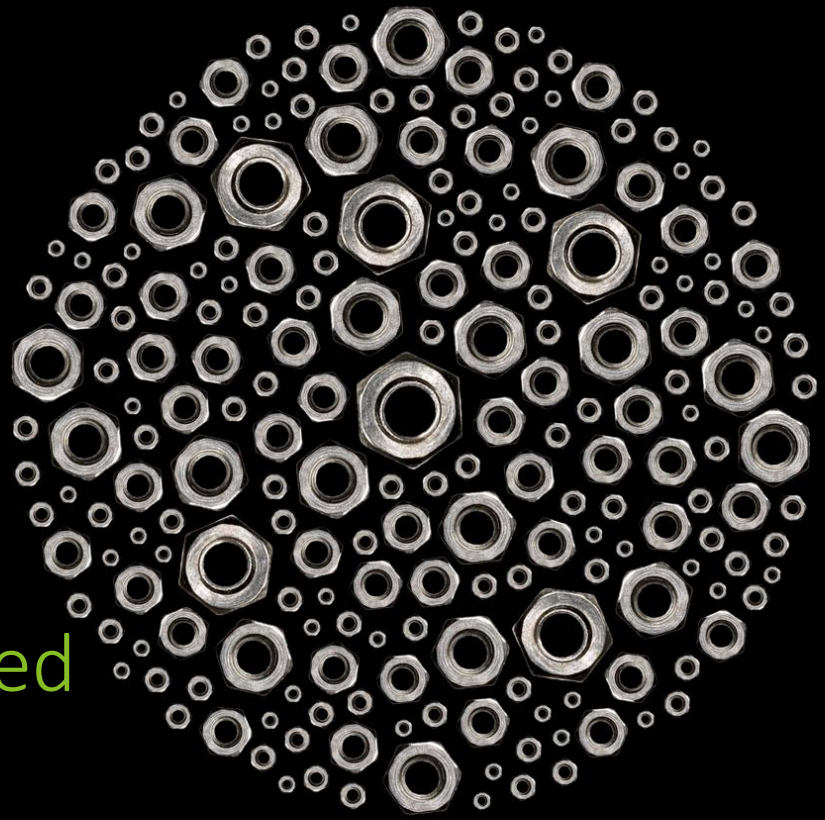
Deloitte.

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken



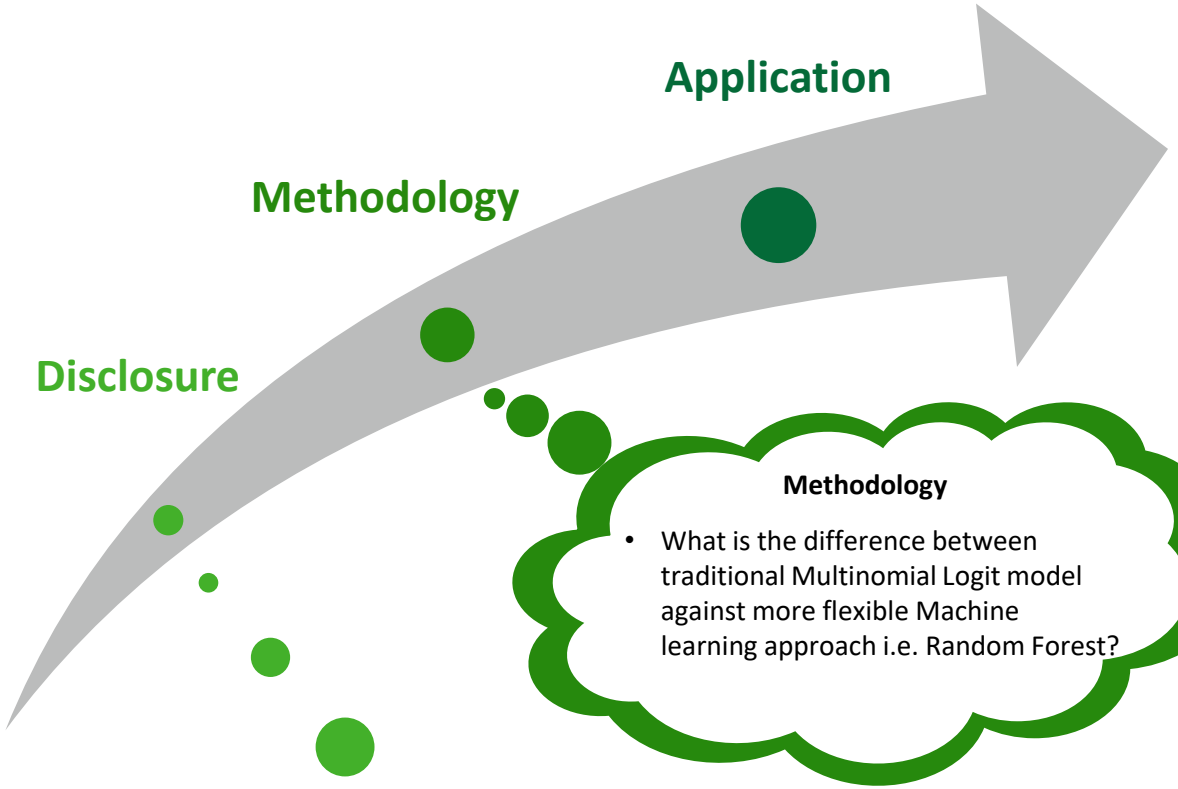
Climate Disclosure: A Machine Learning-Based Analysis of Company-Level Emissions and ESG Data Disclosure

July 26, 2023, Andrej Bajic, Deloitte Germany, University Duisburg-Essen

MAKING AN IMPACT THAT MATTERS since 1845

Table of content

Research questions and key motivation for the analysis



Methodology

- What is the difference between traditional Multinomial Logit model against more flexible Machine learning approach i.e. Random Forest?

Disclosure

- What are significant differences and trends to which ESG and GHG emission data are disclosed?
- Do carbon more intensive sectors disclose more readily information on GHG emissions and ESG data than their less intensive inter sector competitors?

Predictability

- Can we identify blind spots in the financial asset portfolios?

MNLogit Regression Results

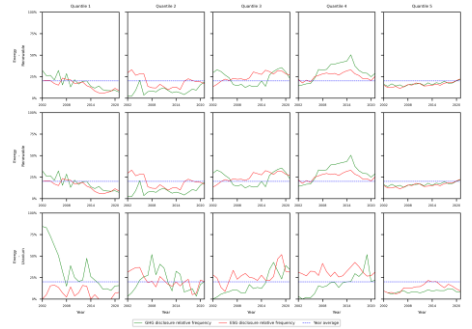
Dep. Variable:	Indicator	No. Observations:	459078
Model:	MNLogit	Df Residuals:	459060
Method:	MLE	Df Model:	16
Date:	Fri, 24 Jun 2022	Pseudo R-squ.:	0.2801
Time:	18:06:02	Log-Likelihood:	-1.5086e+05
converged:	True	LL-Null:	-2.0957e+05
Covariance Type:	nonrobust	LLR p-value:	0.000

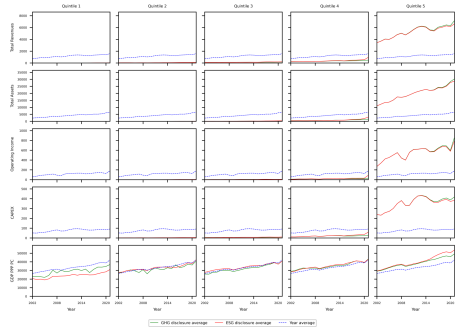
Indicator=1	coef	std err	z	P>|z|	[0.025	0.975]
const	-234.0146	2.809	-83.306	0.000	-239.520	-228.509
totalRevenues	0.0002	2.78e-06	76.915	0.000	0.000	0.000
totalAssets	5.304e-06	3.68e-07	14.403	0.000	4.58e-06	6.03e-06
employees	3.365e-05	6.69e-07	50.269	0.000	3.23e-05	3.5e-05
fiscalYear	0.1135	0.001	81.338	0.000	0.111	0.116
countryNumericCode	0.0008	2.75e-05	28.453	0.000	0.001	0.001
busIdDerived	0.0002	2.54e-05	6.266	0.000	0.000	0.000
gdp	3.694e-05	4.69e-07	78.696	0.000	3.6e-05	3.79e-05
unRegionCode	-0.0011	0.000	-9.757	0.000	-0.001	-0.001

Indicator=2	coef	std err	z	P>|z|	[0.025	0.975]
const	-376.9447	3.476	-108.448	0.000	-383.757	-370.132
totalRevenues	0.0002	2.81e-06	88.851	0.000	0.000	0.000
totalAssets	5.338e-06	3.7e-07	14.427	0.000	4.61e-06	6.06e-06
employees	3.41e-05	6.76e-07	50.413	0.000	3.28e-05	3.54e-05
fiscalYear	0.1857	0.002	107.592	0.000	0.182	0.189
countryNumericCode	0.0007	3.13e-05	22.485	0.000	0.001	0.001
busIdDerived	-0.0004	3.04e-05	-12.999	0.000	-0.000	-0.000
gdp	3.113e-05	5.12e-07	60.790	0.000	3.01e-05	3.21e-05
unRegionCode	0.0051	8.87e-05	57.109	0.000	0.005	0.005

	Precision	Recall	F1-score	Support
No disclosure	0.98	1.00	0.99	154 505
ESG disclosure	0.95	0.76	0.85	9 292
GHG disclosure	0.93	0.87	0.90	7 835
Accuracy			0.98	171 632
Macro avg.	0.96	0.88	0.91	171 632
Weighted avg.	0.98	0.98	0.98	171 632

	Precision	Recall	F1-score	Support
No disclosure	0.92	0.99	0.96	154 505
ESG disclosure	0.40	0.08	0.14	9 292
GHG disclosure	0.70	0.29	0.41	7 835
Accuracy			0.91	171 632
Macro avg.	0.68	0.46	0.50	171 632
Weighted avg.	0.88	0.91	0.89	171 632





Literature overview

Recent academic developments

Bingler J.A., et al. 2021, ***Cheap Talk and Cherry-Picking. What Climate Bert has to say on Corporate Climate Risk Disclosures***

- Develop AI model to question climate disclosure on four core categories.
- Companies supporting TCFD recommendations mostly disclose non-material climate risk information.

Krueger P., et al. 2021, ***The effects of Mandatory ESG Disclosure around the World***

- Mandatory ESG disclosure has a significant and positive effect on the propensity of firms to file ESG reports.
- Quality of reports are higher when mandatory disclosure is in place and where investors demand for ESG related data is largest.

ZHANG Y., Liu J., 2020, Overview of research on carbon information disclosure

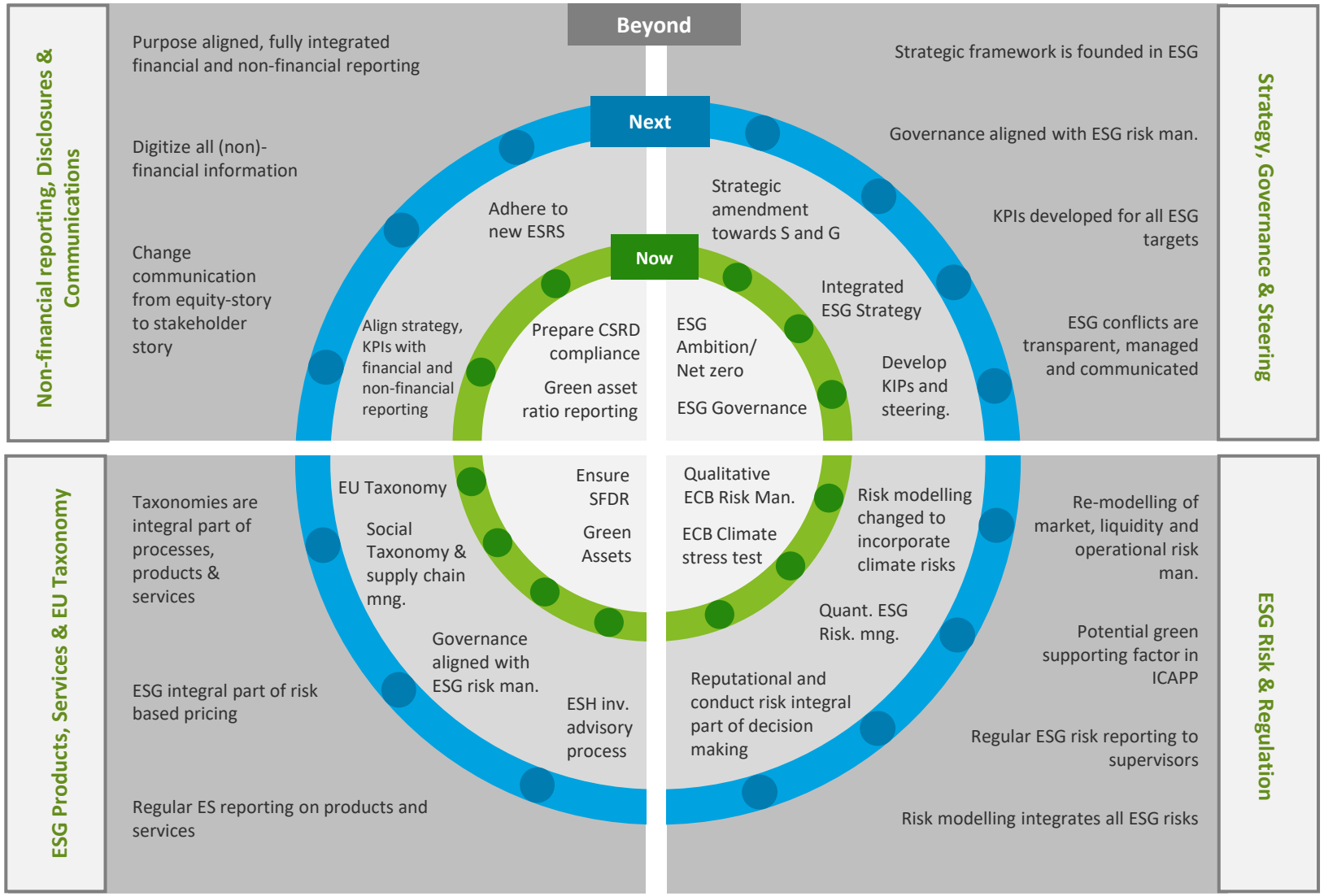
- Systematic overview of literature on carbon information disclosure.
- Disclosure has been increasing, quality and content shouild be improved.
- Discussion on the disclosure influencing factors

Theoretical background of disclosure

- Legitimacy theory: corporations with a high carbon footprint and the potential for environmental damage are likely to volunteer for GHG data disclosure and damage reduction. Lemma et al (2019)
- Voluntary disclosure theory suggests that companies with superior environmental performance are more likely to disclose their ESG/GHG profiles including GHG emission amounts. Giannarakis et al. (2018)

Disclosure development across different segment

ESG and carbon disclosure is essential step for business transformation and reaching net-zero targets



NOW

- According to CSRD all large companies need to publish regular reports on their environmental and social impact activates.
- ECB assesses how prepared are the banks for the shocks caused by the climate change.
- 55% net emission reduction target by 2030.
- SFRD introduced to improve transparency of the sustainable investment products.

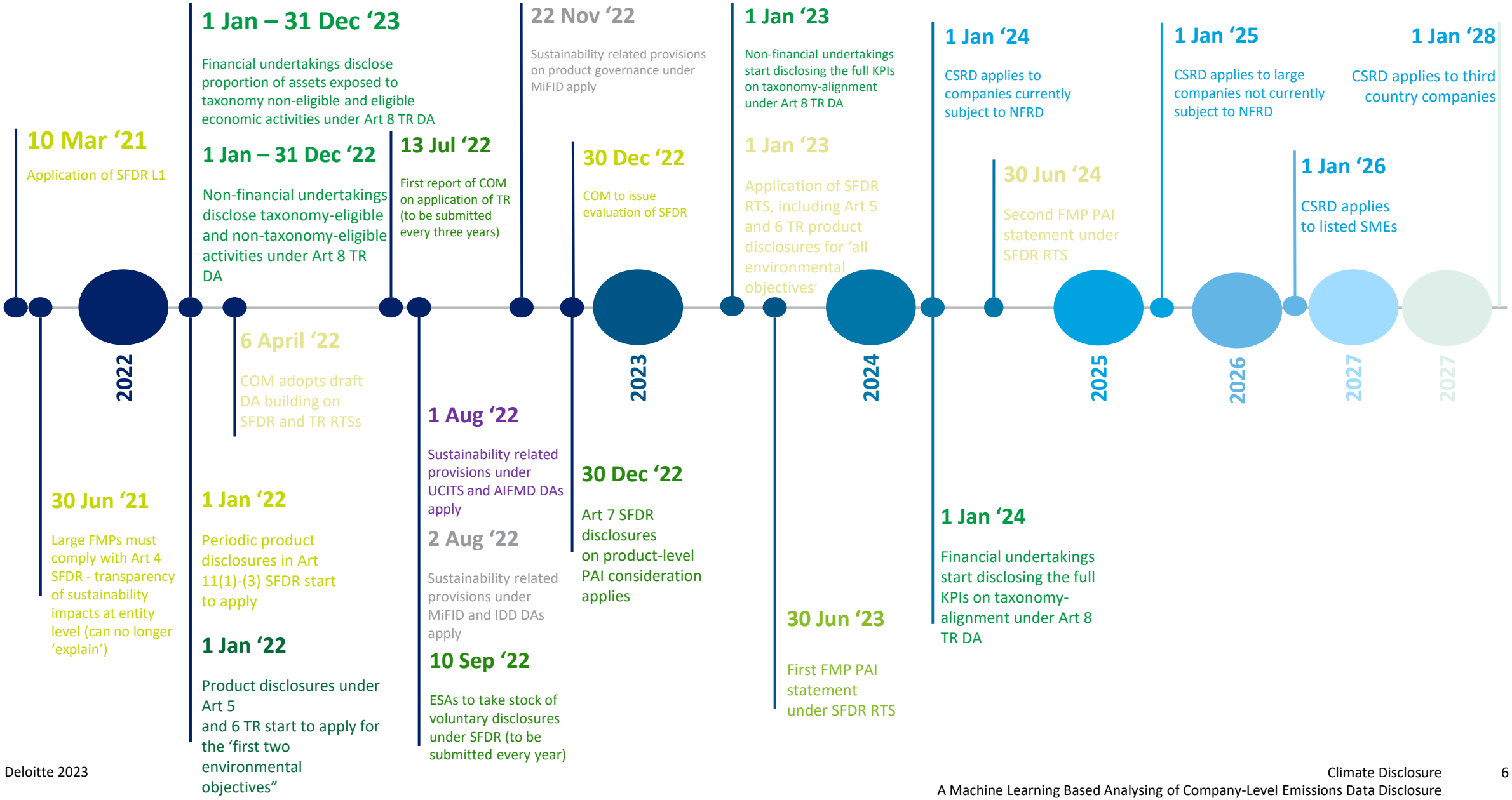
NEXT

- Taxonomy is used for claiming companies and investors as climate-friendly.
- Companies are expected to develop and apply environment KPIs.
- The companies are expected to include climate risks into modeling, decision making and governance.

BEYOND

- Taxonomy has become the integral part of all aspects of products and services.
- All the companies activities take ESG into consideration.
- All the ESG information is available and presented by the companies

EU disclosure development: SFDR & CSRD

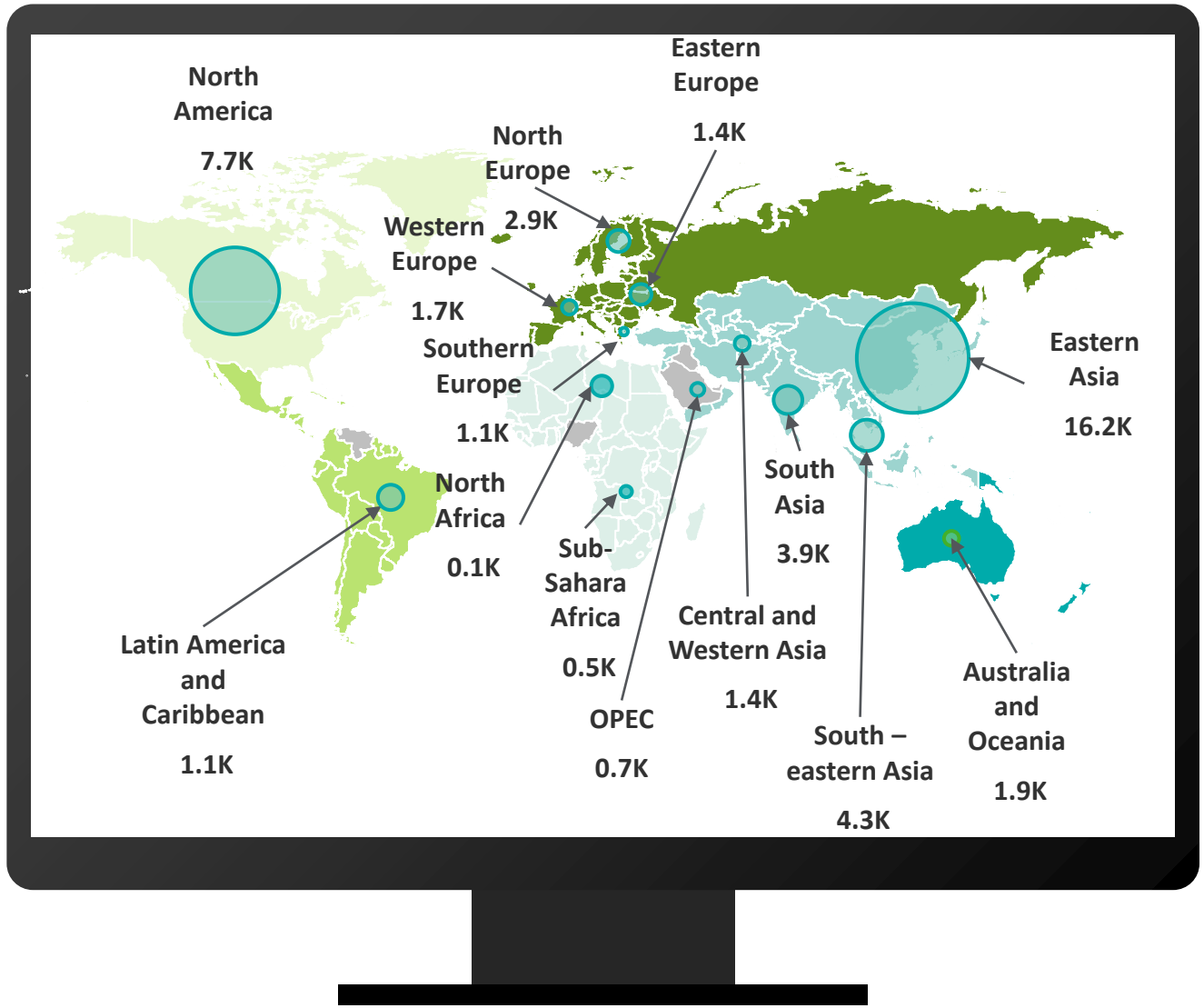


Data overview by indicator type and region

Indicator data table and regional data overview in 2020

Fiscal Year	Ind = 0	Ind = 1	Ind = 2	Total
2002	33 691	751	130	34 572
2003	35 377	724	168	36 269
2004	36 540	1 355	320	38 215
2005	39 380	1 523	564	41 467
2006	41 287	1 388	716	43 391
2007	42 267	1 337	935	44 539
2008	41 965	1 650	1 091	44 706
2009	41 439	1 708	1 447	44 594
2010	40 666	1 995	1 758	44 419
2011	40 501	1 976	1 881	44 358
2012	40 179	1 929	1 999	44 107
2013	39 153	1 970	2 040	43 163
2014	39 097	1 988	2 129	43 214
2015	38 515	2 528	2 371	43 414
2016	38 125	3 139	2 587	43 851
2017	37 864	3 717	2 960	44 541
2018	38 026	3 896	3 500	45 422
2019	37 261	4 218	4 165	45 644
2020	35 323	4 768	4 726	44 817
2021	35 869	3 899	3 688	43 456

Table 1: Indicator type distribution by fiscal year



ESG and carbon disclosure: Research setup

Overview of random forest regression

Model Setup

$$Y = f(X, W, Z) + \varepsilon$$

Dependent variable (Y)

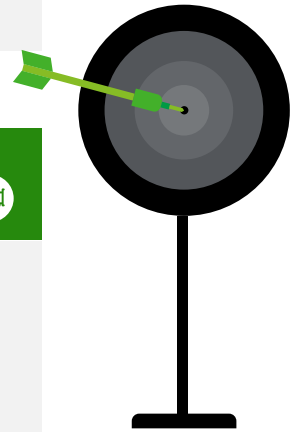
- Disclosure type:
  - non disclosure
  - partial (carbon) disclosure
  - ESG data disclosure

Fundamental variables (X)

- Size variable:
  - Total Revenues,
  - Total Assets
- Profitability variable
  - Operating Income
  - Capital Intensity (CAPEX)

Macroeconomic variables (W)

- Gross Domestic Product Purchasing Power Parity per capita (GDP PPP pc):
  - proxy for the state of development of the country



Sectorial and geographical decomposition variables (Z)

- Classification based on the TRBC four digit codes,
  - added distinction between carbon intensive and non-intensive sectors.
- Region and country of company headquarters.

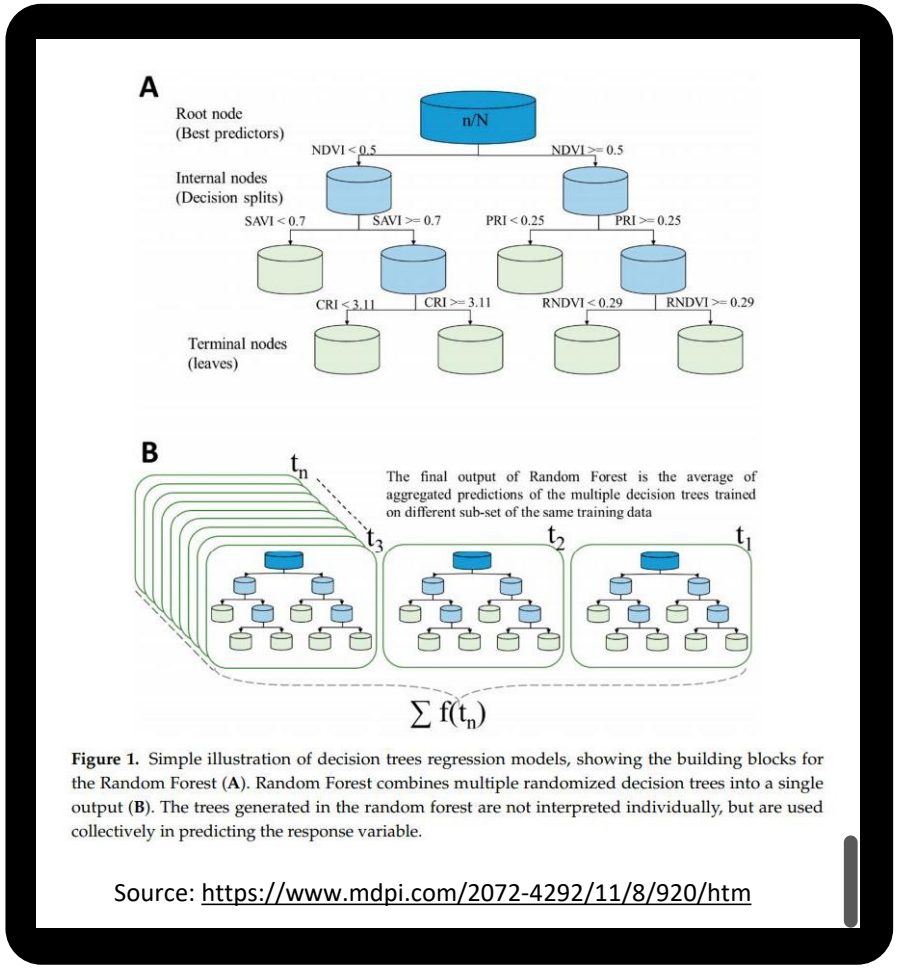


Figure 1. Simple illustration of decision trees regression models, showing the building blocks for the Random Forest (**A**). Random Forest combines multiple randomized decision trees into a single output (**B**). The trees generated in the random forest are not interpreted individually, but are used collectively in predicting the response variable.

Source: https://www.mdpi.com/2072-4292/11/8/920/htm

Methodology Overview

In order to extend the model from the two-class logistic regression, the model is designed to select one of the response classes as the baseline and represent all other classes in relation to the baseline.

For the k-th class of the model, the model can be written as (see James et al. (2013)):

$$Pr(Y = k|X = x) = \frac{e^{\beta_{k0}+\beta_{k1}x_1+...+\beta_{kp}x_p}}{1 + \sum_{l=1}^{K-1} e^{\beta_{l0}+\beta_{l1}x_1+...+\beta_{lp}x_p}}$$

For $k = 1, k = 2, \ldots, k = K - 1$ and:

$$Pr(Y = K|X = x) = \frac{1}{1 + \sum_{l=1}^{K-1} e^{\beta_{l0}+\beta_{l1}x_1+...+\beta_{lp}x_p}}$$

Where log odds can be written as:

$$log(\frac{Pr(Y = k|X = x)}{Pr(Y = K|X = x)}) = \beta_{k0} + \beta_{k1}x_1 + \ldots + \beta_{kp}x_p$$

One of the main advantages that it bears is its computational inexpensiveness, which renders the iterative optimization ("training") procedure required to calibrate the parameters very efficient.

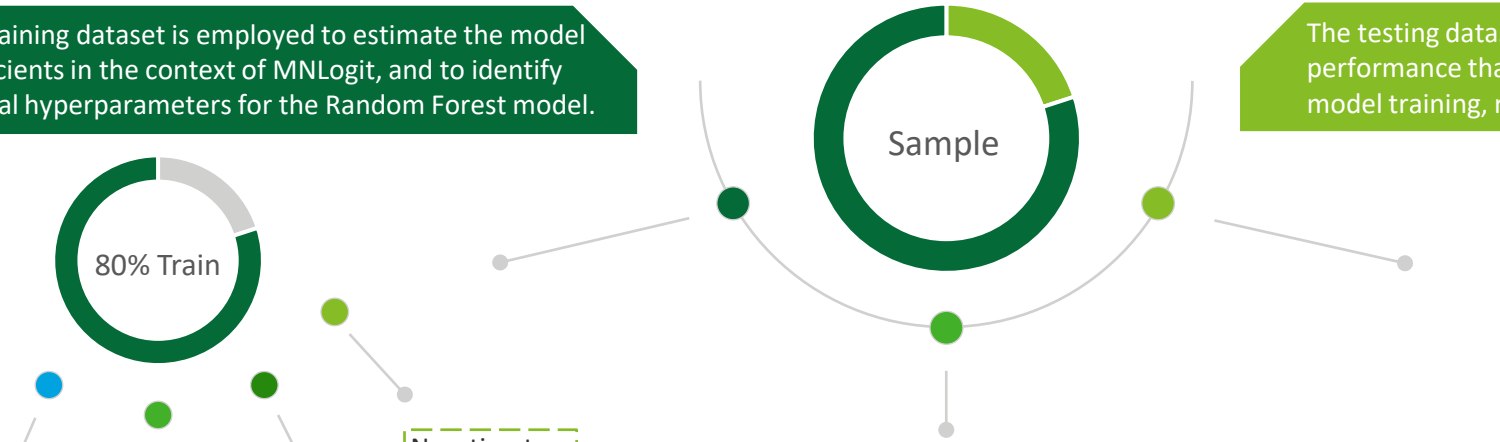
In random forest regression, prediction of random are made with the following three-step procedure steps:

1. The algorithm divides the set of possible values of predictors into distinct non-overlapping regions, $R_1, R_2, \ldots, R_J$
2. For every observation that falls in the region R_J, the predicted value of the response variable is equated with its within-region mean.
3. We grow N trees, considering only a subset of variables for the construction of each tree, and finally average over predictions made by N trees grown.

In each of the N rounds, tree construction is based on subset of predictors that have been chosen randomly from the overall number of predictors. Successive splits are applied in a way that minimizes the Gini impurity.

Data sampling

The training dataset is employed to estimate the model coefficients in the context of MNLogit, and to identify optimal hyperparameters for the Random Forest model.



The testing dataset is utilized to assess the model's performance that is independent of the data used for model training, referred to as out-of-sample evaluation.

We divide the total sample, consisting of more than 850,000 data points, in train and test sets in 80-20 proportions applying stratified sampling, to ensure a proportional representation of all three disclosure indicators.

- Depth: 20
- Maximum features: log2
- Min obs. in a leaf: 3
- Min sample for split: 4
- N. estimators: 1800

\+ Randomized search applied on a more than 1200 hyperparameter combinations

MNLogit

- Results are easily understandable.
- It is computationally inexpensive, which renders the iterative optimization ("training") procedure required to calibrate the parameters very efficient.

MNLogit Regression Results

Dep. Variable:	Indicator	No. Observations:	686527
Model:	MNLogit	Df Residuals:	686389
Method:	MLE	Df Model:	136
Date:	Mon, 23 Jan 2023	Pseudo R-squ.:	0.2910
Time:	18:34:59	Log-Likelihood:	-1.9151e+05
converged:	True	LL-Null:	-2.7010e+05
Covariance Type:	nonrobust	LLR p-value:	0.000

- High frequency of misclassification for partial and GHG disclosure.
- Low predictive accuracy.
- The precision of the "no disclosure" response type is high (92%), but this does not indicate a good model performance since the dataset is highly unbalanced.
- The recall measure shows a very low result 8% for the partial disclosure and only 29% for the GHG disclosure.

- Much better performance than MNLogit.
- Large percentage of correct predictions.
- The precision exceeds 90% for all variables, which is large improvement.
- Substantial improvement in the recall measure. 76% for ESG disclosure and 87% for GHG disclosure.

Multinomial Logit results summary and brief description

	ESG disclosure	GHG disclosure
Continuous Variables		
Total Revenues	1.899***	2.101***
	(0.020)	(0.020)
Total Assets	−0.166***	−0.166***
	(0.014)	(0.014)
Operating Income	1.061***	1.215***
	(0.020)	(0.020)
CAPEX	0.174***	0.179***
	(0.013)	(0.014)
GDP PPP PC	0.216***	0.327***
	(0.009)	(0.010)
Region Dummy Variables		
OPEC	−0.978***	−2.491***
	(0.072)	(0.118)
Eastern Asia	−0.654***	−1.122***
	(0.029)	(0.029)
Eastern Europe	−1.086***	−1.680***
	(0.061)	(0.065)
Northern Europe	0.176***	0.727***
	(0.033)	(0.029)
Southern Europe	−0.190***	0.148***
	(0.048)	(0.042)
Northern America	0.910***	−0.486***
	(0.027)	(0.028)

Sector Dummy Variables		
Banking & Investment Services	1.190***	0.233***
	(0.030)	(0.036)
Energy - Fossil Fuels	0.518***	0.202***
	(0.035)	(0.039)
Renewable Energy	0.194**	0.221**
	(0.094)	(0.091)
Renewable Utilities	0.300**	−0.086
	(0.119)	(0.124)
Uranium	0.372***	0.159
	(0.133)	(0.182)
Utilities	0.330***	0.834***
	(0.047)	(0.043)
Year Dummy Variables		
2003	−0.137*	0.000
	(0.071)	(0.167)
2004	0.703***	1.174***
	(0.061)	(0.145)
2020	2.521***	4.821***
	(0.054)	(0.126)
2021	2.136***	4.293***
	(0.056)	(0.127)

Company Size

Larger firms are more likely to disclose GHG emissions and ESG data due to resources and incentives.

Operating income

Profitability has a significant positive effect on the disclosure likelihood of both ESG and GHG emissions information.

Capex

Capital expenditures have a modest yet significant impact on disclosures.

Sector

Utility companies tend to disclose more readily than the nuclear and renewable utility industries.
Fossil fuel firms tend to disclose ESG data more readily than GHG emissions data, mainly due to the governance and social aspects

Region

Companies in countries with higher economic purchasing power are more likely to disclose their ESG and GHG emissions
Companies incorporated in OPEC countries show the lowest likelihood of disclosing information on ESG and GHG emissions

Result Random forest

For every continuous and discrete variable predictions are analyzed

Case continuous variables

01 STEP

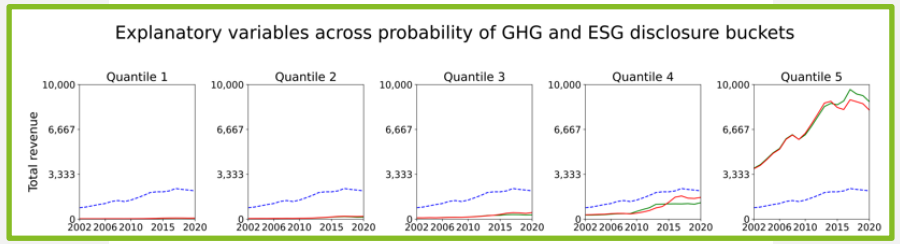
For each year in the sample probabilities of GHG and ESG disclosure are sliced in quintiles (in order to obtain 5 equally populated buckets)

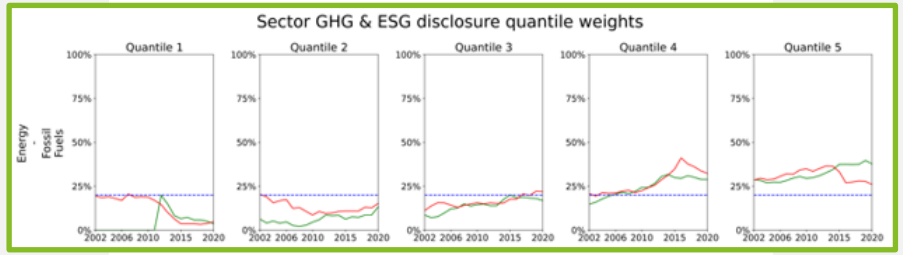
02 STEP

For each of the bucket sample average is calculated

03 STEP

We compare bucket results with yearly averages





Case discrete variables

For each year in the sample probabilities of GHG and ESG disclosure are sliced in quintiles (in order to obtain 5 equally populated buckets)

01 STEP

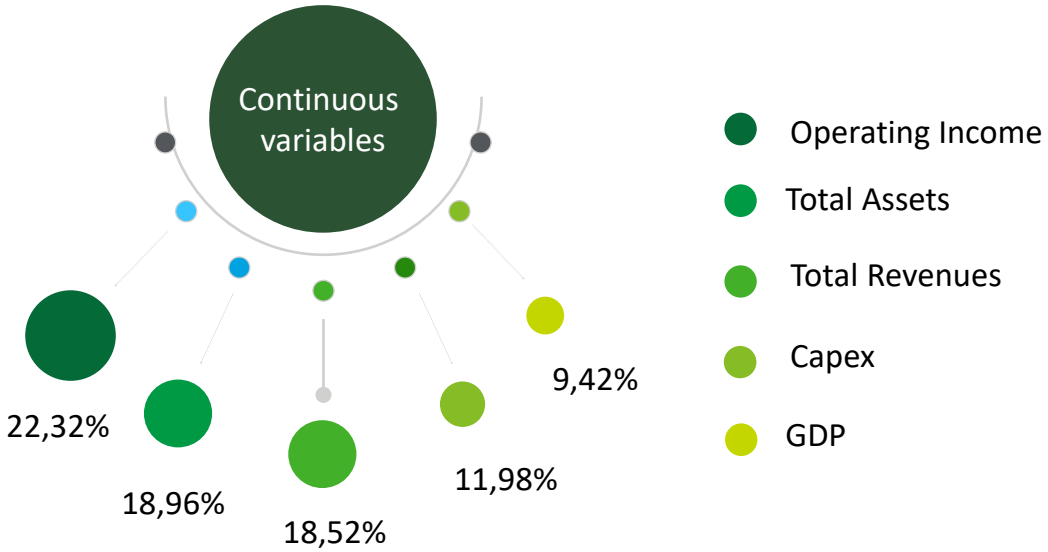
For each bucket, number of companies is counted

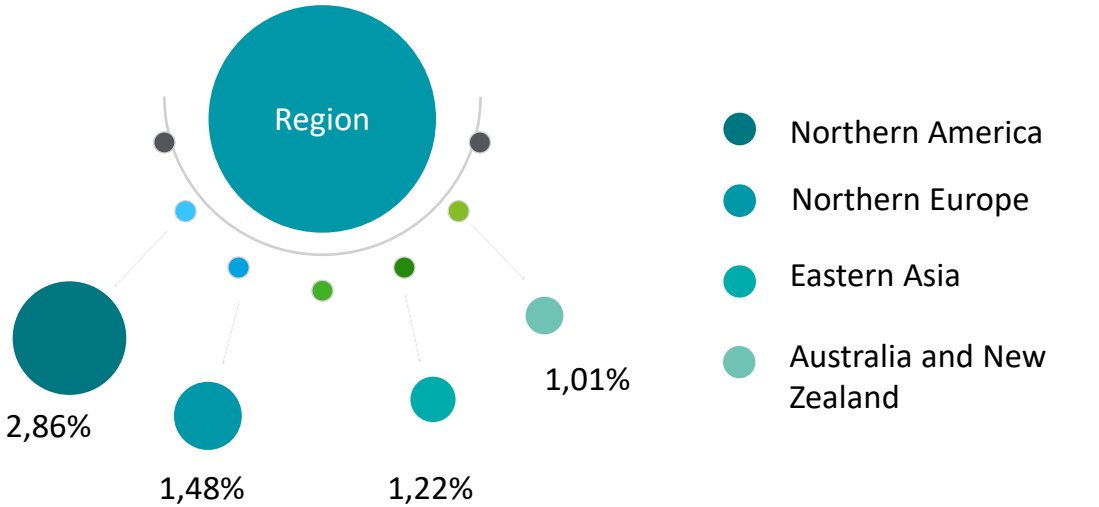
02 STEP

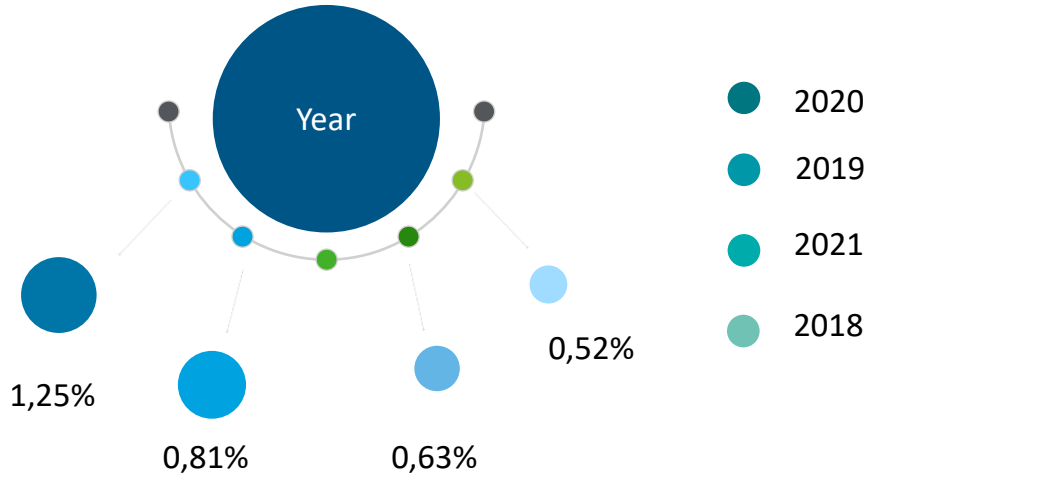
We compare our results with the uniform distribution expectation of 20%

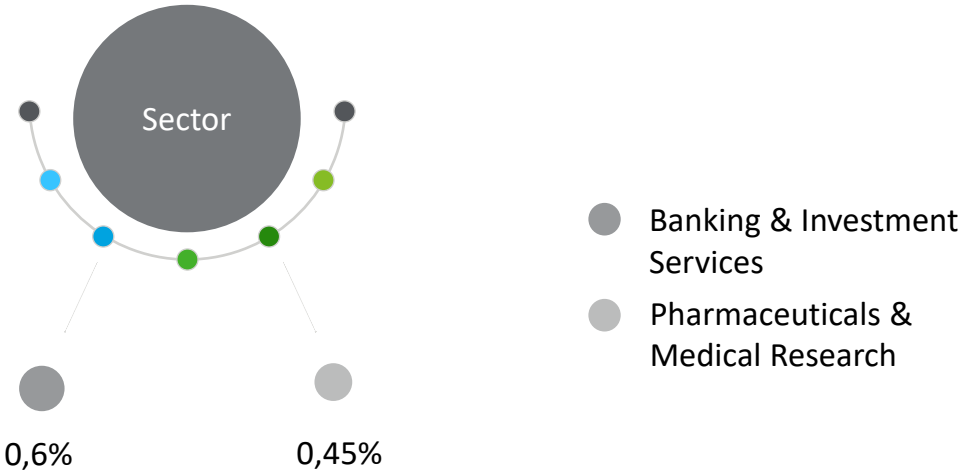
03 STEP

Random Forest variable importance



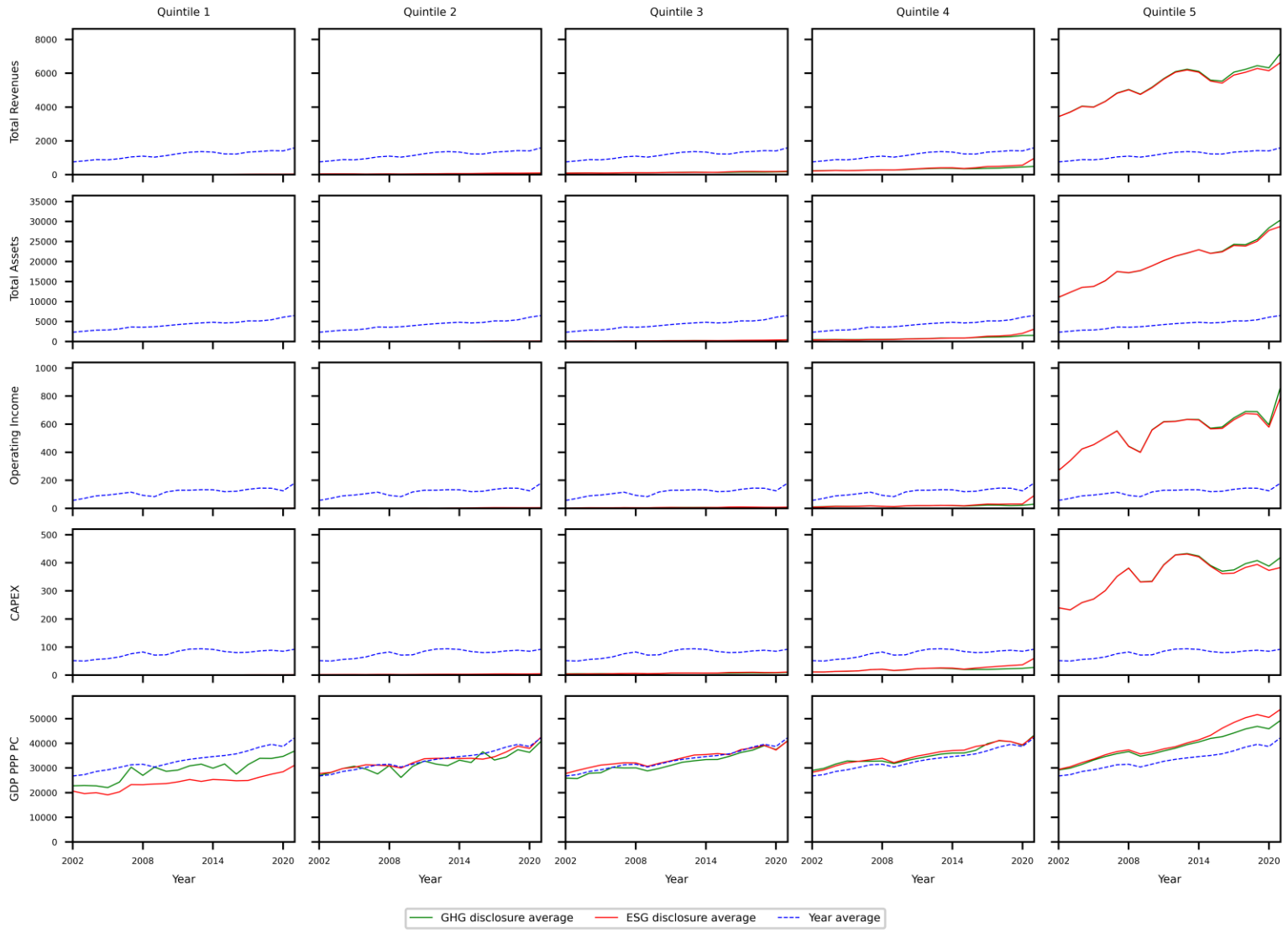






Size variables and GDP

Explanatory variables across probability of GHG and ESG disclosure buckets



Size variables show similar pattern:

- First quantile of the ESG and GHG likelihood of disclosure show very small average of the size variables.
- For every quantile of the ESG and GHG likelihood of disclosure size variables are showing higher average values.
- In the Quantile 5, size variables are significantly larger than the average values for the years.

GDP PPP PC variable show similar pattern, companies that are in the higher probability of disclosure bucket are on average incorporated in countries with higher GDP.

Sector analysis

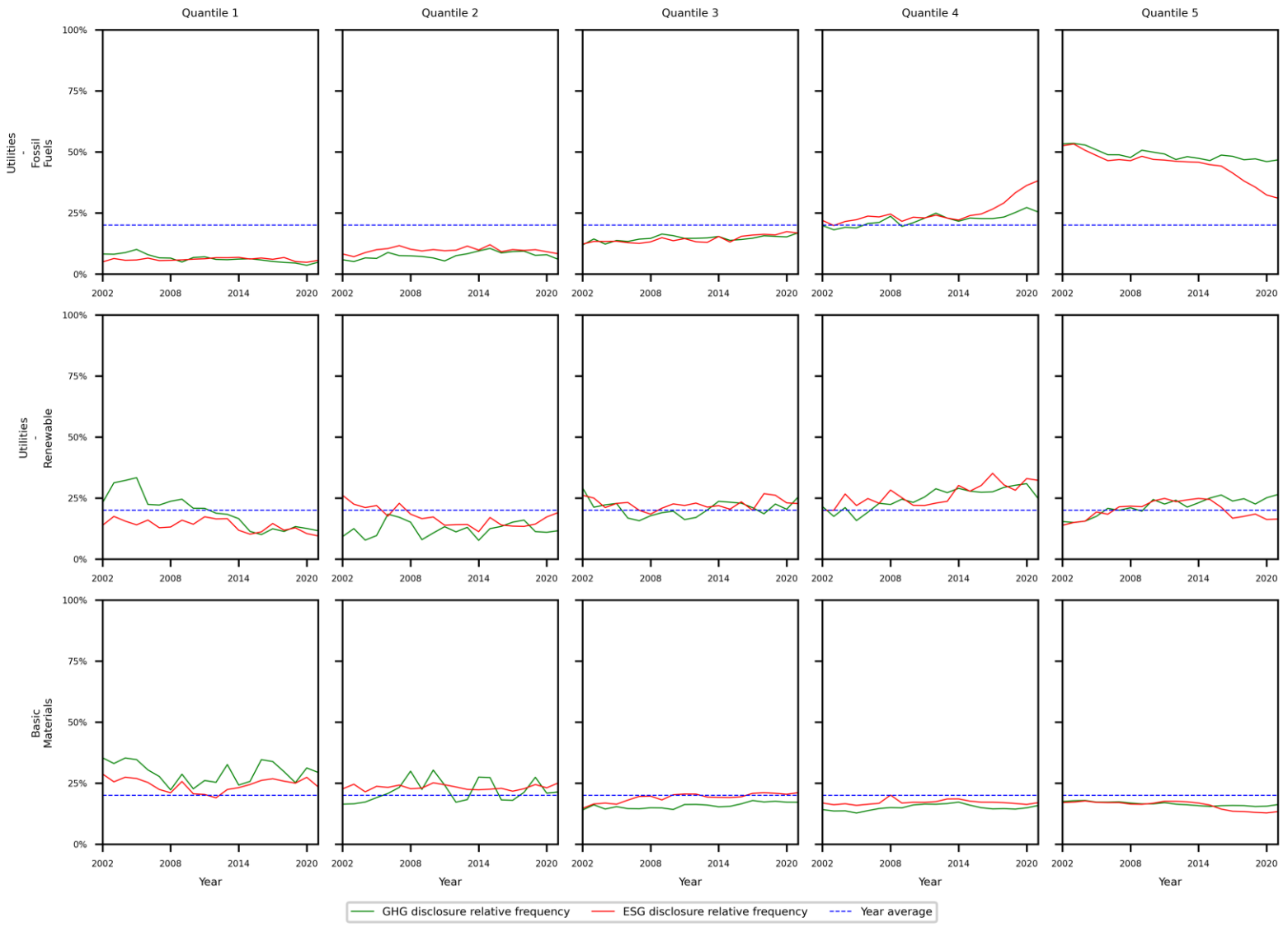
Carbon intensive sectors (ESG and GHG) Q1-Q5:

Sectorial findings:

- The study focuses on 31 business sectors, with greater emphasis on the carbon-intensive energy and utility sectors
- This finding indicates public pressure on the carbon more intensive companies. Results support legitimacy theory.
- Companies belonging to nuclear energy show high ESG disclosure and high fluctuation across probability of disclosure buckets due to low total number of companies.

Sector analysis

Carbon intensive sectors (ESG and GHG) Q1-Q5:

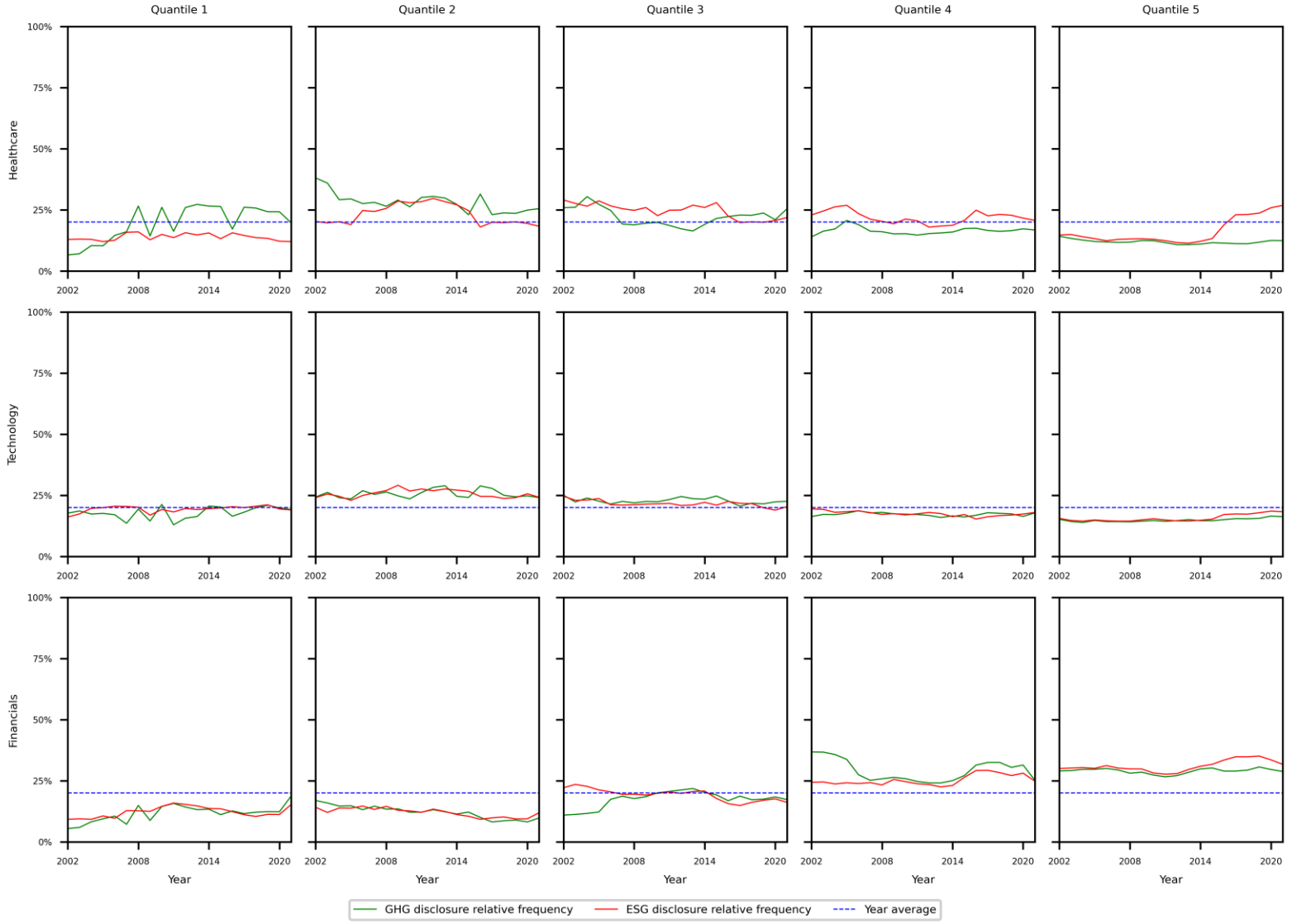


Sectorial findings:

- Traditional utilities (fossil fuel) show very high likelihood to disclose especially in the early years.
- Although utilities are one of the most polluting sectors, regulation and public interest results in high disclosure likelihood.
- Traditional utilities disclose GHG data more readily than ESG data show by the gap in the likelihood of disclosure in recent years.
- Renewable utilities show much lower propensity to disclose compared to their traditional counterparts.
- Basic materials do not deviate too much from the naïve expectations.

Sector analysis

Carbon less-intensive sectors (ESG and GHG) Q1-Q5:

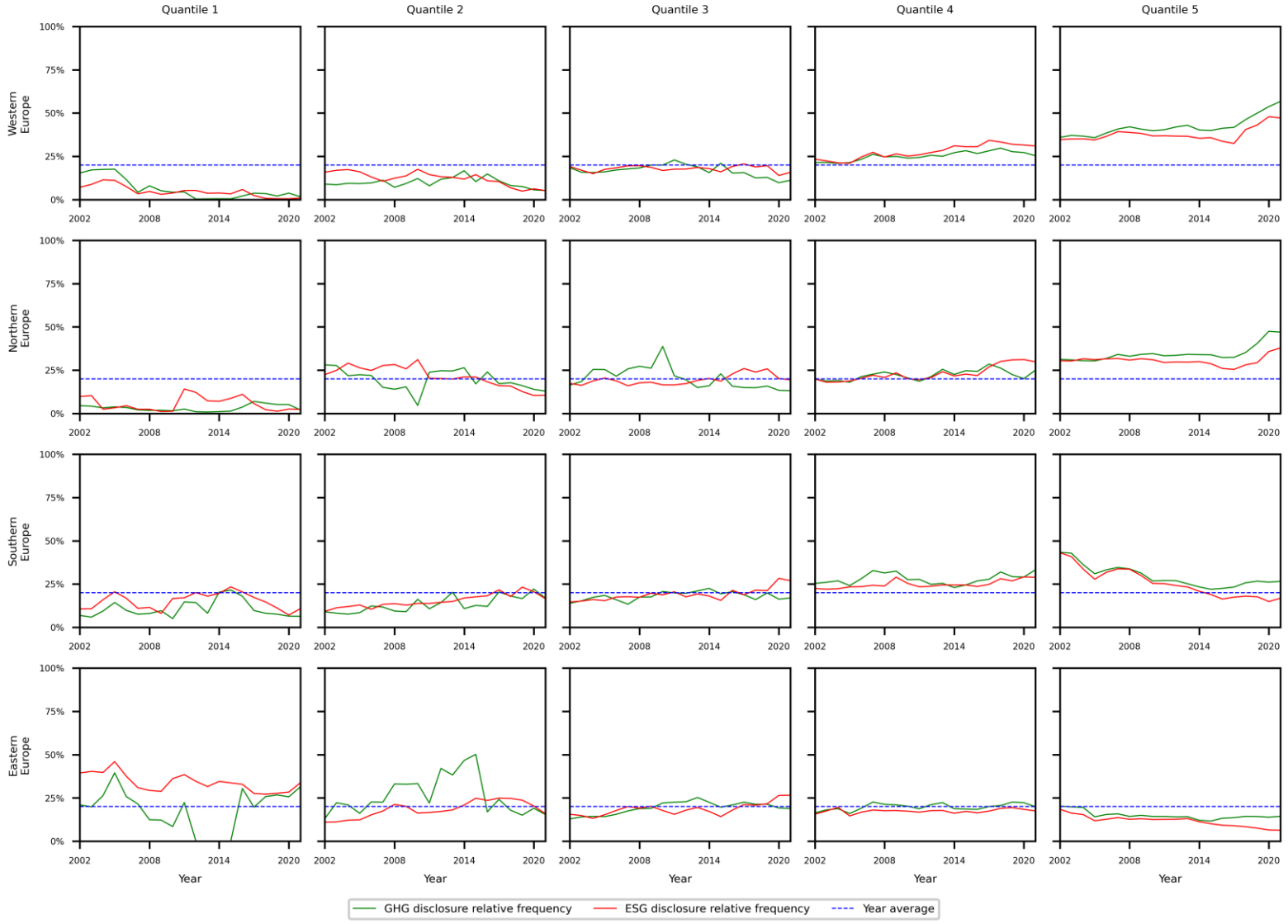


Sectorial findings:

- Financials sector shows high likelihood of disclosure by having disproportionately many companies represented in Q4 and Q5.
- Financials companies put more emphasis in ESG disclosure than in the GHG disclosure.
- Healthcare sector has increased likelihood of ESG disclosure in the recent years.
- Technology companies do not show significant disclosure likelihood.

Region & country analysis

Europe

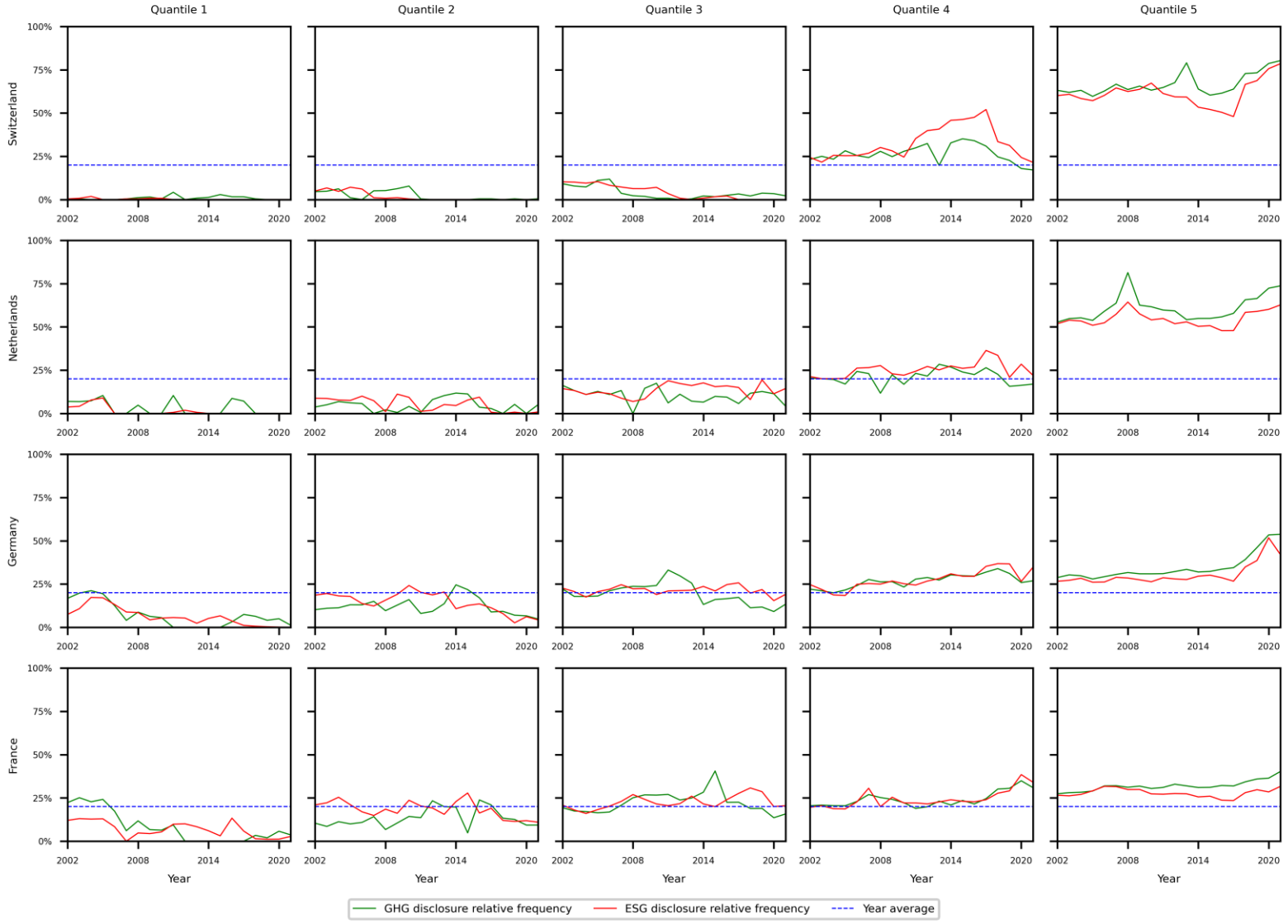


Regional findings:

- Western and Northern European companies show much higher likelihood to disclose GHG and ESG data.
- Emphasis is put on GHG data that is mainly above the ESG disclosure likelihood curve.
- Southern European countries follow their western and northern neighbors in high disclosure likelihood
- Eastern European companies lack in the disclosure.

Region & country analysis

Europe

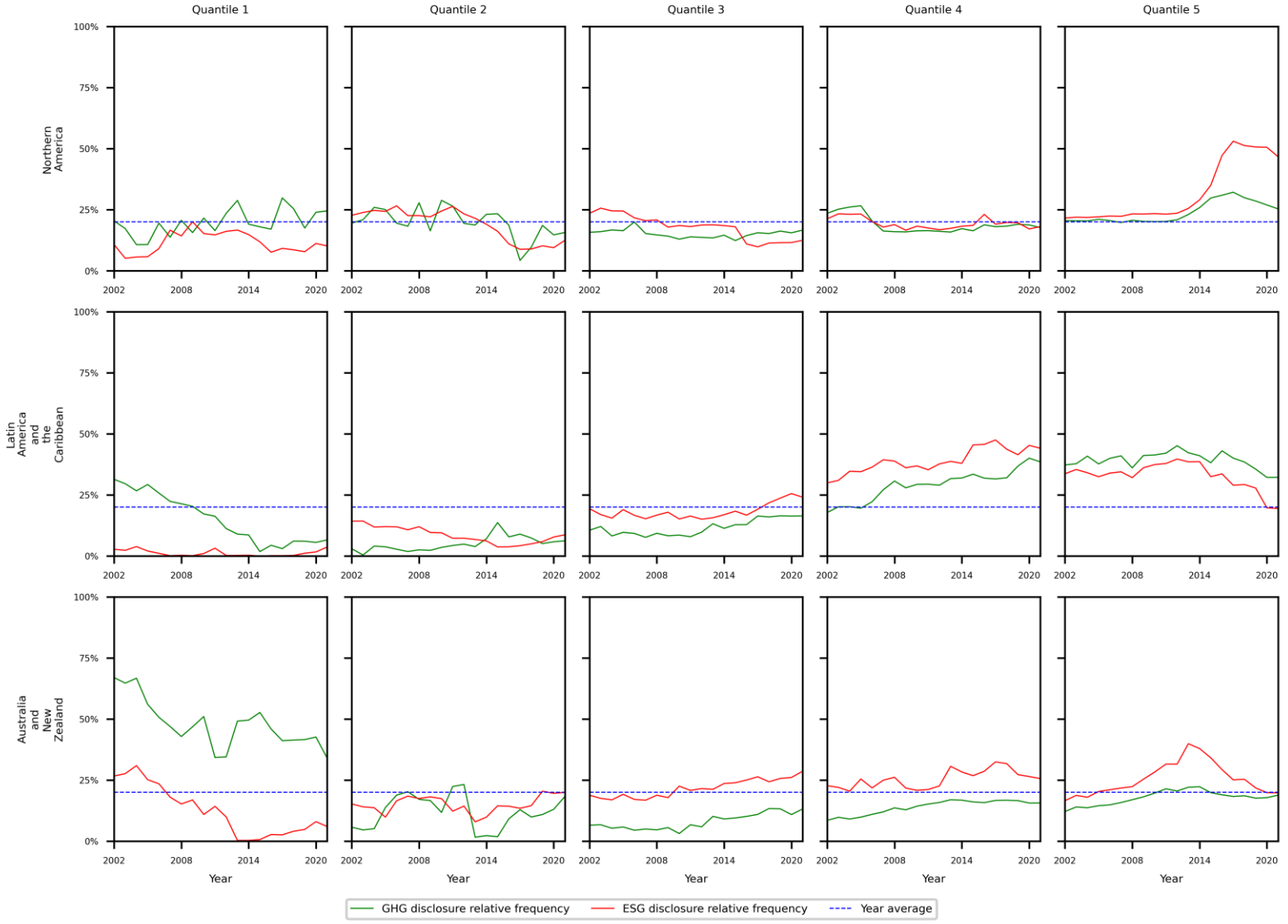


Country findings

- Selected countries from Western Europe show high disclosure likelihood.
- Switzerland is the champion of disclosure, with more than 90% of all companies represented in the Q4 or Q5 bucket of disclosure in recent years.
- Netherlands also show very high degree of disclosure
- Germany and France show higher than average disclosure likelihood, but are lacking behind compared to the mentioned countries.

Region & country analysis

America & Oceania

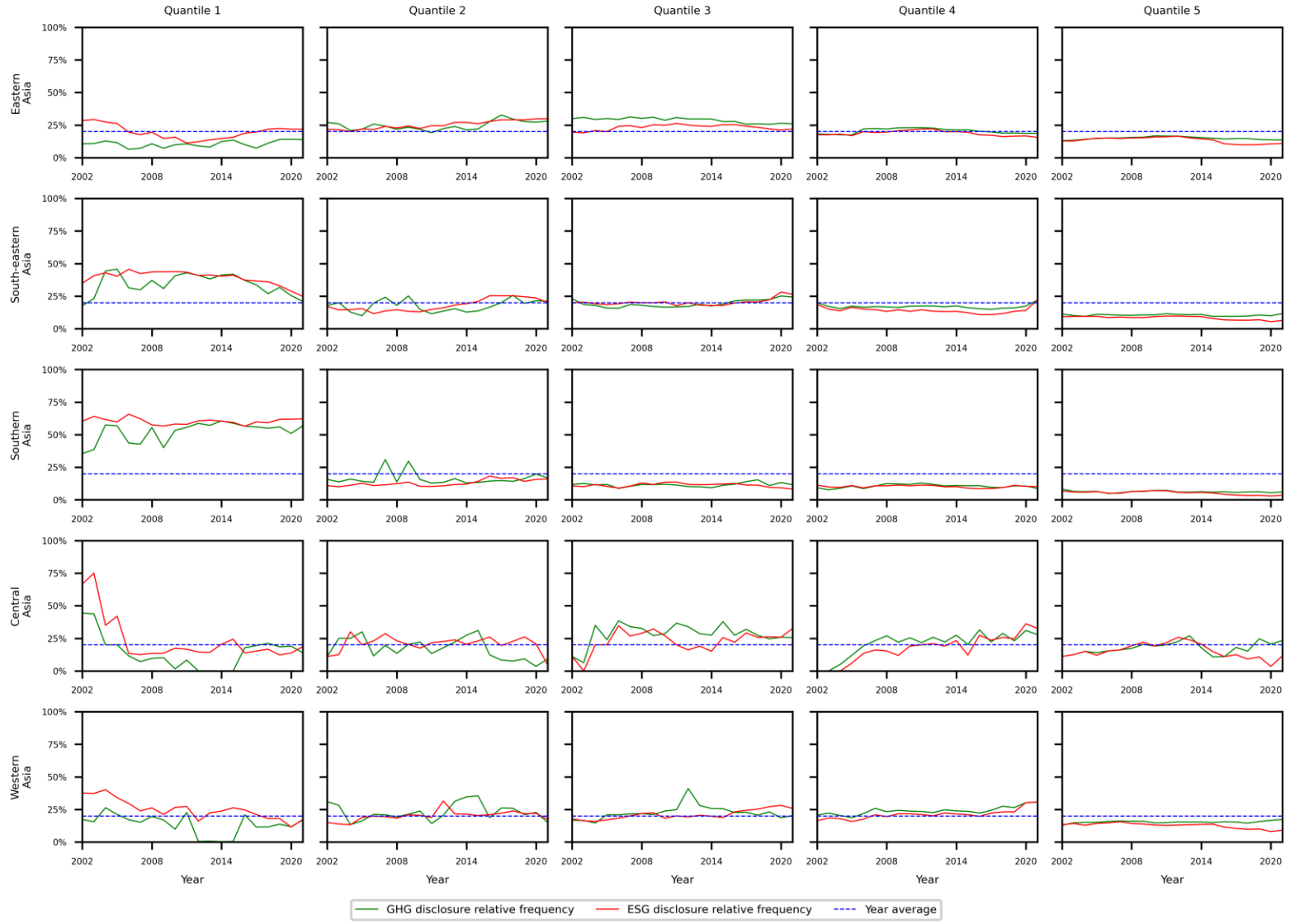


Regional findings:

- Companies from North America show higher than average disclosure likelihood.
- It is obvious that emphasis on the disclosure in the Northern America is put on the ESG in comparison with European companies that are focused more on the GHG reporting.
- Companies from Oceania also show high propensity to disclose, with ESG being dominant in the most of the years.
- Latin American companies disclose more than average, GHG disclosures more dominant than ESG.

Region & country analysis

Asia

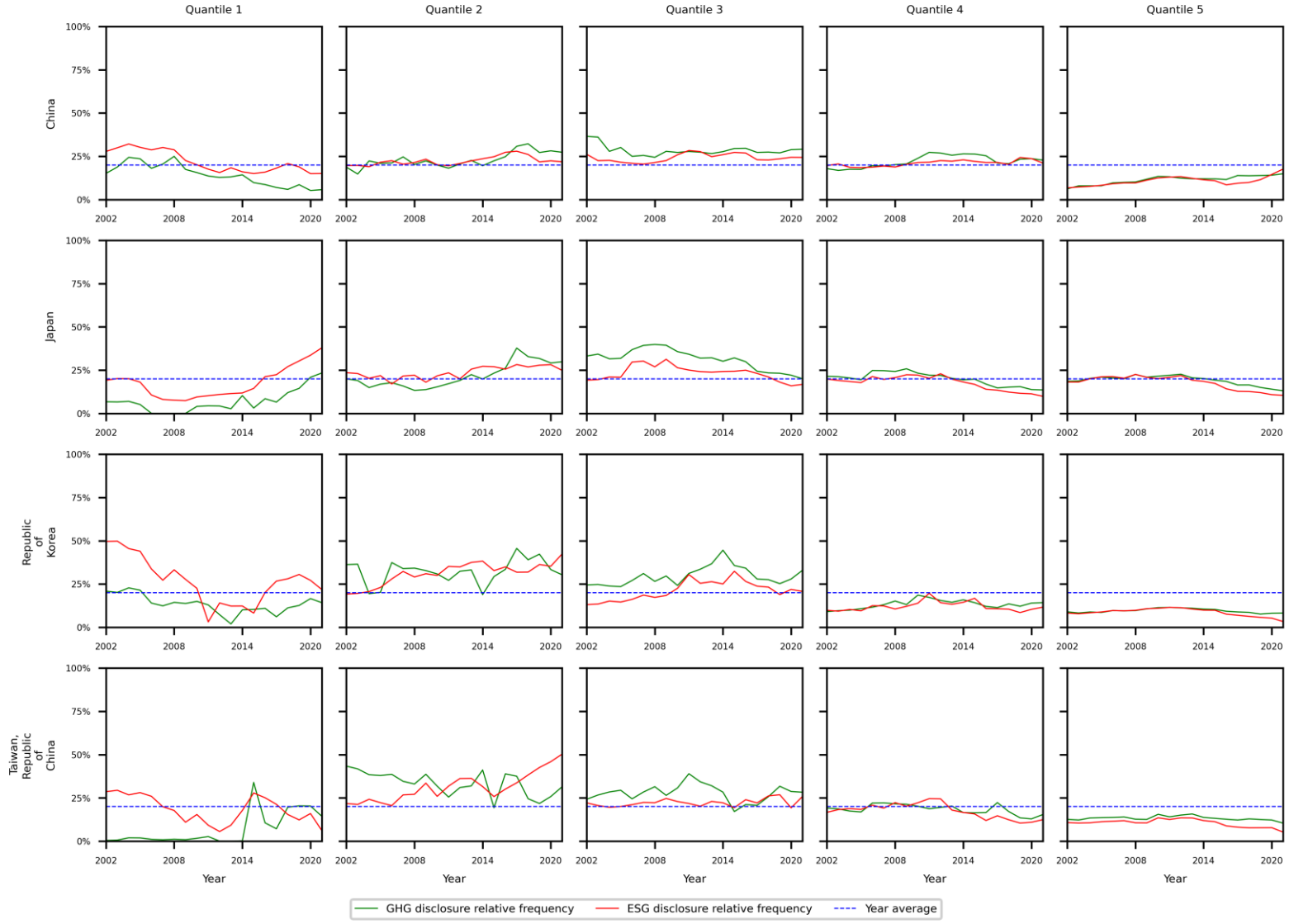


Regional findings:

- Companies from all regions of Asia disclose on average less than the expectation.
- The likelihood of disclosure in Eastern Asia (the biggest sample) is very low, more than 80% of companies are in the first 3 buckets of disclosure.
- The situation is not better in other Asian regions as well, South-Eastern and Southern Asia companies have on average very low disclosure probability.
- This situation will certainly improve with introduction of mandatory disclosure (e.g. Japan 2022).

Region & country analysis

Asia



Eastern Asian companies show very low disclosure likelihood

- Only 5% of Chinese companies are in the quantile 5 of disclosure likelihood.
- Similarly, only 7% Japanese companies are in the most likely to disclosure bucket. The reason can be the big overall sample of Japanese companies and selective disclosure policies that only export oriented companies are fulfilling.

Model performance comparison - Out of sample

MNLogit

Confusion matrix and Classification report

	No disclosure	ESG disclosure	GHG disclosure
No disclosure	153 588	394	523
ESG disclosure	8 089	767	436
GHG disclosure	4 799	761	2 275

	Precision	Recall	F1-score	Support
No disclosure	0.92	0.99	0.96	154 505
ESG disclosure	0.40	0.08	0.14	9 292
GHG disclosure	0.70	0.29	0.41	7 835
Accuracy			0.91	171 632
Macro avg.	0.68	0.46	0.50	171 632
Weighted avg.	0.88	0.91	0.89	171 632

Random Forest

Confusion matrix and Classification report

	No disclosure	ESG disclosure	GHG disclosure
No disclosure	154 090	197	218
ESG disclosure	1 931	7 101	260
GHG disclosure	803	180	6 852

	Precision	Recall	F1-score	Support
No disclosure	0.98	1.00	0.99	154 505
ESG disclosure	0.95	0.76	0.85	9 292
GHG disclosure	0.93	0.87	0.90	7 835
Accuracy			0.98	171 632
Macro avg.	0.96	0.88	0.91	171 632
Weighted avg.	0.98	0.98	0.98	171 632

Results

The findings suggest that the Random Forest model outperforms the Multinomial Logit model in out-of-sample prediction accuracy

Summary of main findings

Disclosure

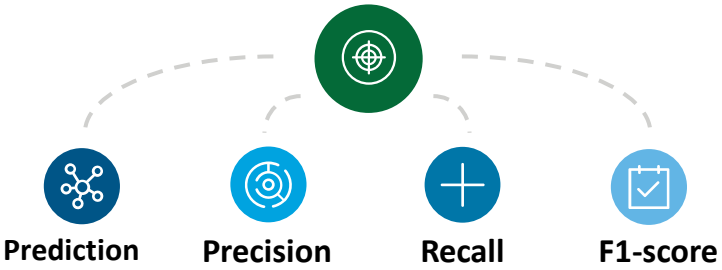
- What are significant differences and trends to which ESG and GHG emission data are disclosed?
- Do carbon more intensive sectors disclose more readily information on GHG emissions and ESG data than their less intensive inter sector competitors?

- Companies with high GHG emissions and significant potential for environmental harm are more likely to disclose their emissions data.
- This finding is consistent with legitimacy theory, which suggests that companies disclose information in response to social and regulatory pressures.
- There is significant public and regulatory pressure on more GHG-intensive companies to disclose their emissions data.

Methodology

- What is the difference between traditional Multinomial Logit model against more flexible Machine learning approach i.e. Random Forest.

Random Forest Outperforms in predictive power Multinomial Logit



Can we identify blind spots in the financial asset portfolios?

- The model estimates disclosure likelihood of companies and can serve purpose for more transparent reporting and portfolio management process
- As assets drop out of sample, companies owning the assets will appear in the books and their disclosure history can be easily assessed

Appendix

Other results

Overview of various Refinitiv data available

Snapshot of currently used data for the purpose of this study

	Asset 4	Datastream	Worldscope	Thompson Reuters Fundamentals	Fixed income EJV data	CDS data
Data type	ESG data Company controversies	Equity prices and indices Futures FX data Commodities Macroeconomic data	Basic company information: Identifiers, industry classification, country of incorporation Company fundamental data: Net Sales, Total Assets, industry specific metrics, etc. Fundamental ratios: Profitability ratios, liquidity ratios, etc. Segment revenue decomposition	Basic company information: Identifiers, Company industry information, NAICS, TRBC Company fundamental data: Net sales, industry specific metrics, etc. Segment revenue decomposition: Regional and sectorial coverage	Interest rates Bond prices / yields Bond static data Benchmark curves	CDS spreads
Coverage	>12,000 companies, 424 active items, 13 score categories Carbon Emission data available for: > 4300 comp. since 2010, > 3000 comp. since 2017	Equity prices: >107.000 companies Index data: > 285.000 indices FX data: > 9.000 pairs Commodities: > 105.000 assets Macroeconomic data: > 8.9 mil. Series	> 89.000 companies > 800 Items Segment data linked to SIC code	> 84.500 companies > 1.116 Items Segment data linked to NAICS code	> 500.000 bonds and convertibles > 20.000 interest rates > 800 benchmark curves	> 96.000 instruments
Time series update frequency	Annual data available Update cycle: biweekly	Daily data available Macroeconomic data: monthly/quarterly Update cycle: daily	Quarterly/Annual data available Update cycle: daily	Quarterly/Annual data available Update cycle: daily	Daily data available Update cycle: daily	Daily data available Update cycle: daily
Application	Emission data -> basis for carbon exposure index calculation ESG scores used to rank companies	Fx data: convert data to common currency Macroeconomic data: used in carbon exposure index Equity and index prices and returns used as response variables to test market significance of the index	Fundamental data and ratios used to estimate carbon exposure index Segment decomposition used to segregate and quantify sector exposure on company level	Fundamental data and ratios used to estimate carbon exposure index Segment decomposition used to segregate and quantify sector exposure on company level	Term structure of interest rates Quantification of effect of carbon exposure index on bond yields	Used as a response variable and as a measure of carbon risk market impact

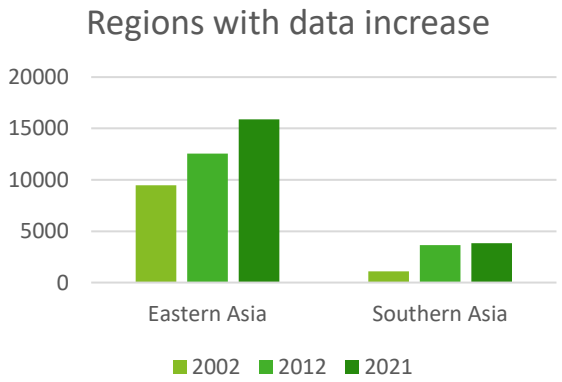
Inclusion of carbon and ESG data for quantitative portfolio analysis is essential in making informed decisions and sustainable investing.

Data overview

Sample imbalance, increase in number of disclosing companies, overall data increase

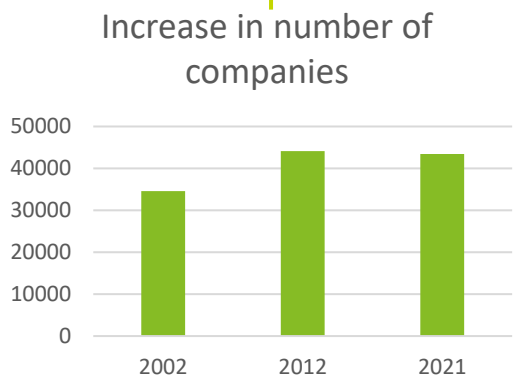
Number of companies by region

- From 2002 until 2020 there is a huge increase of number of companies covered in the sample in Eastern and Southern Asia.
- During the same time, revenue in Eastern Asia increased by 142% and 124% in Southern.



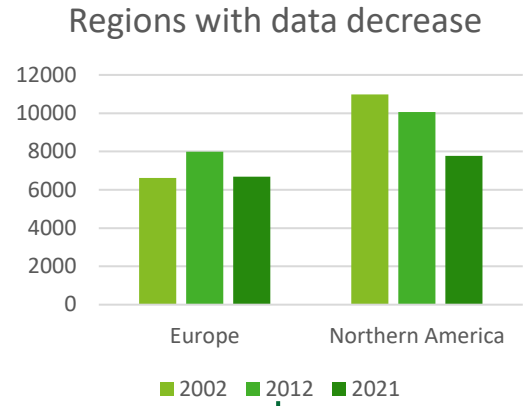
Number of companies by region

- Total number of companies covered in the sample has decreased in Western Europe and Northern America from 2002 until 2020.
- In Northern America average revenue increased by 160%, and 95% in Western Europe.



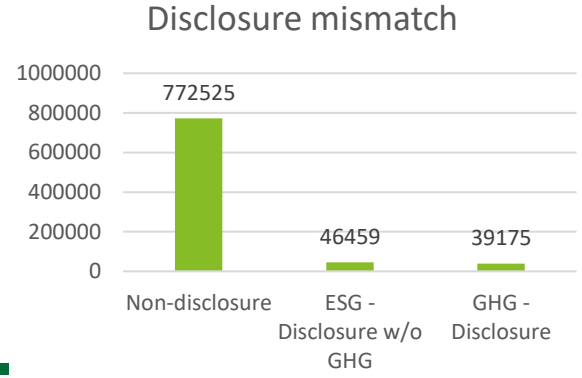
Total number of companies

- 25% increase from 2002 until 2020.
- Much higher increase of the disclosing companies, both ESG and GHG.



Disclosure imbalance

- At every observation year number of disclosing companies is by far lower than the number of non-disclosing firms creating an unbalanced panel.



Carbon emission data disclosure overview

Total data availability: total and country of incorporation decomposition

Fiscal Year	Number of companies	ESG score	CO2e E Total	CO2e E Direct	CO2e E Indirect	CO2e E Scope 3	Derived Emission	Emission Intensity
2002	924	903	135	67	29	1	135	133
2003	935	918	176	89	49	0	176	173
2004	1734	1706	331	154	87	2	331	326
2005	2139	2109	576	325	213	19	577	568
2006	2150	2130	726	470	384	47	726	719
2007	2321	2302	948	534	453	243	948	938
2008	2792	2777	1110	648	561	366	1110	1096
2009	3201	3191	1462	1112	1037	708	1462	1451
2010	3814	3807	1771	1440	1391	988	1771	1758
2011	3906	3904	1895	1551	1515	1110	1895	1881
2012	3975	3973	2017	1659	1638	1212	2017	2002
2013	4080	4076	2059	1626	1609	1057	2060	2041
2014	4192	4192	2142	1681	1668	1021	2142	2125
2015	4955	4955	2378	1914	1900	1122	2378	2362
2016	5791	5791	2593	2130	2110	1261	2594	2579
2017	6569	6569	2948	2482	2457	1417	2952	2935
2018	7295	7295	3470	2992	2967	1674	3476	3458
2019	8387	8386	4026	3518	3496	2025	4034	4011
2020	9081	9072	4306	3795	3802	2251	4311	4262

Table 1: Refinitiv Asset 4 database data availability

- We have seen a constant increase in data availability in recent years: Data on a substantial number of companies are available for analysis
- CO2e total emissions represent the sum of Scope 1 and 2 emissions
- Data availability for Scope 3 emissions is limited and quality of available data is poor due to measurement and attribution problems and the complexity of dynamic input/output linkages.

Fiscal Year	East Asia & Pacific	Europe & C. Asia	L. America & Caribbean	M. East & N. Africa	North America	South Asia	Sub-Sah. Africa
2002	13	87			35		
2003	16	110			50		
2004	88	168	1		74		
2005	151	282	1		143		
2006	182	353	1		190		
2007	262	441	5		236	3	
2008	311	469	17	2	293	7	10
2009	383	587	29	4	430	13	15
2010	513	633	67	10	483	25	39
2011	535	669	72	12	523	28	55
2012	587	693	75	14	527	29	91
2013	625	736	77	14	479	39	89
2014	659	781	87	14	469	40	90
2015	732	844	102	21	544	44	89
2016	817	884	131	23	600	51	86
2017	995	951	149	26	687	49	93
2018	1110	1240	165	33	777	58	91
2019	1302	1383	189	35	961	66	96
2020	1345	1641	171	34	952	76	90

Table 2: Company disclosure by region of incorporation

- Company GHG Emission disclosure by regions is shown above
- Europe in absolute (and especially in relative) terms leads in the number of companies disclosing the data
- Research and analysis needed on the disclosure behavior across different characteristics of corporates: size, country, industry, etc.

Model calibration and parameter risk assessment

Calibration of multi class random forest classifier and discussion on the error matrices

Model calibration

Sample divided into 80-20 train and test 10 fold cross validation performed

Depth of the tree
Model trained not to overfit due to the large number of classes

Maximum features
How many variables are used at each tree split – decorrelating input

Minimum observations in a leaf
Balance between overfitting and underfitting

Minimum samples for split
Number of samples required to allow for additional split of the tree

Number of estimators
Number of trees grown

Randomized search applied on a more than 1200 hyperparameter combinations

Optimal parameters:

Depth: 20
Maximum features: log2
Min obs. in a leaf: 3
Min sample for split: 4
N. estimators: 1800

Out of sample result

Actual class / Predicted class

	No disclosure	ESG disclosure	GHG disclosure
No disclosure	100194	259	252
ESG disclosure	1464	5882	261
GHG disclosure	542	198	5718

	Precision	Recall	F1-score	Support
No disclosure	0.98	0.99	0.99	100705
ESG disclosure	0.93	0.77	0.84	7607
GHG disclosure	0.92	0.89	0.90	6458
accuracy			0.97	114770
macro avg	0.94	0.88	0.91	114770
weighted avg	0.97	0.97	0.97	114770

Confusion matrix

High prediction accuracy for out of sample measurement

Classification report

Slightly higher false negative case for the ESG disclosure case

Full sample result

Actual class / Predicted class

	No disclosure	ESG disclosure	GHG disclosure
No disclosure	500934	1663	928
ESG disclosure	6168	30652	1213
GHG disclosure	2044	1222	29024

	Precision	Recall	F1-score	Support
No disclosure	0.98	0.99	0.99	503525
ESG disclosure	0.91	0.81	0.86	38033
GHG disclosure	0.93	0.90	0.91	32290
accuracy			0.97	114770
macro avg	0.94	0.90	0.91	114770
weighted avg	0.98	0.98	0.97	114770

Confusion matrix

Full sample prediction accuracy as expected better than out of sample

Classification report

Overall very high accuracy with low false positive and false negative predictions

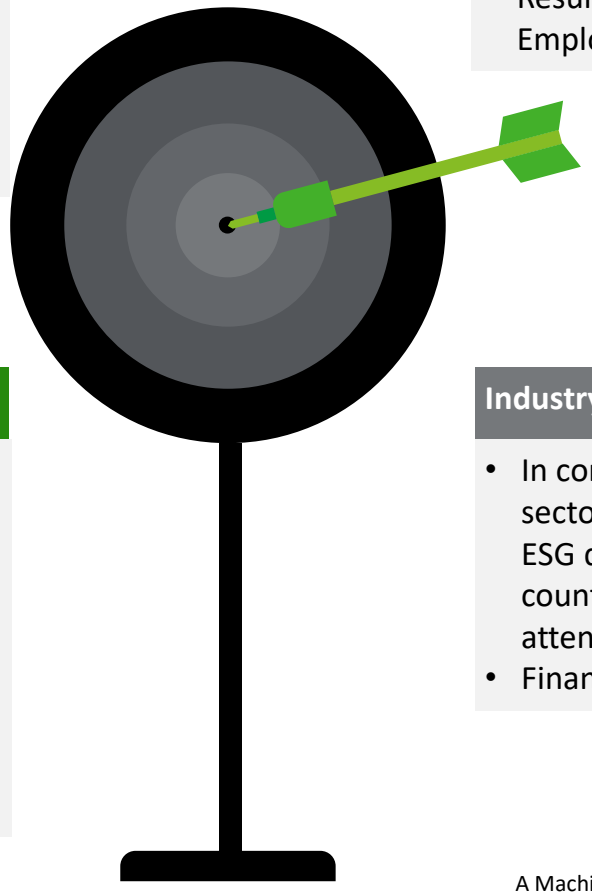
Summary of main results

Novel dataset

- To our knowledge this is a **most comprehensive study of ESG and climate disclosure**.
- In the extended model there are >550k data points from 2002 until 2020.
- Reduced form model consists of 900k data points and the main messages are consistent between both model

Size variables

- As expected, larger companies are more likely to disclose both ESG and GHG data
- Results consistent with Revenues, Total Assets and Employees



Regional variables

- Companies from Western, Northern and Southern Europe are much more likely to disclose ESG and GHG data. Emphasis put on GHG data
- Companies from North America (mostly the US) show high likelihood to disclose ESG data, indicating investor preference for ESG data over GHG.
- Asian region lacks in the disclosure, more regulation will push the regulation higher.

Industry variable

- In contrast to expectations, carbon more intensive sectors show higher likelihood to disclose GHG and ESG data compared with their renewable counterparts. Possible explanation is high public attention on the carbon intensive companies
- Financial sector mostly focused on ESG