Online Appendix

Appendix A – The machine learning process applied to companies' classification

To identify cleantech companies, the VICO dataset was enhanced by adding the extended textual business description of each VC-backed company. This information was collected from two sources, BVD Orbis and S&P Capital IQ datasets, to maximise the number of companies with an explicit business description. The extended business description is a standardised description of the company activity, written in the same language (English) for all companies. Out of the total VICO dataset that includes 46,966 deals related to GVC and IVC and 19,415 VC-backed companies between 1998 and 2014 in the 10 countries identified, we found an extended business description for 11,769 companies (60.06% of companies). The following tables show the distribution of the two subsamples along the different dimensions: Table A1 by country, Table A2 by industry and Table A3 by year of foundation.

Table A1 - Distribution of the sample by country

Country	VICO		VICO companies with business description	
	n	%	n	%
United Kingdom	5,185	26.71%	3,475	29.53%
France	3,426	17.65%	2,210	18.78%
Germany	2,804	14.44%	2,001	17.00%
Sweden	1,182	6.09%	530	4.50%
Spain	1,247	6.42%	723	6.14%
Netherlands	882	4.54%	495	4.21%
Denmark	576	2.97%	244	2.07%
Ireland	570	2.94%	278	2.36%
Finland	988	5.09%	699	5.94%
Italy	582	3.00%	319	2.71%
Portugal	306	1.58%	68	0.58%
Poland	381	1.96%	153	1.30%
Hungary	209	1.08%	16	0.14%
Austria	300	1.55%	110	0.93%
Belgium	469	2.42%	332	2.82%
Norway	101	0.52%	47	0.40%
Czech Republic	85	0.44%	36	0.31%
Greece	60	0.31%	14	0.12%
Slovakia	40	0.21%	14	0.12%
Slovenia	19	0.10%	4	0.03%
Switzerland	3	0.02%	1	0.01%
Total	19,415		11,769	
Pearson chi2(21)=479.94, Pr=0.000				

Table A2 - Distribution of the sample by industry

Industry	VICO		VICO companies with business description	
	n	%	n	%
Information and communication	6,678	34.40%	4,088	34.74%
Manufacturing	4,460	22.97%	3,026	25.71%
Professional, scientific and tech.	1,528	7.87%	994	8.45%
Wholesale and retail trade	1,031	5.31%	663	5.63%
Financial and insurance activities	663	3.41%	395	3.36%
Administrative and support service	479	2.47%	294	2.50%
Human health and social work act.	250	1.29%	163	1.38%
Electricity, gas, steam and air cond.	189	0.97%	124	1.05%
Construction	162	0.83%	120	1.02%
Transportation and storage	149	0.77%	99	0.84%
Accommodation and food service activities	145	0.75%	85	0.72%
Mining and quarrying	130	0.67%	82	0.70%
Education	103	0.53%	55	0.47%
Arts, entertainment and recreation	102	0.53%	65	0.55%
Water supply; sewerage, waste management	94	0.48%	63	0.54%
Real estate activities	70	0.36%	46	0.39%
Other service activities	65	0.33%	29	0.25%
Agriculture, forestry and fishing	57	0.29%	35	0.30%
Public administration and defence	18	0.09%	6	0.05%
Not classified	3042	15.67%	1337	11.36%
Total	19415		11769	
Pearson chi2(19)=200.17, Pr=0.000				

Table A3 - Distribution of the sample by period of establishment

Period of establishment	VICO		VICO companies with business description	
	n	%	n	%
Until 2000	4,684	24.13%	3,085	26.21%
Between 2001 and 2007	6,369	32.80%	4,399	37.38%
After 2007	8,362	43.07%	4,285	36.41%
Total	19,415	100%	11,769	100%
Pearson chi2(2)=217.51, Pr=0.000				

Even if broad definitions of clean technologies are used both at governmental and academic level, the punctual identification and classification of cleantech innovative companies is more difficult. As for many emerging/innovative sectors, including cleantech, standard industry classification such as NACE, is not typically able to capture the sustainable characteristics of a particular business or activity (Criscuolo and Menon, 2015, Cumming et al., 2016, Christensen and Hain, 2017, Mazzucato and Semieniuk, 2017). [12]

Therefore, the first step of our analysis is devoted to the definition and identification of cleantech companies. Based on common definitions (Migent et al., 2017), "cleantech are products, services and technologies able to improve the productive and responsible use of natural resources, to reduce or eliminate negative environmental impacts, and to provide superior performance at a lower cost compared to existing solutions". According to UNFCC[13], the cleantech sector consists of energy efficiency, renewable energy, waste beneficiation, water efficiency, green buildings, transport, advanced materials and chemicals.

All definitions are extremely general and are not able to support the identification of cleantech projects among the entire sample of VC-backed companies. Therefore, any analysis of cleantech must forego a specific punctual industry reclassification of the sample or database utilised in the analysis. Several previous studies (Malen and Marcus, 2017, Criscuolo and Menon, 2015, Polzin et al., 2015) analysed VC cleantech investments already collected and classified by a third-party information provider (Cleantech Group, Bloomberg New Energy Finance). Some other authors (Shapira et al., 2014, Petkova et al., 2014, Gaddy et al., 2017, Cumming et al., 2016, Mazzucato and Semieniuk, 2017) applied a punctual reclassification of start-ups to identify cleantech ones through an exogenous dictionary of cleantech relevant words. As highlighted by Butticè et al. (2019), the exogenous

[12] The definition issue also arises at EU level. The final report on sustainable finance published in 2018 by the High-Level Expert Group on Sustainable Finance - Secretariat provided by the European Commission[12] identifies a common taxonomy on sustainability at EU level as one of the key recommendations to guarantee a common framework to financial institutions. The subsequent technical report on taxonomy published in June 2019 focuses on climate change risk taxonomy. It defines relevant climate change risks for private companies (climate change adaptation and climate change mitigation of underling activities) and identifies a first subset of sub-industries characterised by a high level of these two specific risks. This report represents not only the first comprehensive taxonomy on sustainability, but also underlines the importance of taxonomy as the first step for providing a common ground to investors and financial institutions who are approaching the topic.

[13] United Nation Framework Convention on Climate Change - UNFCC http://bigpicture.unfccc.int/#content-the-paris-agreemen

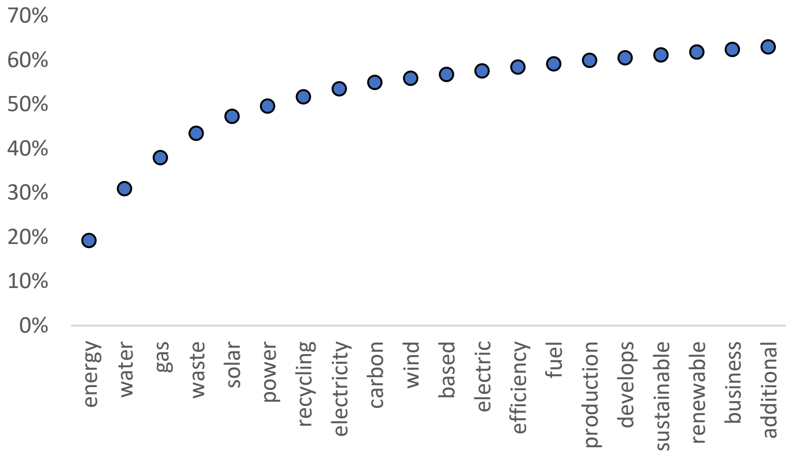
selection of keywords may introduce a relevant bias: a dictionary of keywords is arbitrarily defined by the authors without any procedure of validation of such dictionary.

In this paper, we build a fully replicable methodology that, starting from the full sample of VC-backed companies for which we collected the extensive business description, identifies cleantech and non-cleantech companies. We employ a machine-learning algorithm to create a content-specific classifier of cleantech companies based on their extensive business description. This approach addresses the need for transparent analysis based on publicly available data: the methodology applied can be fully replicated on the sample utilised thanks to the identification process transparent and based on specific algorithm. Text classification firstly analyses a set pre-classified business description (training set), derives a decision function, and then applies it to predict the category of description whose class is unknown. A widely-used approach consists in the bag-of-words model (Sebastiani, 2002), where each business description (or document) is treated as a set of terms and converted to a numeric vector containing the frequency of occurrence of each term in the document.

We apply this technique, analysing the extended business description of each company in our dataset to identify cleantech companies. The first step of the procedure consists of randomly splitting the dataset by identifying a subset of descriptions to create the training set. The training set consists of the description of 380 companies which were manually tagged as "cleantech" or "non-cleantech" according to the definition of cleantech set by Migent et al. (2017) where cleantech includes companies which focus on green and sustainable technologies with products, processes or services able to reduce the amount of greenhouse gas emissions. Manual tagging has been made by two research assistants separately. When differences in the tagging arose (< 5% of cases), one of the authors classified the document and then discussed the tagging with the research assistants until agreement was reached. Each labelled text was then analysed using natural language processing (NLP) filters and was converted into a numeric vector. A machine-learning algorithm was finally implemented to identify the optimal classification function, which was used to predict the classification of the remaining companies in the sample. Different machine learning algorithms have been utilised to classify texts; among these, the Random Forest algorithm (Breiman, 2001) was selected thanks to its accuracy, efficiency and robustness. Random forest has shown great potential in several domains, ranging from risk assessment in social lending (Malekipirbazari and Aksakalli, 2015) to bank failure forecasts (Barboza et al., 2017).

In addition to the forecast properties, however, two other characteristics put forward its implementation in the present research. First, unlike other machine-learning algorithms, it requires a limited number of iterations for tuning its parameters. Second, it generates internal estimates of the importance of the variables, such as the mean decrease in accuracy (MDA), which measures the relevance of the predictors both in individual and in multivariate interactions. This property has been utilised to identify the words that, among others, contributed most to the accurate discrimination between cleantech and non-cleantech companies. Not surprisingly, terms such as "energy", "water", "waste" and "solar" emerged as the most influential according to the best classification model. The list of the 30 most relevant words generated by the learning process is provided in Figure A1.

Figure A1 - Mean decrease in accuracy



Cleantech companies, identified by RF algorithm, represents 9.21% of the entire sample of companies: our sample is finally composed of 1,066 cleantech companies and 10,703 non-cleantech companies.

Appendix B- Environmental policy stringency measures

To evaluate the effect of policy stringency effect, we utilised indicators developed by Botta and Koźluk (2014).

They identify, on a yearly and country basis, the stringency level of 5 environmental policy indicators: three market-based, and two non-market based. Market-based indicators are trading schemes of environmental certificates (ETS), environmental taxes (Taxes) and feed-in tariff (FIT) mechanisms. Non-market based policies are emission limits and R&D subsidies.

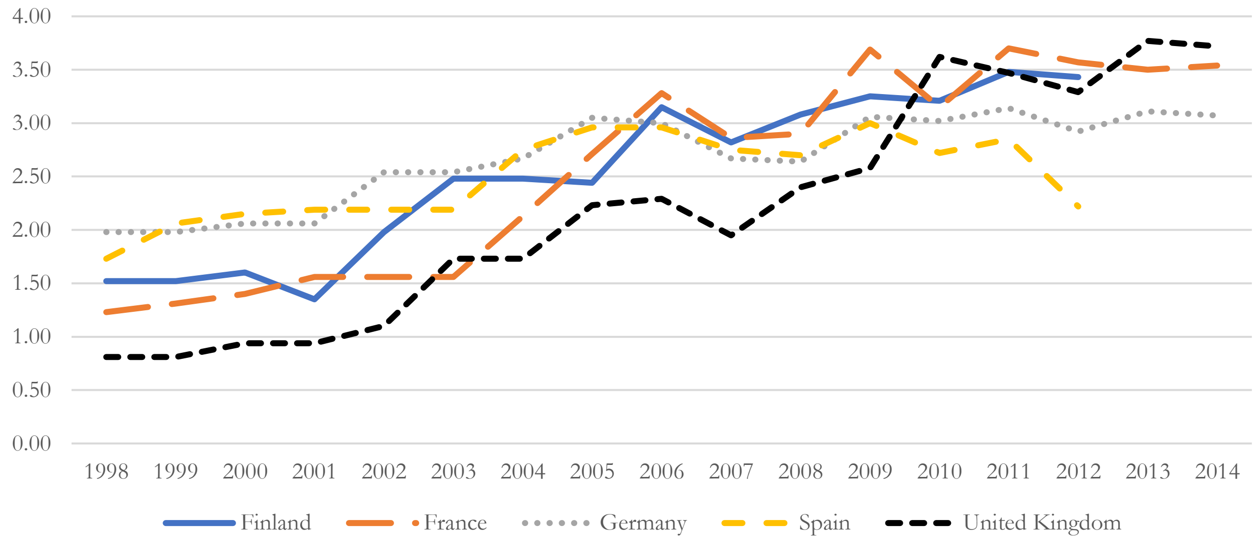
For each indicator, authors analysed multiple information and quantitative measures and then scored and aggregated them into the five policy indicators reported and described in Table B1.

Table B1 – Description of policy stringency variables

Policy category	Policy indicator	Policy instrument	Information considered for scoring
Market-based policy	ETS	Emission Trading Scheme(CO2)	Price of one CO2 allowance
Market-based policy	ETS	Renewable Energy Certificates Trading Scheme	% of renewable electricity that has to be procured annually
Market-based policy	ETS	Energy Certificate Emission trading Scheme	% of electricity saving that has to be delivered annually
Market-based policy	ETS	Emission trading Scheme for SO2	Price of one SO2 allowance
Market-based policy	Taxes	CO2 tax	Tax rate in EUR/ tonne
Market-based policy	Taxes	NOx Tax	Tax rate in EUR/ tonne
Market-based policy	Taxes	SOx Tax	Tax rate in EUR/ tonne
Market-based policy	FIT	Feed-in tariff for wind	EUR/kWh
Market-based policy	FIT	Feed- in premium for wind	EUR/kWh
Market-based policy	FIT	Feed-in tariff for solar	EUR/kWh
Market-based policy	FIT	Feed-in premium for solar	EUR/kWh
Non-market policy	Emission limit	Particulate Matter Emission Limit Value for newly built coal-fired plant	Value of Emission Limit in mg/m3
Non-market policy	Emission limit	SOx Emission Limit Value for newly built coal-fired plant	Value of Emission Limit in mg/m3
Non-market policy	Emission limit	NOx Emission Limit Value for newly built coal-fired plant	Value of Emission Limit in mg/m3
Non-market policy	R&D Subsidies	Government R&D expenditures for renewable energy technologies	Expressed as % of GDP

Authors also developed a comprehensive synthetic indicator: OECD PSI Index. Figure B1 shows the evolution of OECD PSI index in the first 5 countries in terms of the number of cleantech companies financed by VCs. A positive long-term trend of environmental policy stringency characterised all countries, but yearly variations evidence that specific governmental interventions determine a peculiar short-term behaviour in each country. The graph demonstrates that resorting to dummy variables as a proxy of policy intervention may bias the analysis, given that the level of policy intensity shows a highly variable trend over time. If policy stringency increases over time, institutional investors may significantly modify their investment strategy and this phenomenon may only be captured by a variable indicating the stringency level of the policy over time instead of a dummy indicating the mere presence of a policy in the period under consideration.

Figure B1 – OECD PSI index in five European countries



Analysis of the synthetic PSI index can provide only a general framework of environmental policies' evolution at country level, but it is not able to unravel the impact of each category of environmental policy on VC investments. We then focus our analysis on policy indicators identified by Botta and Koźluk (2014). Figures B2 - Figure B6 show the minimum, maximum and average value of stringency indicator for each policy included in our study. Minimum and maximum do not frequently represent the first and last observation of indicator in a given country[14]; the stringency level of a single instrument evolves during the horizon analysed, also based on the general environmental country policy: therefore, for several countries/policies, we can observe an initial increase and subsequent decrease, with, eventually, the last value lower than the first one. For this reason, we opt for a minimum-maximum representation of each stringency indicator. Figures B2 - Figure B6 show that the heterogeneity among countries is high: a country's rank, based on average value, is different across indicators, proof of different global environmental strategies applied in each country. Not only the average level, but also the volatility around this value, varies at country and instrument level. The volatility of stringency level is also influenced by specific government laws or regulations, which, on a yearly basis, can modify incentives to develop clean technologies and impose penalties for pollution. Finally, Table B2 presents the detailed data of each stringency indicator at country level.

[14] i.e. for taxes, minimum and maximum represent the first and last observation in only 12 out of 42

Figure B2 – Maximum, minimum and average value of environmental stringency of taxes in the period 1994-2014

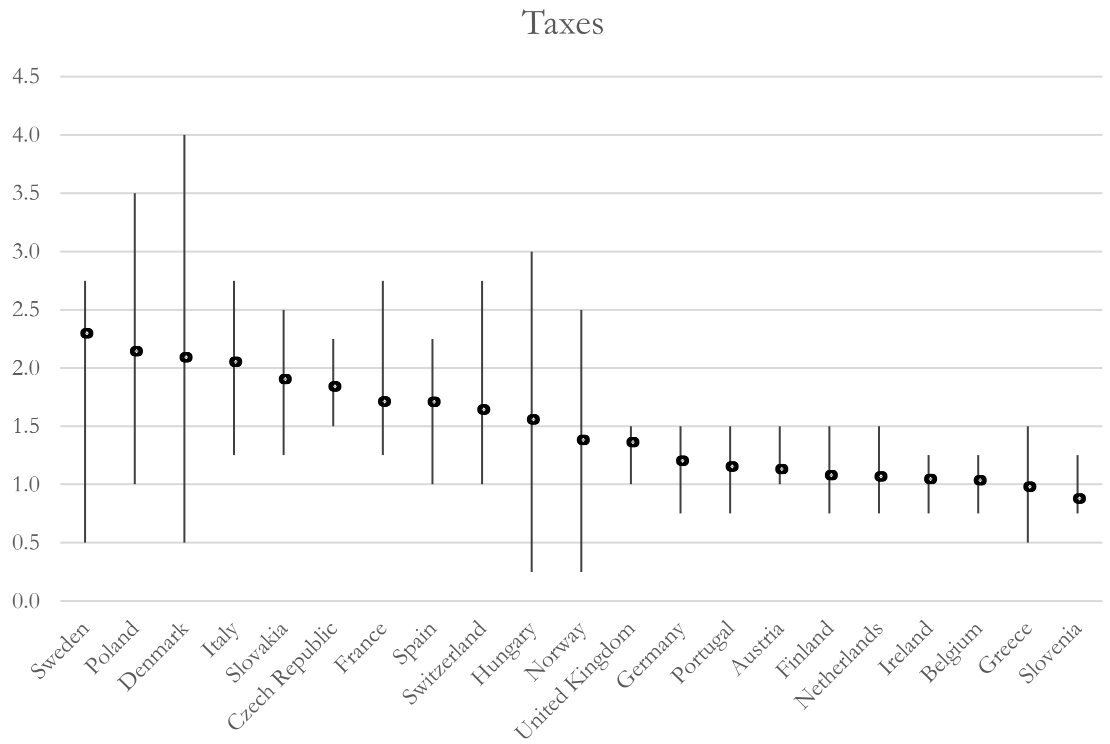


Figure B3 – Maximum, minimum and average value of environmental stringency of emission limits in the period 1994-2014

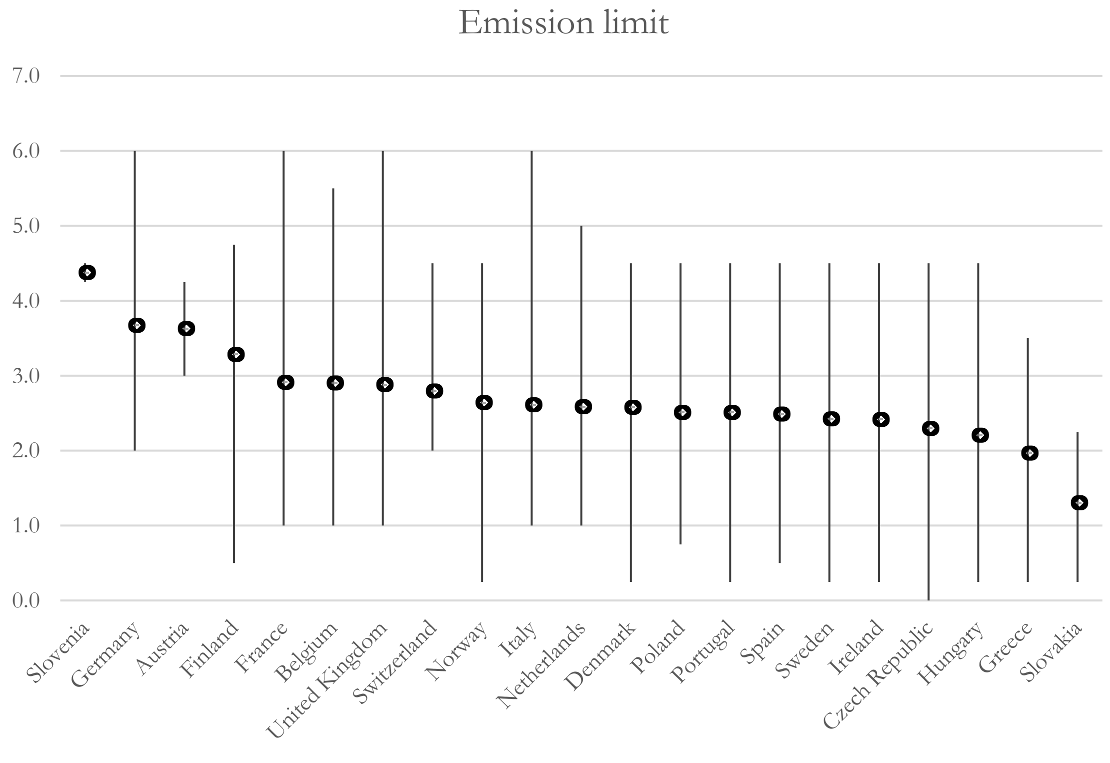


Figure B4 – Maximum, minimum and average value of environmental stringency of ETS in the period 1994-2014

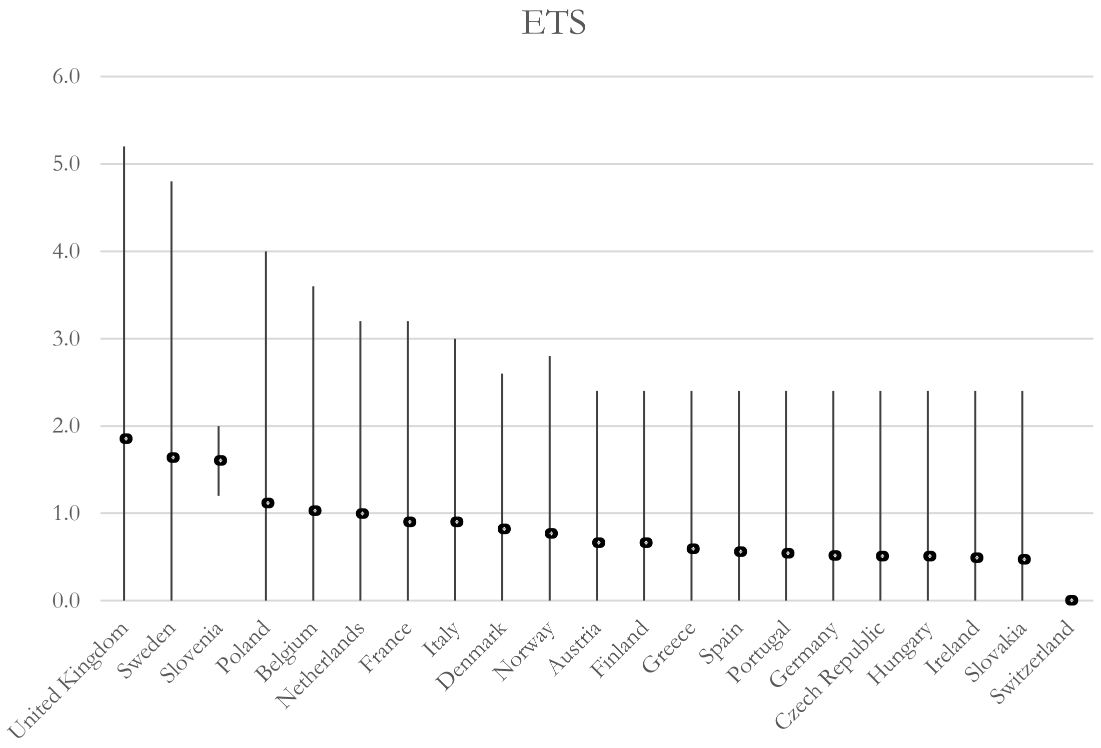


Figure B5– Maximum, minimum and average value of environmental stringency of FIT in the period 1994-2014

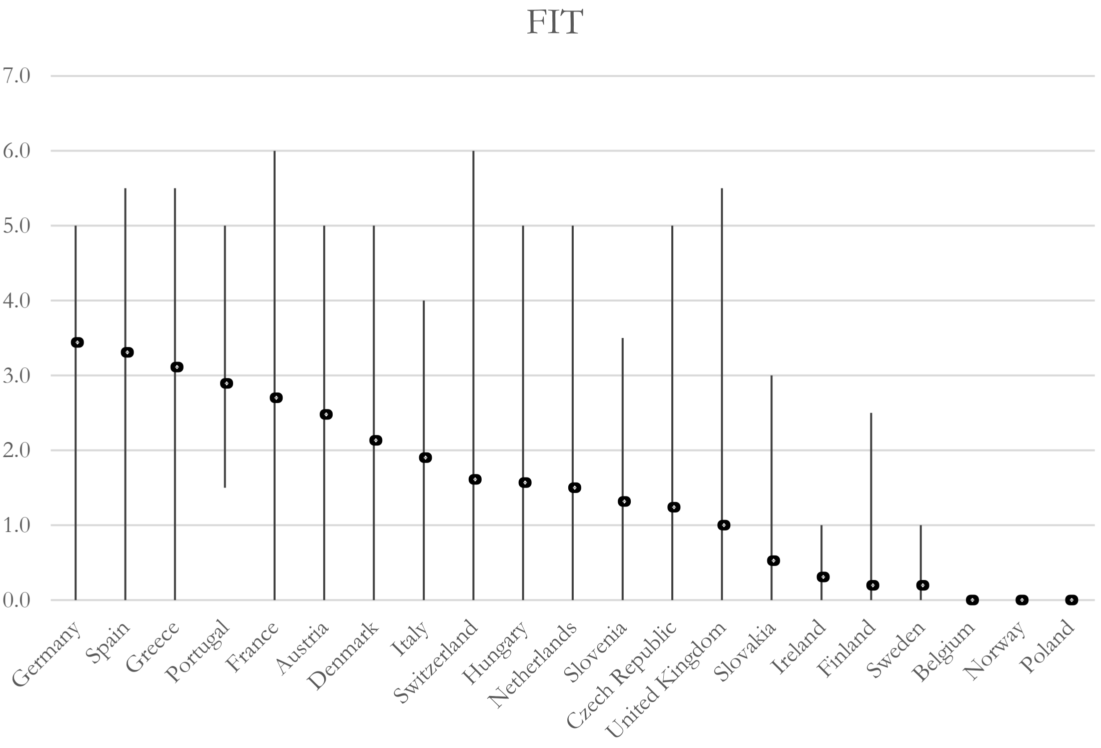


Figure B6– Maximum, minimum and average value of environmental stringency of R&D subsidies in the period 1994-2014

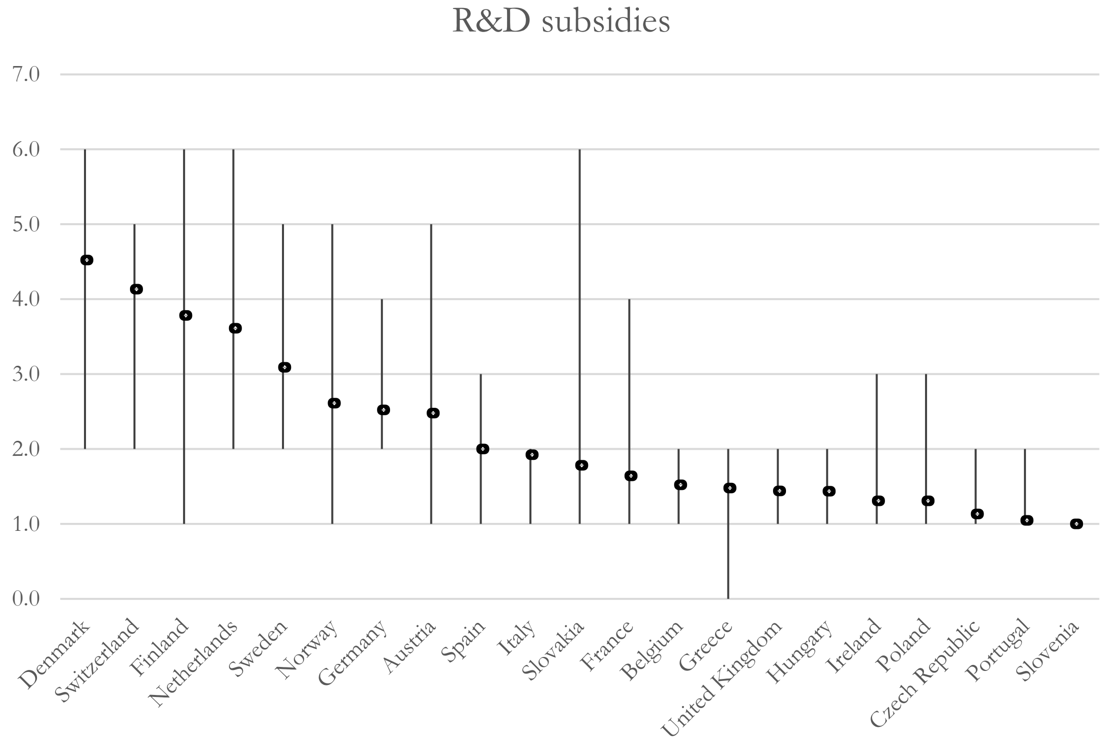


Table B2 – Descriptive statistics of Taxes, Emission limits, ETS, FIT and R&D subsidies by country

			Taxes		
	average	first obs.	last obs.	min obs.	max obs.
Austria	1.13	1.00	1.00	1.00	1.50
Belgium	1.03	1.00	0.75	0.75	1.25
Czech Republic	1.84	1.75	1.75	1.50	2.25
Denmark	2.09	0.50	3.75	0.50	4.00
Finland	1.08	1.25	0.75	0.75	1.50
France	1.71	1.25	2.75	1.25	2.75
Germany	1.20	1.25	1.00	0.75	1.50
Greece	0.98	0.50	0.75	0.50	1.50
Hungary	1.55	0.25	2.75	0.25	3.00
Ireland	1.04	1.25	0.75	0.75	1.25
Italy	2.05	1.25	2.25	1.25	2.75
Netherlands	1.07	1.00	0.75	0.75	1.50
Norway	1.38	0.25	2.50	0.25	2.50
Poland	2.14	1.25	2.50	1.00	3.50
Portugal	1.15	1.25	0.75	0.75	1.50
Slovakia	1.90	1.25	1.75	1.25	2.50
Slovenia	0.88	1.00	0.75	0.75	1.25

Spain	1.71	1.00	1.75	1.00	2.25
Sweden	2.29	0.50	2.25	0.50	2.75
Switzerland	1.64	1.50	2.50	1.00	2.75
United Kingdom	1.36	1.25	1.25	1.00	1.50
Average	1.49	1.02	1.67	0.83	2.15

	Emission limit				
	average	first obs.	last obs.	min obs.	max obs.
Austria	1.13	1.00	1.00	1.00	1.50
Belgium	1.03	1.00	0.75	0.75	1.25
Czech Republic	1.84	1.75	1.75	1.50	2.25
Denmark	2.09	0.50	3.75	0.50	4.00
Finland	1.08	1.25	0.75	0.75	1.50
France	1.71	1.25	2.75	1.25	2.75
Germany	1.20	1.25	1.00	0.75	1.50
Greece	0.98	0.50	0.75	0.50	1.50
Hungary	1.55	0.25	2.75	0.25	3.00
Ireland	1.04	1.25	0.75	0.75	1.25
Italy	2.05	1.25	2.25	1.25	2.75
Netherlands	1.07	1.00	0.75	0.75	1.50
Norway	1.38	0.25	2.50	0.25	2.50
Poland	2.14	1.25	2.50	1.00	3.50
Portugal	1.15	1.25	0.75	0.75	1.50
Slovakia	1.90	1.25	1.75	1.25	2.50
Slovenia	0.88	1.00	0.75	0.75	1.25
Spain	1.71	1.00	1.75	1.00	2.25
Sweden	2.29	0.50	2.25	0.50	2.75
Switzerland	1.64	1.50	2.50	1.00	2.75
United Kingdom	1.36	1.25	1.25	1.00	1.50
Average	1.49	1.02	1.67	0.83	2.15

	ETS				
	average	first obs.	last obs.	min obs.	max obs.
Austria	1.13	1.00	1.00	1.00	1.50
Belgium	1.03	1.00	0.75	0.75	1.25
Czech Republic	1.84	1.75	1.75	1.50	2.25
Denmark	2.09	0.50	3.75	0.50	4.00
Finland	1.08	1.25	0.75	0.75	1.50
France	1.71	1.25	2.75	1.25	2.75
Germany	1.20	1.25	1.00	0.75	1.50

Greece	0.98	0.50	0.75	0.50	1.50
Hungary	1.55	0.25	2.75	0.25	3.00
Ireland	1.04	1.25	0.75	0.75	1.25
Italy	2.05	1.25	2.25	1.25	2.75
Netherlands	1.07	1.00	0.75	0.75	1.50
Norway	1.38	0.25	2.50	0.25	2.50
Poland	2.14	1.25	2.50	1.00	3.50
Portugal	1.15	1.25	0.75	0.75	1.50
Slovakia	1.90	1.25	1.75	1.25	2.50
Slovenia	0.88	1.00	0.75	0.75	1.25
Spain	1.71	1.00	1.75	1.00	2.25
Sweden	2.29	0.50	2.25	0.50	2.75
Switzerland	1.64	1.50	2.50	1.00	2.75
United Kingdom	1.36	1.25	1.25	1.00	1.50
Average	1.49	1.02	1.67	0.83	2.15

			FIT		
	average	first obs.	last obs.	min obs.	max obs.
Austria	2.48	0.00	3.50	0.00	5.00
Belgium	0.00	0.00	0.00	0.00	0.00
Czech Republic	1.24	0.00	3.50	0.00	5.00
Denmark	2.13	0.00	2.00	0.00	5.00
Finland	0.20	0.00	2.50	0.00	2.50
France	2.70	0.00	3.00	0.00	6.00
Germany	3.44	0.00	2.00	0.00	5.00
Greece	3.11	0.00	4.50	0.00	5.50
Hungary	1.57	0.00	2.50	0.00	5.00
Ireland	0.30	0.00	1.00	0.00	1.00
Italy	1.90	0.00	4.00	0.00	4.00
Netherlands	1.50	0.00	4.00	0.00	5.00
Norway	0.00	0.00	0.00	0.00	0.00
Poland	0.00	0.00	0.00	0.00	0.00
Portugal	2.89	2.00	3.00	1.50	5.00
Slovakia	0.52	0.00	3.00	0.00	3.00
Slovenia	1.31	0.00	3.50	0.00	3.50
Spain	3.30	0.00	2.50	0.00	5.50
Sweden	0.20	0.00	0.00	0.00	1.00
Switzerland	1.61	0.00	4.50	0.00	6.00
United Kingdom	1.00	0.00	4.50	0.00	5.50
Average	1.50	0.10	2.55	0.07	3.74

R&D subsidies					
	average	first obs.	last obs.	min obs.	max obs.
Austria	2.48	1.00	4.00	1.00	5.00
Belgium	1.52	1.00	2.00	1.00	2.00
Czech Republic	1.13	1.00	1.00	1.00	2.00
Denmark	4.52	3.00	6.00	2.00	6.00
Finland	3.78	2.00	6.00	1.00	6.00
France	1.64	1.00	3.00	1.00	4.00
Germany	2.52	2.00	4.00	2.00	4.00
Greece	1.48	2.00	1.00	0.00	2.00
Hungary	1.43	1.00	2.00	1.00	2.00
Ireland	1.30	1.00	2.00	1.00	3.00
Italy	1.92	2.00	2.00	1.00	2.00
Netherlands	3.61	5.00	5.00	2.00	6.00
Norway	2.61	2.00	5.00	1.00	5.00
Poland	1.30	1.00	2.00	1.00	3.00
Portugal	1.04	1.00	1.00	1.00	2.00
Slovakia	1.78	1.00	6.00	1.00	6.00
Slovenia	1.00	1.00	1.00	1.00	1.00
Spain	2.00	2.00	1.00	1.00	3.00
Sweden	3.09	3.00	4.00	2.00	5.00
Switzerland	4.13	5.00	4.00	2.00	5.00
United Kingdom	1.44	2.00	2.00	1.00	2.00
Average	2.18	1.90	3.05	1.19	3.62