What is the most important indicator for the Environmental Performance Index?

Supervised Machine Learning – final paper

Jingwen Zhang (17-742-487)

17/07/2020

1 Introduction

In this paper, I will assess the relative importance of different factors of predicting nations' Environmental Performance Index (EPI). To answer this question, two different datasets have been used, one is the Environmental Performance Index calculated by Yale Center for Environmental Law & Policy annually, the other is replicated from the research by Baettig and Bernauer (2009) which examines whether democracies are more cooperative in climate change policy by studying a cross-section of 185 countries in 1990–2004. To evaluate the relative importance, several supervised machine learning methods are applied. Firstly, by running a General Linear Model, an overview of the relationships between all variables are shown; after that, Partial Least Squares and Random Forests are both applied to compare and find out the most important explanatory variables. Based on the three most important features selected by the previous steps, the C5.0 algorithm is used to further assess the most important factor for predicting EPI, which is the variable 'income'. In the end, I plotted 4 classes of EPI and 4 classes of income onto world maps to summarize the findings.

2 Data

To find out which variables contribute most to environmental performance, I chose the most updated Environmental Performance Index 2020 as the dependent variable calculated by the Yale Center for Environmental Law & Policy. By using 32 performance indicators across 11 issue categories (Figure 1), including air quality, sanitation & drinking water, heavy metals, waste management, biodiversity & habitat, ecosystem services, climate change, pollution emissions, agriculture, water resources, fisheries, the EPI ranks 180 countries on environmental health and ecosystem vitality (Environmental Performance Index, 2020). As for the explanatory variables, which are considered as factors influencing the environment in

current environmental policy studies, these include trade openness, democracy index, policy output, policy outcome, CO2 emissions per capita in 1990, climate change index (CCI), GDP growth rate, income, those are replicated from the research by Baettig and Bernauer (2009). Trade openness is the ratio of the sum of exports and imports to GDP, it is included as one of those explanatory variables because there is a tradeoff between gains from a cleaner environment and losses from lower exports is more adverse for more open economies (Bernauer et al., 2010: 518). The democracy index is included because the effect of democracy on levels of political commitment to climate change mitigation (policy output) is positive, and policy output leads to emission reduction, which is policy outcome (Baettig and Bernauer, 2009). The variable 'CO2 emissions per capita in 1990' is used as a proxy to estimate mitigation costs, however, there is no agreement on this (Stern, 2007; Baettig and Bernauer, 2009), some argues that countries with higher per capita emissions may be less energy-efficient so it is easier to reduce emissions, but some states that those emission-intensive countries might be less willing to reduce emissions to avoid the sacrifice of the economies. We will check this disagreement with the results of this paper. Climate Change Index (CCI) is also included, it is the index of climate change risk exposure, higher values indicate greater risk for the respective country, countries which are faced with a higher risk in the future are expected to more cooperative in environmental policies, which will contribute to better environmental performance (Baettig and Bernauer, 2009). Other economic factors such as income and GDP growth rate are covered in this paper. The variable 'income' is measured as the natural logarithm of GDP per capita, the higher the level of income of a country, it is likely to be associated with nonincreasing or decreasing emissions (Baettig and Bernauer, 2009); the variable 'GDP growth' is the average annual growth of GDP per capita calculated by the World Bank, it is expected that the higher the growth rate is, the more intensive the economic activities are, therefore it would contribute to higher emissions which will worsen the environmental performance.

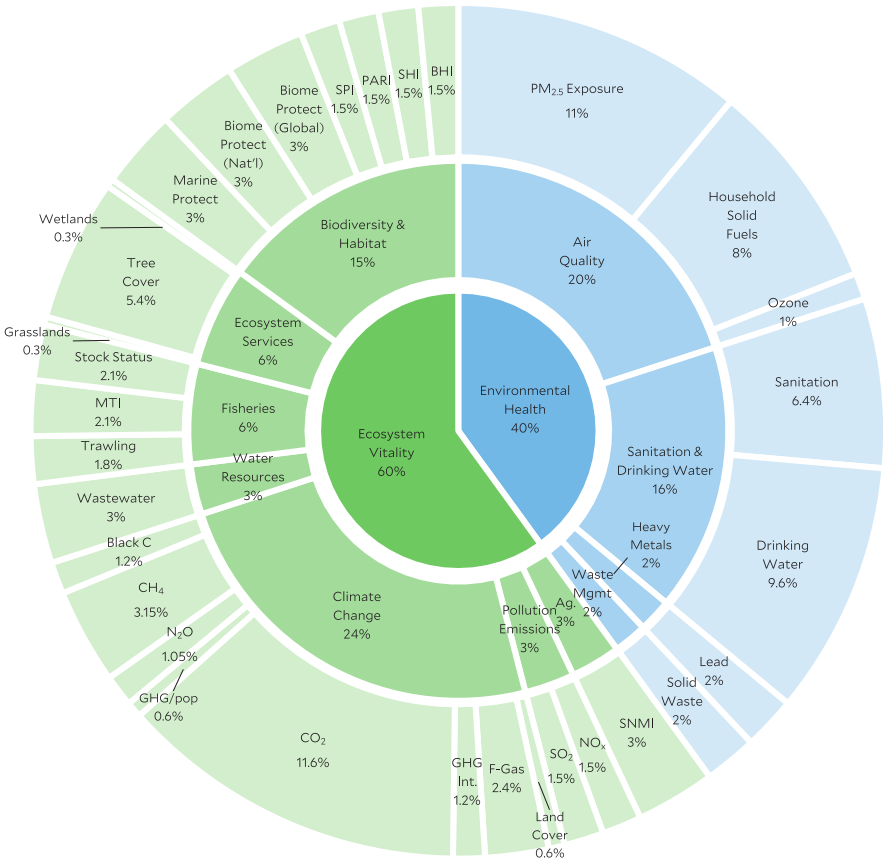

Figure 1: The EPI framework of indicators and weights

(from https://epi.yale.edu/downloads/epipolicymakersummaryr9.pdf)

Combining these 2 datasets together, in total, the dataset covers 173 countries worldwide with 9 variables, all missing values are replaced with the median of each column since there are just several missing values which have no relationship to the outcome because they are just small regions. The values of them are missing because of the different coding systems of world regions of these 2 different datasets. The basic characteristics of different variables are shown in Table 1.

Descriptive statistics

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
GDP Growth	173	3.22	2.97	-4.31	1.77	4.40	19.47
Democracy Index	173	64.36	29.09	14.29	39.00	95.29	100.00
Income	173	8.12	1.11	5.74	7.14	8.91	10.26
Outcome	173	0.59	0.23	0.09	0.42	0.79	1.00
Output	173	0.63	0.19	0.00	0.54	0.75	1.00
CCI	173	0.50	0.14	0.00	0.40	0.55	1.00
Trade openness	173	34.28	42.27	0	11	42	371
CO2cap1990	173	4.50	5.88	0.00	0.40	6.80	31.10
EPI	173	46.72	15.56	22.60	34.40	55.20	82.50

Table 1: descriptive statistics of variables

3 Methods and Results

In this paper, several methods are applied step by step, therefore, the explanation of methods chosen and discussion of results will be analyzed together. Firstly, I used General Linear Model to have an overview of the relationships between all variables; after that, Partial Least Squares and Random Forests are both applied to find out the most important explanatory variables. Based on the most important features selected by these methods, I classify EPI into 4 classes and use the C5.0 algorithm to compare theses important factors and find out which one is the most important factor to predict EPI classes.

General Linear Model

First, I run a linear regression model to have an overview of the relationship between different variables. As we can see in table 2, variables 'democracy' and 'income' have statistically positive relationships with EPI, while 'climate change index' and outcome' have statistically negative relationships with EPI, which mean the higher climate change risk exposure a country has, the more likely it will have a worse environmental performance; the higher the environmental outcomes (such as higher GHG emissions), the worse the environmental performance of a country is.

	Dependent variable:
	EPI
gdpgrowth	−0.258
	(0.189)
democracy	0.080***
	(0.026)
income	8.991***
	(0.895)
outcome	−10.271***
	(2.931)
output	5.200
	(3.440)
cci	−17.016***
	(4.056)
tradeopen	−0.015
	(0.014)
CO2cap1990	0.101
	(0.137)
Constant	−19.247***
	(6.879)
Observations	173
Log Likelihood	−576.372
Akaike Inf. Crit.	1,170.745
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 2: regression result of the GLM

Partial Least Squares

I chose dimension reduction because it is an efficient way for feature selection because the first component accounts for most of the variance of the features. There are 2 common methods of dimension reduction, which are Partial Least Squares, and Principal Components Regression. PLS is selected here as it overcomes the shortcomings of PCR, the latter one is an unsupervised method and it is applied without the consideration of the correlation between the dependent variable and the independent variables themselves (Maitra and Yan, 2008). In comparison, PLS finds components that maximally summarize the variation of the predictors while simultaneously requiring these components to have a maximum correlation with the response, therefore, the PLS direction contains highly predictive information for the response (Kuhn and Johnson, 2013).

Required by PLS, the predictors should be centered and scaled, especially if the predictors are on scales of differing magnitude (Kuhn and Johnson, 2013), this is done within the regression. Since there are in total 173 observations, I chose leave-out-one cross-validation instead of the commonly used 10-fold cross-validation. The first component consists of more than 76% of the total variance (Table 3), which is satisfying. As shown in Figure 2, we can observe an "elbow" clearly around the first component.

	1	2	3	4	5	6	7	8
X	33.60	47.31	58.06	67.70	76.40	84.75	94.78	100.00
EPI	76.59	79.27	80.44	81.00	81.16	81.17	81.17	81.17

Table 3: variance explained (%)

gdpgrowth	democracy	income	outcome	output	cci	tradeopen	CO2cap1990
-0.13	0.39	0.56	-0.36	0.30	-0.19	0.24	0.45

Table 4: loadings of variables on the first component

When setting the cutoff of loadings to 0.35, variables 'democracy', 'income', 'outcome', 'CO2cap1990' stand out, as shown in Table 4, which suggests focusing on these 4 variables further on.

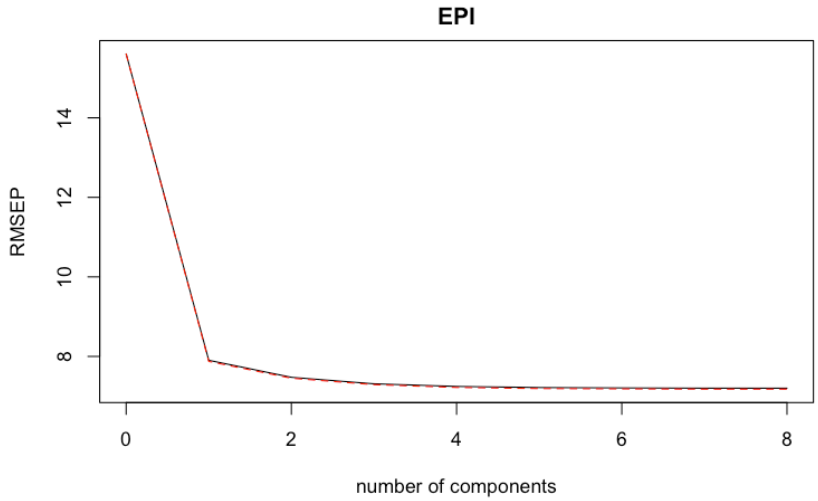

Figure 2: Plot of components

Random Forests Regression

In addition to the method Partial Least Sqaure applied above, random forest regression is also an efficient way to select the important features. In general, random forest regression belongs to ensemble learning which combines the predictions from multiple machine learning algorithms to have more accurate predictions than any individual model alone, which are expected to have higher accuracy. I will compare the results from the random forests with results from PLS and see whether the same features are considered as important by both methods. Random Forest Regression is a bagging technique (Bootstrap aggregation) which involves random sampling of a small subset of data from the dataset, results in a distribution of predicted values for each sample/tree, those predictions are averaged to give the forest's prediction (Kuhn and Johnson, 2013). Because of the same reason mentioned above, I used the leave-out-one-cross-validation, and 2 features are selected at each split because it is recommended that the number of features per split should be around one-third of total number of features when running a regression.

The result shows that the variance explained is around 84%. By checking the importance matrix, variables 'income', 'CO2cap1990', 'democracy' are the most important features, as shown in Figure 3.

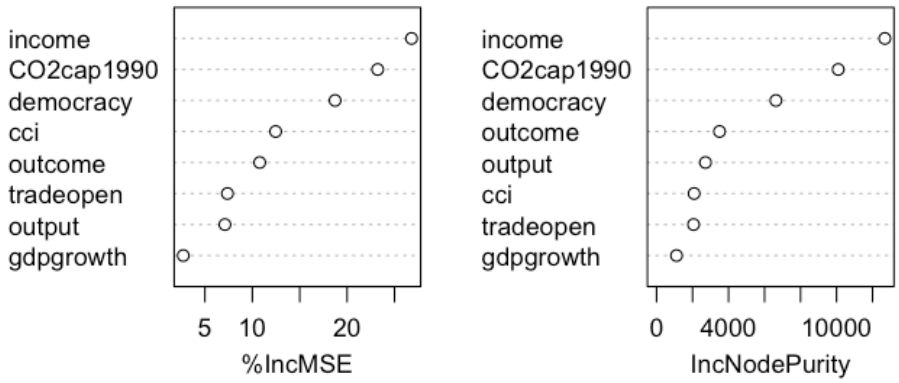

Figure 3: Importance of variables

In the previous part, results of PLS analysis suggests variables 'democracy', 'income', 'outcome', 'CO2cap1990' are 4 most important feature, compared to the results of random forests regression, I will focus on variables 'income', 'CO2cap1990', 'democracy' for further analysis.

The C5.0 Algorithm

Based on the results from feature selection, I create a subset of data that excludes variables other than variables 'income', 'CO2cap1990', 'democracy'. At the same time, I divide the dependent variable EPI into 4 classes based on the quantile of the data distribution, this is because in the conclusion part I want to visualize them in the world map. The new subset includes 3 explanatory variables and the classes of EPI.

The C5.0 algorithm is utilized to decide how splits should be made and the numbers of splits, which offers the information for the visualization of the classification tree. The process of the C5.0 algorithm is first growing an initial tree is, collapsed into rules, then the individual rules are simplified via pruning and a global procedure is used on the entire set to potentially reduce the number of constituent rules (Kuhn and Johnson, 2013). This is helpful for the research question in this step because I am curious to see which feature will remain as the most important one. I split the sample on a 0.8 threshold and used the C5.0 algorithm. The overall accuracy rate is around 45%, which is not optimal (Table 4).

Accuracy	Kappa	AccuracyLower	AccuracyUpper	AccuracyNull	AccuracyPValue
0.45	0.26	0.28	0.64	0.48	0.70

Table 4: accuracy of the prediction

However, the plot tells a different story, we can see the variable ‘income’ is the most important and it appears twice. Moreover, as long as it is higher than 9.16, EPI class absolutely belongs to the highest class (= 4), which has the best environmental performance. In this case, the accuracy rate of income for EPI class 4th is 100%.

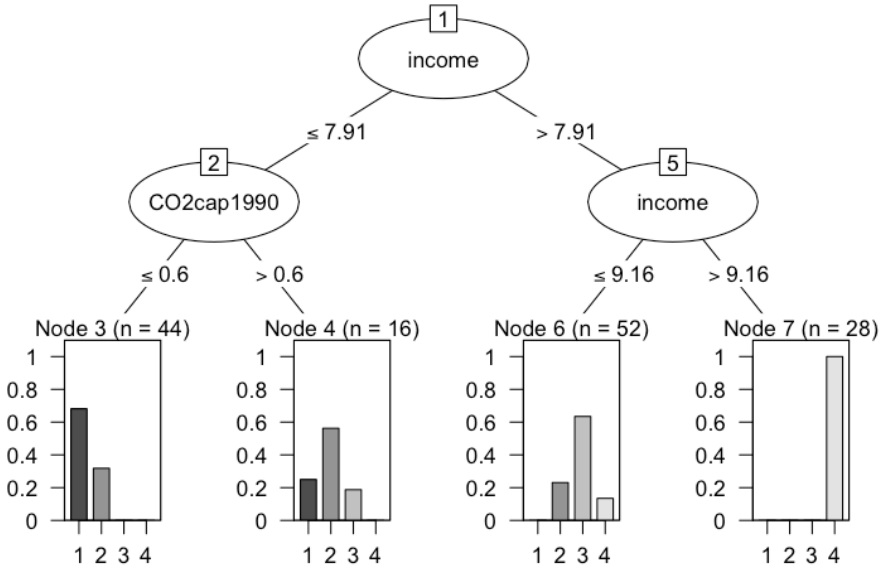

Figure 4: The decision tree

As for other EPI classes, the accuracy rate is low probably because I only keep 3 most important variables, and also classify all EPI into only 4 classes, which is somewhat arbitrary, however, this is because I want to find out the most important feature and to simplify the decision tree.

4 Dicussion and conclusion

Since it is suggested by the decision tree that the variable 'income' is the most important indicator to predict the higher classes of EPI, I also divided 'income' into 4 classes, and plot them onto the world map. At the same time, I also plot 4 classes of EPI onto the world map to offer a comparison.

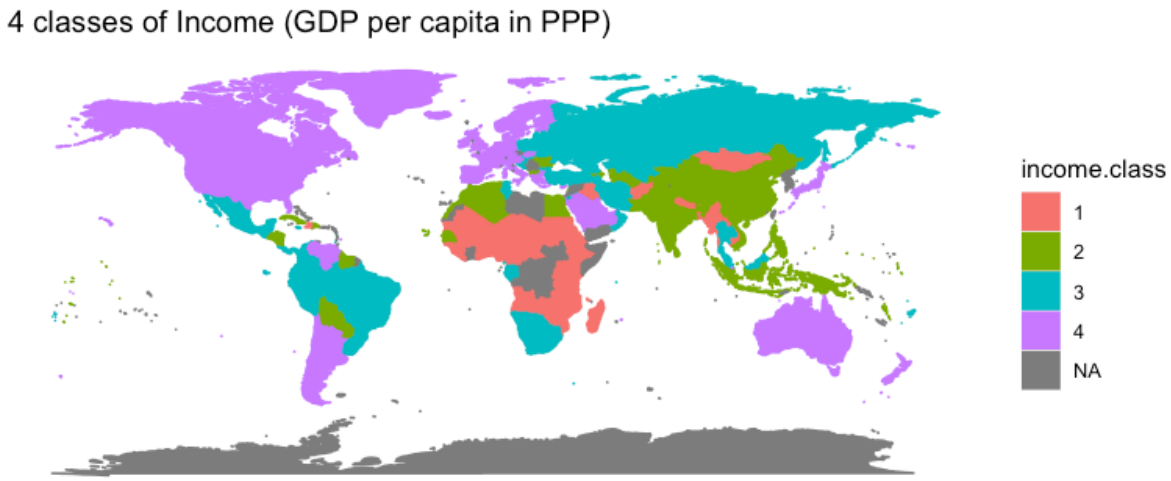

Figure 5: 4 classes of income

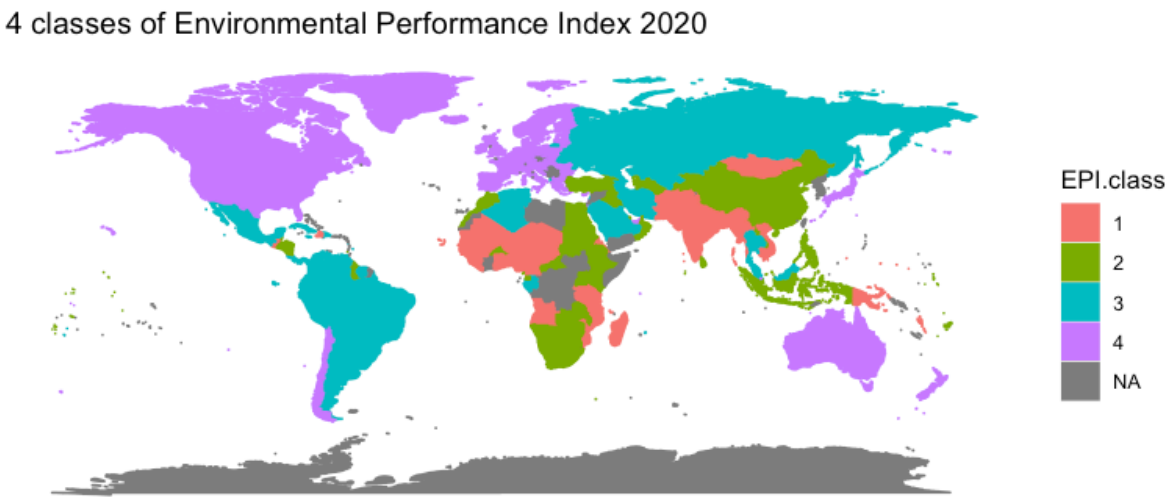

Figure 6: 4 classes of EPI

At a first glance, these 2 maps look all most the same, which indicates that the variable 'income' indeed is the most important factor to predict EPI class; especially for the higher classes 3 and 4, they are mostly correct, which is the same as shown by the decision tree.

When talking about environmental performance, we always expect there are other indicators should be more important for the environment, such as governance level, democracy levels, trade openness of a country, etc., less likely to expect income as the most important factor among those since sometimes it is expected that the development of the economy would worsen the environment.

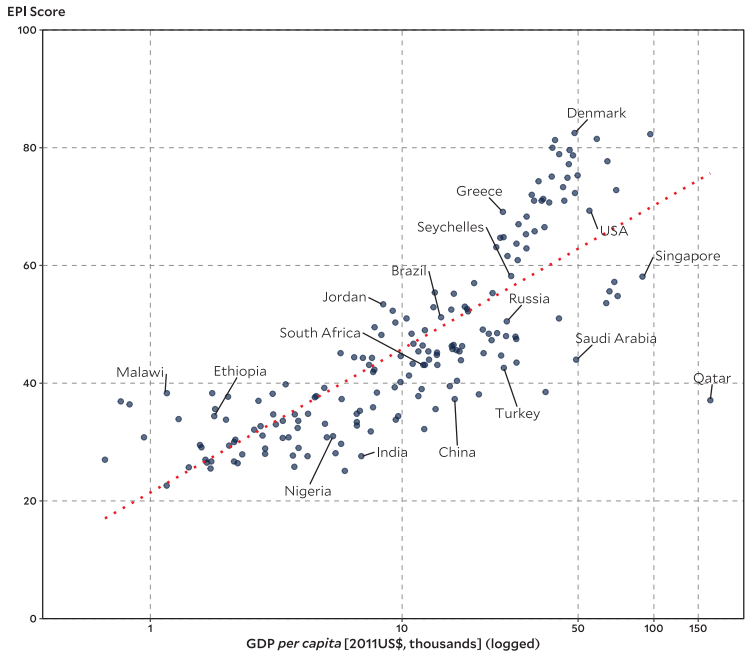

Figure 7: the scatter plot of the relationship between EPI and GDP per capita

(from https://epi.yale.edu/downloads/epipolicymakersummaryr9.pdf)

After I finished the analysis section of the paper, I examined the policy suggestion report by the EPI center, and I find out the conclusion reached in my paper is in line with the report of the EPI center, it also shows a strong positive correlation between the wealth and environmental performance (Figure 7), the center explained it as that economic prosperity makes it possible for nations to invest in policies and programs that lead to desirable outcomes. This trend is especially true for issue categories under the category of environmental health, as building the necessary infrastructure to provide clean drinking water and sanitation, reduce ambient air pollution, control hazardous waste, and respond to public health crises yields large returns for human well-being.

Another possible reason is that the variables included in this paper are limited, such as, social awareness, which is considered as a major contribution to improve the environment and always associated with higher income or higher levels of democracy, those factors have been proven statistically significant for the environmental performance by current studies.

Literature

Baettig, Michèle, Bernauer, Thomas (2009): National Institutions and Global Public Goods: Are Democracies More Cooperative in Climate Change Policy. *International Organization*. 63(2): 281-308.

Bernauer, T., A. Kalbhenn, V. Koubi and G. Spilker (2010): A comparison of international and domestic sources of global governance dynamics. *British Journal of Political Science*. 40(3): 509–538.

Kuhn, Max, Johnson, Kjell (2013): *Applied Predictive Modeling*. New York: Springer.

Maitra, Saikat & Yan, Jun (2008): Principle Component Analysis and Partial Least Squares: Two Dimension Reduction Techniques for Regression. *Casualty Actuarial Society*, 2008 Discussion Paper Program, https://www.casact.org/pubs/dpp/dpp08/08dpp76.pdf.

Stern, Nicholas, Sir (2007). Written Testimony to the U.S. Senate Committee on Energy and Natural Resources. *Stern Review of the Economics of Climate Change*. 110th Cong., 1st sess. 13 February.

Yale Center for Environmental Law & Policy: Environmental Performance Index. (https://epi.yale.edu/downloads/epipolicymakersummaryr9.pdf [17.07.2020]).