific effects.
- Top- down methods based only on industry and employment data may overlook **firm-level variation**
- Firms with a higher share of **intangible assets**—often associated with innovation and R&D—appear to have a lower **emissions footprint**, hinting at a potential link between innovative capacity and carbon efficiency.

Possible extensions

- **Linkage:** Integrate **Energy consumption** from structural business statistics X **emission factors** (from UBA)
- **Generalisability:** Include other countries (ETS covers EU-countries and linktables are available)
- **Novel approaches:** Test further ML approaches such as **XGBoost** or **Transformer Models** (VIM will be extended)

Thank you for your attention

Dr. Susanne Walter

Susanne.Walter@bundesbank.de



References

Rubin, D.B., 1976. Inference and missing data. *Biometrika* 63, pp.581-592.

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.

Liu, H., Fan, X., 2017. Value-added-based accounting of co2 emissions: a multi-regional input-output approach. *Sustainability* 9, pp. 1-18.

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).

Kalesnik, V., Wilkens, M., Zink, J., 2020. Green data or greenwashing? do corporate carbon emissions data enable investors to mitigate climate change? *SSRN Electronic Journal* doi:10.2139/ssrn.3722973

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.

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](https://doi.org/10.18637/jss.v074.i07).

Stekhoven DJ (2022). missForest: Nonparametric Missing Value Imputation using Random Forest. R package version 1.5.

Stekhoven DJ, Buehlmann P (2012). "MissForest - non-parametric missing value imputation for mixed-type data." *Bioinformatics*, 28(1), 112–118.

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

Icons from flaticon and freepik (Alfredo Creates)