Heterogeneity in the income-emission relationship: Debunking homogeneity assumptions of the Environmental Kuznets Curve for CO_2

Aldo Javier Gómez Valdez

2661629

Thesis presented for the:

MSc in Spatial, Transport and Environmental Economics

Vrije Universiteit Amsterdam

Supervisor:

Prof. Dr. Henri de Groot

Abstract

The Environmental Kuznets Curve is a hypothesis stating that emissions initially rise as income per capita increases, then decrease after some level of development has been achieved. Graphically, the income-emission relationship should follow an inverted U-shape curve. Empirical estimations have resulted in mixed shapes of the curve and the income level that defines a turning point. Most studies estimate linear or panel versions of the model. These assume an homogeneous income-emission relationship across countries. Although some authors have pointed out that the relationship is heterogeneous across countries, no further research has been done to examine the sources. Despite non-parametric models can find evidence of the EKC, they are limited to unveil the interactions of other variables. We present a regression tree technique to analyze and classify CO_2 emissions for countries with differentiated economic characteristics. Our findings suggest that as countries develop, the increasingly complex economy configures an income-emission relationship in which per capita income is insufficient to account for different emission levels in countries with the same income level. Relevant variables are renewable electricity generation, the share of exports and industrial activity in GDP, population growth, and urban population.

Acknowledgements

I would like to express my appreciation: To my parents, siblings and the rest of my family for their continuous love and support. To Stina, who has been filling every step of this journey with love. To Prof. Dr. Henri de Groot for his advice on this research. To the VU Fellowship, the Orange Tulip Fellowship, the Holland Scholarship and Banco de México for their financial support. To my friends, especially Daniel, Fayçal, Dominika, Marta and Georgia, who along many others made this year delightful. To all the teachers in the STREEM programme for their dedication to our education and their efforts during the current COVID pandemic.

1 Introduction

One of the main concerns of environmental economics is to investigate under which conditions countries can reduce the impact of economic activity on environmental quality. Since there is no suitable scale to measure good or bad environmental quality, many studies examine variables that negatively affect the environment. Greenhouse Gas Emissions (GHE) are the primary pollutants contributing to Climate Change. CO_2 is the major release from human activities, accounting for two-thirds of the total GHE. [69]

Early works analyzing the implications of economic growth on the environment[1] point out the incompatibility of increasing the volume of economic activities with preserving natural resources and environmental quality. The Environmental Kuznets Curve (EKC) is a hypothesis formulated by Panayotou [54]. based on the empirical analysis carried out by Grossman and Krueger [32]. The hypothesis states that developing countries with low levels of income will initially increase their emissions alongside economic growth, and then decrease after some level of development has been achieved. Graphically, the income-emission relationship would follow an inverted U-shape curve. The EKC can be better understood by considering three different stages of development. First, economic growth in countries with low levels of income will decrease the quality of their environment as higher consumption and industrial activity augments emissions [63]. The second stage commences when the larger income and better distribution results in higher welfare. Society is more aware of environmental problems and demands better environmental quality.[41] When those demands result in either policy regulations [27], or market-driven shifts in the goods produced and the production quality [62], a turning point might arise. Emissions reach a maximum level and decrease as the economy grows beyond this level. Either because services become predominant in the sectoral composition of the economy [22], technological development improves production processes to become less emission-intensive [9, 19], or more resources from the society are allocated for cleaning the environment [25].

Since the EKC hypothesis was conceived, empirical estimations have yielded controversial findings. Different shapes of the income-emission relationship have been found for different

[1] Such as The limits to Growth [50] or the Brudntland report [10].

pollutants, geographical areas and periods.[2] Even studies finding the inverted U-shape curve rarely agree on the threshold value of income that represents the turning point [14, 26]. These differences have resulted in debate about the validity of the hypothesis. Supporters argue that results are affected by data limitations [2, 13, 16] and econometric methods [38, 71, 72]. A third research strain has raised the question of whether the turning point can be assumed as homogeneous across countries and periods. For instance, De Bruyn [20] estimates total CO_2, SO_2 and NO_x emissions for the Netherlands, West Germany, the United Kingdom and the United States as a function of total GDP. By performing a homogeneity tests on the coefficients associated with the turning point, the hypothesis of homogeneity is rejected. Hence, the income-emission relationship does not follow the same path for different countries, even if they have similar development levels. Under a Bayes estimator consistent with long time-series panel data only for CO_2, Musolesi et at. [53] find different relationships across countries according to the level of development. Less developed countries display an increasing monotonic relationship, while an N-shaped curve emerges for industrialized countries. Piaggio and Padilla [55] argue that imposing a homogeneous specification of the functional form and parameters dismisses the underlying differences in the relationships of countries with the environment. In their study, only 18 of 31 countries (OCDE, Brazil, China and India) are found to have a long-run relationship between CO_2 and economic activity, while only 14 of those achieve a within-sample turning point. Even by grouping countries according to the shape of the EKC curve, the estimated parameters prove to generate different shapes of the income-emission curve.

Although the previously mentioned studies provide evidence of heterogeneity of the EKC-related parameters, they do not investigate potential sources. The research question that motivates this thesis is to investigate how different variables apart from income can alter the shape of the income-emission curve, particularly the EKC. To our knowledge, the only study moving in that direction is carried by Jobert et al. [35] who make an attempt to classify countries according to the shape of the curve. However, their classifications are based only on the value of the estimated income parameters. Further interactions of other variables and classifications are discussed without providing statistical evidence.

We start by conducting several tests to the classical linear reduced form. Volleberg, et al. [71] suggest applying non-parametric techniques to deal with the identification problems and restric-

[2]For instance, see de Groot [21].

tions imposed by linear specification. We also estimate non-parametric and semi-parametric models and compare. Bernard et al. [5] discuss that, although non-parametric approaches are an improvement over linear methods by providing more robust tipping points, there is still work to be done on the identification of the curve. We contribute to the EKC literature by applying a regression tree model, a statistical tool used in data mining and machine learning, to investigate the heterogeneity of the income-emission relationship across countries as an alternative to identify the income-emission relationship..

The subsequent sections are divided as follows. Section 2 conducts a literature survey where the most notorious theoretical developments and econometric methods are addressed. Section 3 defines the models to be compared and describes the dataset, complimented by a preliminary data analysis. Section 4 presents the results of estimating the different specifications of the model and the outcome of the regression tree technique. Section 5 discusses the results and concludes.

2 Literature survey

2.1 Theoretical approximations

The EKC hypothesis was developed following empirical analysis unsupported by theory. Consequently, many authors have tried to develop theoretical explanations for the conditions required for an economy to display an inverted U-shape behavior. Three main approaches have been formulated. From the consumer perspective, a trade-off between consumption and pollution is unavoidable. This approach has been the concern of authors such as McConnell [49], Stokey [65], Andreoni and Levinson [4], Di Vita [24]; more recently, Ma and Shi [46] and Figueroa and Pastén [29]. Although the consumer behavior is an essential factor influencing not only consumption patterns but also environmental regulation [27], the estimation of such parameters in a cross-country study is difficult to achieve because of the idiosyncratic features.

The seminal works of Grossman and Krueger [32] and Panayotou [54] discussed the economic forces and interactions that could explain the EKC. First, a *scale effect* derived from the expansion of the economic activity that yields more pollution. Second, a *composition effect*

could occur due to differences in the sectoral distribution of economic activities and environmental regulations. The lowest-income countries will transit from an agriculture-intensive economy to develop industrial activity. Afterward, the services sector emerges. Third, the *technique effect* will change the embodied pollution per unit of production.

According Bousquet and Favard [7] and Mitić et al. [51], the *scale effect* is dominant in early stages of development. The second stage of the income-emission relationship starts when enough development has been achieved and the economy changes its structure[3]by *composition effect* shifting production towards less emission-intensive sectors [22]. On the final stage, economic growth will foster technological development and encourage the adoption of cleaner production processes and the implementation of abatement technologies [62].

More thorough theoretical approaches examine the *scale effect* by applying growth theories to relate the aggregated production with emissions or pollution. Such models are constructed based on neoclassical growth models. Macroeconomic formulations of the EKC incorporate technological parameters to depict its implications, regardless if it is treated as endogenous or exogenous. Thus these types of studies also tend to consider the *technique effect*.

A useful tool is the decomposition analysis used to disentangle changes in emissions derived from changes in structural economic variables such as sectoral contributions to the production and their energy intensity. Usually the analysis of the *composition effect* is investigated under this framework, relying more in empirical examinations, rather than theoretical developments.

2.1.1 Growth and the environment

The most prominent approaches to develop a theoretical framework for the EKC start from expanding the Neoclassical Growth Model (the Solow model) to either include the environment as a production factor, pollution as an outcome, or both. The nature of this model allows emphasizing the role of technological development in the income-emission relationship.

López [43] is the first author to develop a model focused on the elasticity of substitution between inputs and pollution shaping the EKC. He argued that under the extreme case of a zero elasticity of substitution, economic growth centered around improved productivity of cap-

[3]In a similar way as the one described by the original Kuznets curve [39].

ital and labor would lead to a monotonic increase increase in pollution. The *scale effect* will continuously dominate. Thus, the only mechanism to reduce pollution is reducing economic growth. However, "pollution-saving" technological change can foster "clean growth". He started by specifying a production function $y = G[f(K, L), x]$ where G is the production function, $f(\bullet)$ stands for an aggregator function of traditional inputs and x is the environmental factor.[4] A social welfare function is defined as $\mu = \mu[R[p; f(\bullet), x], x, p]$ where $R[\bullet]$ is the revenue function and p are prices. If pollution is priced at the marginal social cost, the equation $R_3 \equiv \partial R/\partial x = -\frac{\partial \mu/\partial x}{\partial \mu/\partial f} = q$ depicts the optimal internalization of pollution. Under non-homothetic preferences and separability between environmental and traditional factors of production with a CES function $R = A[\gamma_2 f^\rho + \gamma_3 x^\rho]^{(1/\rho)}$, by defining $a \equiv -(\mu_{11}R)/\mu_1$ as the coefficient of relative risk aversion of social welfare derived from the production function $G[\bullet]$, the EE relationship depends on $dx/df \gtreqless 0$ if $1 - \rho \gtreqless a(f)$. As López explains [43, pp. 171-172]:

> *"Economic growth increases the value of the environment for consumers. If this increased value is manifested in the market, firms will have to pay an increasing price for pollution. [...] The coefficient a, on the other hand, shows how the marginal utility of income declines as income expands. [...] Thus, if a is large the pollution price that consumers will demand will increase much more as income increases than if a is small, and vice versa".*

Since a captures the increase in welfare derived from production (considering the environment as a factor), it is a function of the aggregate production f. As production increases, $a(f)$ will approach the threshold value of $1 - \rho = 1/\sigma$ inherited by the CES preferences. When the increased production generates less welfare than the loss caused by diminishing environmental quality, the income-emission relationship and the EKC arises, as shown in Figure 1.

Stokey [65] was concerned about the implications of environmental policy on growth. Her model introduces technological change as both an endogenous and exogenous variable in an AK model. Policymakers try to maximize social welfare derived from consumption and pollution by setting technological standards z. The formulation allows for the possibility of sustained growth of the economy without environmental protection, but this is not an optimal outcome.

[4] He also considers index variables for technological change. Since they are exogenously determined and for simplicity of explanation they are omitted.

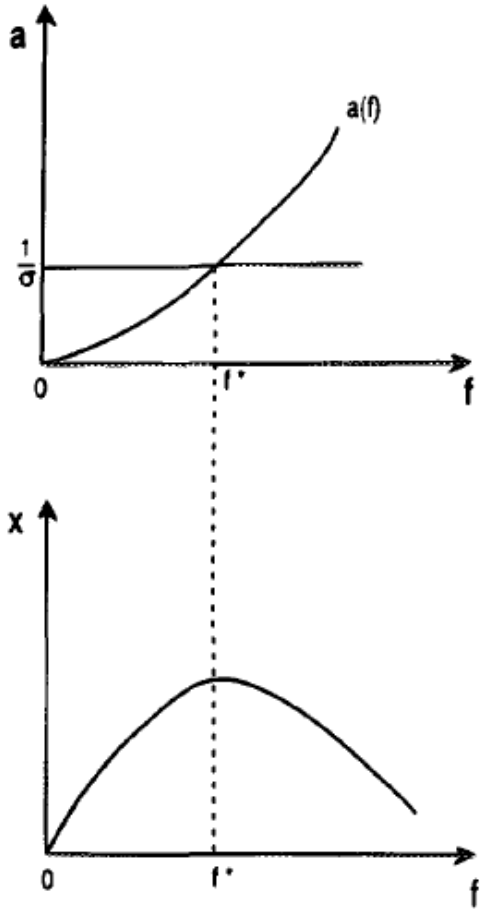


Figure 1: The EKC as a relationship of production, substitutability and consumers' risk aversion. *Source:* López [43, p. 172]

The system must converge to a steady state. With endogenous technological change, defining pollution during the transition phase as $x = AKz^{\beta}$ and if the capital stock surpasses a critical level, pollution can only decline if $\beta\frac{\dot{\lambda}}{\lambda} + (\beta - 1)\frac{\dot{k}}{k} < 0$, where β is the elasticity parameter of pollution w.r.t. technology, and λ is the shadow value of capital. To sustain the rate of return to capital as capital stock grows, increasing pollution is required. Imposing stricter regulation in z lowers the rate of the return of capital; accumulation eventually ceases and growth stops. The exogenous technical change is not only set by the social planner but also improves at a constant rate, surpassing the growth rate of capital. Hence, growth is possible even with strict environmental regulation.

The previous models' conclusions regarding the EKC heavily relied on explaining the properties of the utility function for both producers and the social planer. This analysis required strong assumptions not only on the specification of equations but also on the behavior of the parameters. Thus, the capacity to analyze the interactions of the *scale*, *composition* and *tech-*

nique effects is limited [44] and the results of empirical examinations could be inaccurate. This issue is addressed by Brock and Taylor [9], who adapt the Solow model to incorporate technological progress in abatement and abatement costs. Although Stokey [65] and Dinda [25] included technological progress as part of their models, it was tied to the availability of capital, whereas Brock and Taylor [9] assume that technology is developed with the specific purpose to reduce pollution.

The Green Solow model is elaborated to illustrate a single sector economy. The pollution equation is specified as $E = \Omega F - \Omega A(F, F^A) = \Omega F a(\theta)$, where F is the scale of economic activity, F^A is the abatement efforts, A is the abatement level, Ω is the share of pollution abated from the total created, and $a \equiv [1 - a(1, F^A/F)]$, $\theta = F^A/F$,. The intensive measures of output, capital and pollution are $y = f(k)[1 - \theta]$, $\dot{k} = sf(k)[1 - \theta] - [\delta + n + g_B]k$ and $e = f(k)a(\Omega)$. By proceeding as in the regular Solow model, the balanced growth path is given by $g_E = g_B + n - g_A$, where $g_B + n$ represents the *scale effect* of the rate of production raising effective labor g_B and population growth n. g_A captures the *technique effect*.[5] Figure 2 graphs the rates of change of capital and emissions by α times. As capital increases, the accumulation process causes economic growth accompanied by an increase in pollution. As the economy approaches the steady state in T,the increase in pollution is slowed down. Continuous economic growth is only possible through technological development. Improvements in abatement technology will accompany.

Since the Green Solow model assumes that abatement technology is exogenous, it does not provide information about how producers and policymakers respond to environmental degradation. That is precisely the aim of Smulders et al. [62], who further explore the incentives of heterogeneous producers to adopt specific technologies that have two implications. They can either improve the quality of goods produced or be pollution-saving. In any case, they require investment in research and development. Thus, firms choose the technology best suited to maximize profits.

As seen in Figure 3, Smulders et al. [62] further divide the phases of economic development into four sections according to not only the shape of the EKC but also on the associated production-pollution characteristics of the technology. In the first phase, a single technology

[5]The Green Solow model is elaborated to illustrate a single sector economy. Hence, the *composition effect* cannot be evaluated. The authors recognize that it can be further extended to illustrate a more diversified economy.

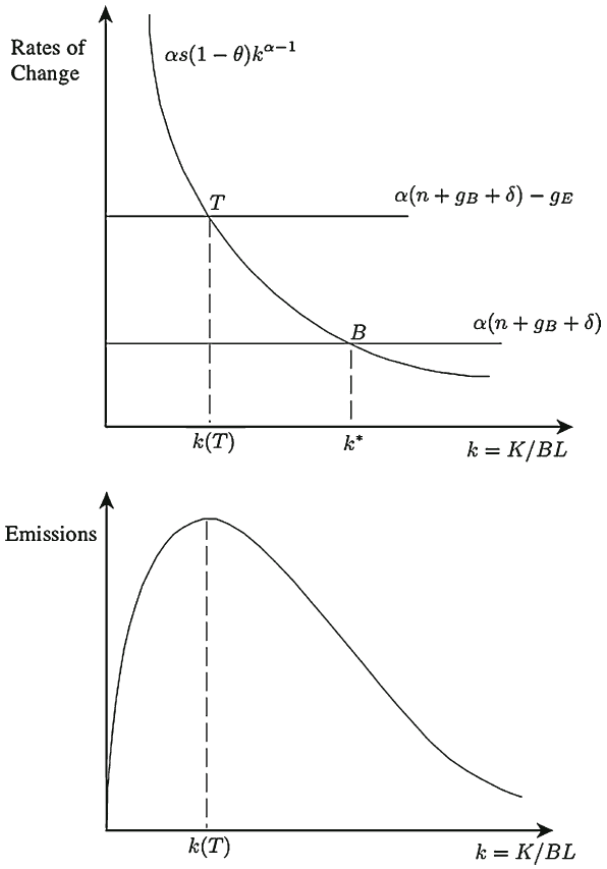


Figure 2: The EKC relationship of total emissions by pollutant and year generated by growth rates. *Source:* Brock and Taylor [9, p. 137]

is used and generates no pollution. In the second phase, new technology is available and it is labor augmenting. Immediately, the increased competition among producers incentivizes the development of product improvements. Pollution rises along with the *scale effect*. On the alarm phase, environmental degradation is evident and the governments develop environmental policy, leading to the cleaning-up phase, where environmental technology is adopted. Since technological affects the market performance of producers, the less efficient ones will eventually leave the market.[6] Thus active policy might generate market incentives for the adoption of cleaner technologies and achieve the cleaning of the environment. Empirical studies, such as Popp [56] and Aghion [1] support this argument.

[6]This argument is further explored by Cherniwchan et al. [15] who found that in open economies, less efficient and dirty exporters might even exit the market due to cost constraints.

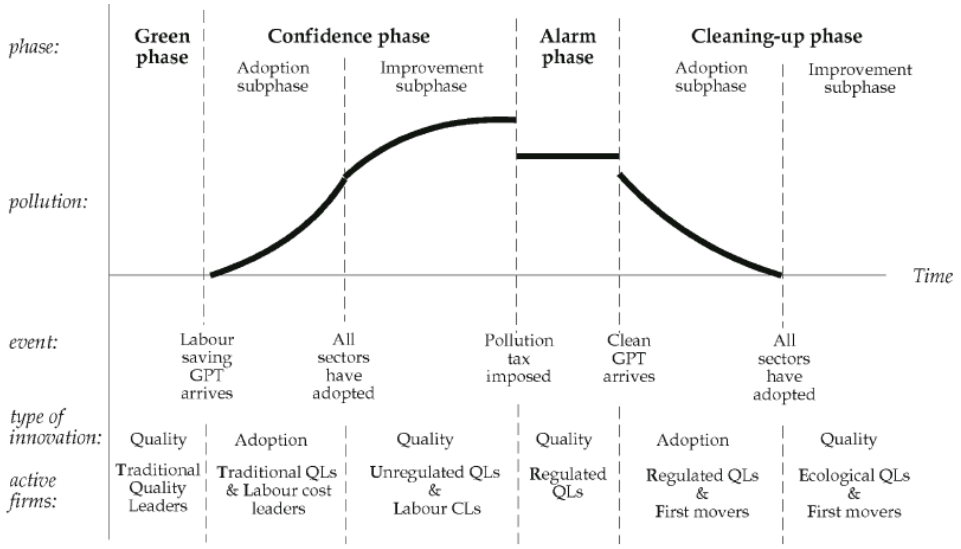


Figure 3: The four phases of technological development and the EKC. *Source:* Smulders et al. [62, p. 86]

2.1.2 Decomposition analysis

The previously exposed theoretical approaches consider only the general specification of a single sector economy. To further examine the implications of the sectoral composition of the economy, the decomposition analysis of emissions can be done by defining an identity equation. Grossman [31] proposed to use $E_t = \sum_i a_{it} s_{it} Y_t$, where Y_{it} is the scale of economic activity at time t, s_{it} is the sector i share of output and a_{it} is the pollution generated by a unit of output in the corresponding sector. Since $s_{it} = Y_{it}/Y_t$ and $a_{it} = E_{it}/Y_{it}$, it is clear that $E_t = \sum_i E_{it}$. By differentiating with respect to time and dividing by E_t the equation $\hat{E} = \hat{Y} + \sum_i e_i \hat{s}_i + \sum_i e_i \hat{a}_i$. Where $\hat{x} = (dx/dt)/x_t$ and e_i is the sectoral share of emissions.[7] The final equation allows to examine the *scale* ($\hat{Y}$), *composition* ($\hat{s}$), and *technique* ($\hat{a}$) *effects*.

A second specification of the identity equation is the Kaya identity[8] [36]
$E_t = \frac{E_t}{FEC_t} \frac{FEC_t}{TEC_t} \frac{TEC_t}{GDP_t} \frac{GDP_t}{P_t} P_t$ where FEC_t is fossil energy consumption, TEC_t is total energy consumption and P_t is population. Papers who use this methodology usually find that the *technique* offsets the *scale* [64]. This means either a low emission intensity of the main economic sectors of a country or that some components of the $\frac{E_t}{FEC_t} \frac{FEC_t}{TEC_t} \frac{TEC_t}{GDP_t}$ part of the identity dwindle

[7] This approach has also been used by Ekins [26], de Bruyn [20], Bruvoll and Medin [11], and Tsurumi and Managi [68].

[8] More commonly used in studies focusing on energy production. Luzzati, Orsini and Gucciardi [45] combine the Kaya analysis with the empirical estimation of the EKC.

while GDP_t and population increase or are held constant.

De Groot [22] develops a link between growth theory and the decomposition analysis through developing a multi-sector general equilibrium model to analyze the implication of different technologies and income elasticities for the demand for goods produced by each sector. To introduce the demand for different goods, the utility function is specified as $U = [\sum_{i=1}^{S} a_i(C_i - \bar{C}_i)^{\rho}]^{1/\rho}$, where $i \in S$ denotes each sector, a_i is the distribution parameter, C_i is consumption and $\bar{C}_i$ is a subsistence requirement of consumption. The budget constraint is given by $C_i P_i \leq Y_i$. Production is defined as $Q_i = [b_{Li}(h_{Li}L_i)^{\sigma} + b_{Ei}(h_{Ei}E_i)^{\sigma}]^{1-\sigma}$ where b_{Li} and b_{Ei} are share parameters of labor L and emissions E respectively, and are affected by labor-augmenting technological progress in the form of productivity denoted by h_{Li} and h_{Ei}. Profit maximization allows to specify emissions as a Marshallian demand for productive inputs. Thus, under market equilibrium, emissions can be expressed as a function of labor and its relative prices, consumption and its shares, as well as productivity parameters.

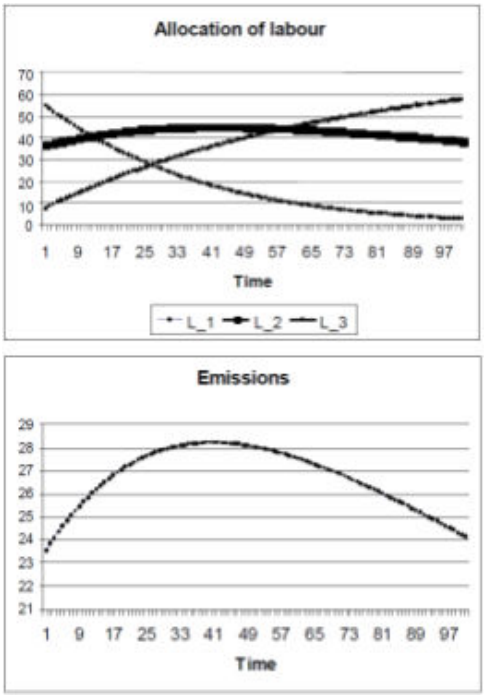


Figure 4: Example of the structural composition and the emergence of the EKC. *Source:* De Groot [22, pp. 23-24]

The decomposition analysis used by De Groot [22] formulates the identity $E = C\sum_{i=1}^{S} \frac{E_i}{L_i}\frac{L_i}{C_i}\frac{C_i}{C}$ resulting in the growth rate of emissions as $\hat{E} = \hat{C} + \sum e_i \hat{p}_i \sum e_i l_i + \sum e_i \hat{s}_i$, where p_i is the emission labor ration and l_i is the inverse of the productivity of labor. Similar to Grossman [31], this formulation can be divided into three components. Albeit the

technique effect can be approximated through labor productivity. This method can yield different results that can result (or not) in an EKC according to the allocation of labor in the three economic sectors. Figure 4 illustrates one of these cases.

2.1.3 Critiques

The EKC hypothesis is the subject of several critiques. Most studies rarely take into account that the income-emission relationship might be a dynamic process on which pollution affects economic performance [64]. Irreversibility on the damages caused to the environment affects its capacity to absorb pollutants. Thus, even if the flow of emissions declines, the environment is not necessarily improved [41].

The fact that the inverted U-shape curve has been found for some pollutants but not for others might be due to evolving production processes that, over time, shift pollution from one pollutant to another. For instance, the decline in diesel in favor of gasoline engines shifts emissions from CO to CO_2. This might be attributed to technology changes or unpairementd regulation that focuses on one type of pollutant, which might cause aggregate emission levels to rise [64].

An explanation for the N-shaped curve is that as the economy internalizes the costs of pollution, it effectively decreases. However, when the internalization process is complete, the scale effect resurges and pollution rises once more. [41] The costs to society merely become monetary transfers.

The *Pollution Heaven Hypothesis* itself constitutes a critique to the EKC since it argues that environmental regulations in one country will motivate the migration of dirty industries to less-developed countries [17]. The *Race to the Bottom Hypothesis* is an extension in which less-developed countries compete to reduce the stringency of environmental regulations in order to attract investors.[9]

One of the critiques that have not received enough attention in the literature is the analogy between the original Kuznets Curve (KC) to the EKC. The original Kuznets work focused on income distribution. Although including the Gini coefficient or other inequality measures could contribute to the EKC examination [7], cross-country studies rarely have access to consistent

[9]Both are confronted by Rasli et al. [58].

data. Even our dataset extracted from the World Development Indicators does not provide enough data points to include this consideration.[10]

2.2 Econometric methods

The seminal papers by Grossman and Krueger [31] and Panayotou [54] set the traditional polynomial equation that many posterior studies will replicate. The assumed functional form of the income-emission relationship is $E_{it} = \beta_0 + \beta_1 Y_{it} + \beta_2 Y_{it}^2 + \beta_3 Y_{it}^3 + \gamma Z + v_i + \varepsilon_t$, where E are emission variables, Y can be economic activity or income measures such as GDP or gross added value, Z is a vector of control variables and v_i are geographical fixed effects. In order to control for different economy sizes, emission and income variables are usually expressed in per capita units. When $\beta_1 \neq 0$ and $\beta_2 = \beta_3 = 0$, a monotonical relationship is found. The peak of the curve or the turning point is computed by partially differentiating E with respect to Y_{it} and seeking the maximum. That is $Y_{it}^* = -\beta_1/2\beta_2$ and $\beta_3 = 0$.

The linear specification entails some basic pitfalls. As List and Gallet [42] argue, the panel approaches usually assume that the EKC shape is the same for different countries, while the true values can be conditional on the time and geographical areas of study, as shown by [5, 35, 55]. If there is no heterogeneity in the level of development of the countries within the sample, omitting lower-income countries will result in models sensitive to sample characteristics. Estimations will be biased and inconsistent [38, 41] and lack external validity [64]. Furthermore, the quadratic specification implicitly assumes that the shape of the EKC is completely symmetrical [38].

Many authors have noted that the controversy on the EKC estimation can be attributed to weak or restrictive econometric analysis [41]. For instance, Grossman and Krueger [32] found the inverted N and U-shape curves emerging between income per capita and emissions of SO_2 and smoke, depending on the inclusion of fixed effects or not. Suspended particles were found to decline as income increases in both specifications. Nevertheless, they estimated turning points ranging from \$2,000 to \$5,000 1985 per capita USD. Panayotou's [54] estimates for

[10] Ridzuan [60] utilizes the Standardized World Income Inequality Database which is a collection of inequality data from different sources such as national databases, published articles, and imputation methods. As the author notes, the dataset has been criticized for its reduced reliability. For this reason, this thesis does not include this variable.

SO_2, NO_x, CO_2, suspended particles and deforestation on a larger sample of countries without country fixed effects found lower turning points in the range of $800 to $5,500 per capita 1987 USD. Many subsequent works also find different results [14, 21, 38, 51, 57, 63].

One common issue across early empirical analyses of the EKC is that they either do not include additional control variables and do not report cointegration statistics to test the presence of omitted variable bias [64]. The omitted variable bias is probably one of the reasons why results are different across studies. The most common way around is to include higher-order polynomials of income. Although imperfect multicollinearity is not a severe problem for econometric estimations, the standard errors will be larger and increase the chances to commit a Type II error [20, 41]. Because of the multicollinearity caused by the high order polynomials, if control variables are included, they will have little variance left to explain the model. Reduced-form estimations will operate under correlation, not because of causality [38].

Other potential issues are heteroskedasticity, simultaneity and cointegration issues [64]. Countries with high GDP and population will usually display a smaller residual term, causing heteroscedasticity. The simultaneity arises both from the logic of the hypothesis and one of its critiques, as pollution might cause a feedback effect on production. Cointegration problems are pointed out by Vollebergh, et al. [71] who argue that one explanation of the lack of robustness on the estimates comes from the fact that both pollution and income trends are time-dependent variables and many attempts to isolate the effect through imposing restrictions on the functional form will throw results highly dependent on the chosen functional form. Time effects are assumed to affect equally all countries and, thus, the relationship is assumed to be equal.

Panel approaches improve upon the pooled estimations as they allow for incorporating cointegration and unit root considerations [23, 53, 55, 59, 74] in the modelling process. In combination with improvements in the availability and quality of the data, recently published papers achieve more accurate estimations of the EKC [57]. However, the panel estimations of the EKC still face many issues, such as different integration orders, cross-sectional dependence, endogeneity and parameter heterogeneity [72, 57].

Earlier limitations of data noted by Carson [13] are no longer a concern. International organizations such as the World Bank, The OCDE and the European Union have improved upon limitations on the availability and comparability of data across countries and periods. At the

same time, they offer significant information about data quality for researchers to make decisions about the sampling.

The problems present in the linear estimation of the EKC can be faced by testing the presence of non-linear relationships, or by combining parametric and non-parametric estimations [57]. Volleberg, et al. [71] demonstrate that linear EKC estimations can result in differentiated outcomes according to the assumed functional form of the independent variables and income, leading to non-robust estimations. By applying a Bayesian estimator combined with a pairwise estimation improves the robustness of results[11] to avoid imposing restrictions on the functional form, the equation to estimate is $E_{rt} = f(Y_{rt}, r) + \lambda(r, t) + \varepsilon_{rt}$, where $f(Y_{rt}, r)$ is the non-parametric function of income for the geographical area r, and $\lambda(r, t)$ isolates the time effects from the true income-emission relationship. Under a similar formulation, Bernard et al. [5] find that time and cross-sectional inconsistencies can generate differences in the shape of the curve. In their study, even non-parametric techniques are also limited for generating an homogeneous EKC across different subsamples. Andreé, et al. [3] reaches this conclusion and argues that local economic conditions can be relevant to determine environmental results.

The heterogeneity of the parameters can be dealt with by estimating a random-coefficient model as carried out by Jobert [35]. Because many countries lack consistent emission data for more extended periods, such approximation might not be possible for studies focusing on many countries due to limitations on the rank condition necessary to achieve identification. This condition creates a double-edged sword for the estimation of the EKC. In order to estimate heterogeneous parameters, the time T should be significantly larger than the number of countries N included in the sample. As noted by Zoundi [74], traditional panel methods applied to such data can generate spurious results as their closeness to time series will indicate cointegration rather than causality. In the opposite case (large N and small T), researchers must assume parameter homogeneity.

[11] A similar approach taken by Jobert, et al. [35] was to group countries according to the shape of their individual curves.

3 Empirical strategy

We propose a four-step procedure to answer the research question of this paper. First, we conduct several tests to the linear specification of the EKC to analyze whether it is an appropriate approach or not. The linear tests follow Chapter 6 of de Bruyn [20] complemented by contributions by other authors. Second, we develop a semi-parametric model similar to Vollebergh et al. [71] and Bernard et al. [5]. We omit the pairwise approach and include linear effects on variables different than income to formulate a semi-parametric specification. On the third step, we conduct the Hsiao [33] test to analyze the heterogeneity of the income-emission relationship in our sample.[12] Because the heterogeneity hypothesis cannot be rejected, we apply a regression tree[13] method to examine the differentiated influence on emissions of the variables considered in our models.

3.1 Data description

The dataset was extracted from the World Bank Development Indicators, which contains data for 217 countries. After filtering the dataset following criteria similar to Mankiw, Romer and Weil [47], the final sample includes information for 91 countries for the period 1990-2014.[14]

The dependent variable is CO_2 emissions (metric tons per capita). The explanatory income variable is GDP per capita, PPP (constant 2017 international USD). Power Purchasing Parity units of GDP were selected to keep consistency across observations and avoid measurement errors caused by transforming currencies. To address the relationships considered by the decomposition analysis, the added value of the three-sector model is measured as a percentage of GDP. Those are agriculture, forestry, and fishing (Sector 1); industry, including construction (Sector 2); and services, including public sector activities (Sector 3). We attempted to proxy the *technique effect* by including the variables of secondary and tertiary educational attainment

[12] The procedure of the test is offered in Appendix A.2.

[13] AppendixA.4 summarizes the technique. For further explanation Breiman et al. [8] or James et al. [34] are suggested readings.

[14] The criteria to filter the sample can be consulted in Appendix A.1.

as bot the the percentages of total population over 25 years old and working population. Results were often not significant or robust enough to be included in the final model. Additional controls are exports of goods and services as a percentage of GDP (to illustrate the *Pollution Heaven* and *Race to the Bottom* hypotheses), the percentage of the urban population, population annual growth rate, and the renewable electricity output as a share of total electricity output.

The resulting panel, although not perfectly balanced, contains full information for the CO_2 emissions and GDP per capita for all countries during the period 1995-2014. 72 countries have observations from 1990, and 10 from 1992. The remaining countries' observations start at different years in the interval 1991-1996. Because the 72 countries for which we have full information represent 79.12% of the observations, there are no discontinuous jumps in the data, and to avoid complications not required by the non-parametric or the regression tree methods, the panel data will be treated as balanced.

3.2 Summary statistics and stylized facts

Figure 5 presents the income-emission relationship for the full sample of 91 countries in the 1990-2014 period. Despite the different sources, the scatter resembles the one presented by Jobert, et al. [35]. As expected, emissions per capita generated by low-income countries are lower than those of the rest of the world. Unfortunately, this country group is the lesser represented in the dataset due to the sample removal criteria. Most of the low-income countries lacked an acceptable data quality score, according to the World Bank's Statistical Capacity Indicator.

It is important to note that the heterogeneity of the income-emission relationship increases alongside income. A divergent speed in the increase of emissions with respect to income can be observed. An upper (faster increase of emissions) and lower (less emission-intensive increase of income) bounded tendencies can be observed. For the lower bound an apparent EKC is emerging. These patterns suggest that some countries are linking their economic growth to the environment. In other words, those countries either rely on natural resources or pollution-intensive activities to sustain economic growth. Other countries' growth is sustained by different mechanisms.[15]

[15]The linking-delinking hypothesis is similar to the EKC. This approach, however, is less theoretically devel-

Figure 5: Scatter plot of CO2 emissions per capita in metric tons and GDP in 1997 constant international USD. Full sample.

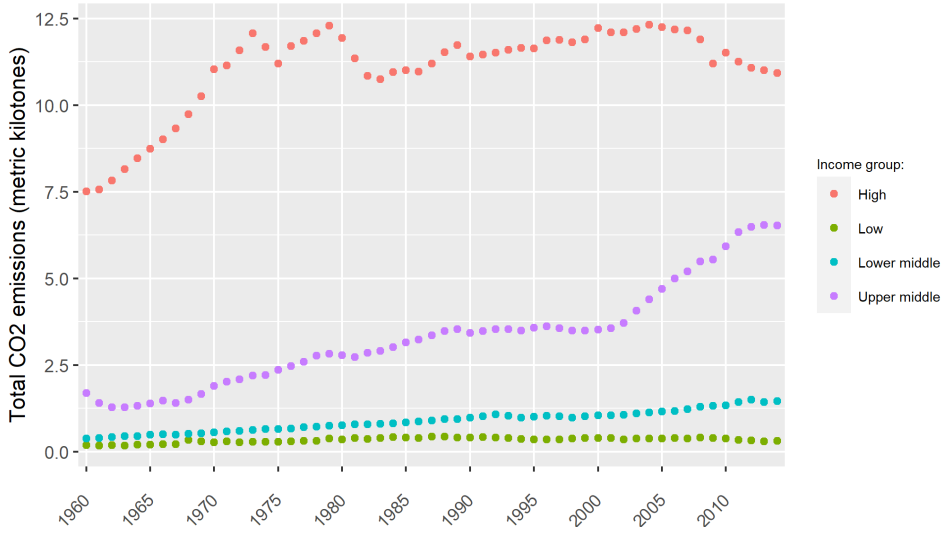
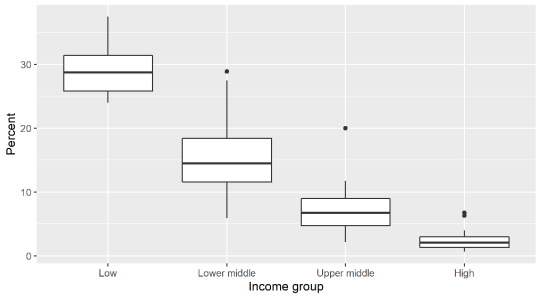


Figure 6: Evolution of total CO2 emissions across time per income group.

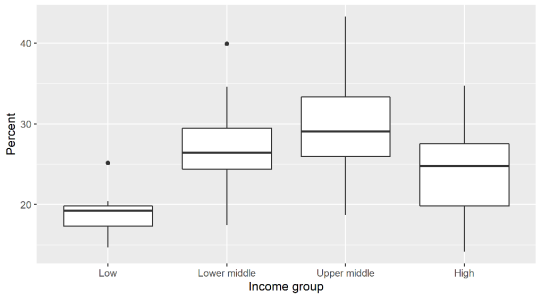
More developed countries in the upper bound of the scatter plot, with CO_2 emissions above 15 metric tons per capita are Australia, Canada and the United States for almost all the periods. Although not entirely oil producers, they are known for their exhaustive energy use and extraction of natural resources. The High income developed countries with CO_2 emissions lower than 5 metric tons per capita for most periods are Panama, Uruguay, Chile, Latvia and Lithuania. Croatia and Portugal, whose emissions were less than 6.5 metric tons per capita. Figure 6 illustrates how these "newly" high-income countries might be driving down total emissions for their income group in recent years. Middle income countries have increased emissions,

oped. See, for instance, de Bruyn [20], Cumming and von Cramon-Taubadel [18] and Marin and Mazzanti [48].

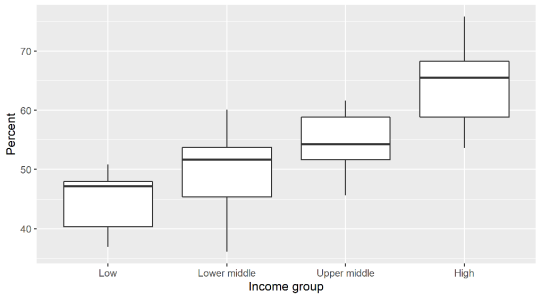
arguably because of industrialization processes.



(a) Agriculture, forestry, and fishing



(b) Industry (including construction)



(c) Services

Figure 7: Sectoral share of the value added in the GDP per income group, 2014.

Figure 7 shows how the share of industrial activity is relatively larger on Upper-middle income countries, drawing a hum shaped curve. The agriculture and the services sectors are configured similarly to the expectations by Kuznets [39] and de Groot [22]. These facts illustrate how industrial, developed and developing economies might have linked their economic growth to take advantage of the environment, while service-oriented economies might have delinked their growth. The standard empirical estimation of the EKC might undermine the different types of processes that explain these interactions and divergence across economic sectors.

Table 1 presents the summary statistics divided by income group. Besides the increase in emissions as income increases, exports also represent a progressively larger share of the GDP for high-income countries. Population is more concentrated in urban areas for these countries too. However, population growth rates decline with income. These facts are widely spread in economics but necessary to keep in mind for further analysis.

Table 1: Summary statistics by income level.

Low income	Mean	S.D.	Min.	Max.
CO_2 per capita	0.088	0.035	0.036	0.321
GDP per capita	1,235.345	469.973	436.719	2,536.6
Share of S1	32.299	6.705	18.478	53.283
Share of S2	17.977	3.396	10.415	28.372
Share of S3	42.822	5.942	22.328	62.117
Exports	18.041	6.668	5.585	35.66
Population growth	2.98	0.767	0.251	8.118
Urban population	21.704	6.923	11.076	39.196
Lower-middle income				
CO_2 per capita	1.281	1.653	0.05	13.447
GDP per capita	4,422.258	2,038.589	1,109.236	10,980.33
Share of S1	20.508	9.03	5.488	51.853
Share of S2	26.251	5.704	12.646	52.152
Share of S3	46.118	7.391	22.956	61.064
Exports	32.574	14.558	5.908	86.405
Population growth	1.858	0.848	-1.007	3.541
Urban population	42.944	14.491	15.437	68.968
Upper-middle income				
CO_2 per capita	3.866	3.181	0.223	15.94
GDP per capita	11,469.56	5,059.502	1,411.806	26,603.01
Share of S1	10.503	6.32	2.098	52.346
Share of S2	30.802	6.929	15.347	48.53
Share of S3	51.509	7.199	22.922	68.162
Exports	33.229	19.719	6.598	121.311
Population growth	0.983	1.36	-9.081	7.786
Urban population	59.958	16.805	18.196	91.377
High income				
CO_2 per capita	8.237	3.874	1.09	20.179
GDP per capita	34,313.5	13,041.51	9,492.153	66,038.73
Share of S1	3.007	2.046	0.554	10.997
Share of S2	26.078	4.862	13.682	41.107
Share of S3	61.372	6.077	46.609	76.444
Exports	39.863	19.25	8.972	110.025
Population growth	0.517	.774	-3.848	6.017
Urban population	74.58	11.808	47.915	97.833

Note: Full sample, all years.

3.3 Specification of the models

The first specification to be tested is a pooled OLS estimation of the EKC:

$$E_{it} = \beta_0 + \beta_1 Y_{it} + \beta_2 Y_{it}^2 + \beta_3 Y_{it}^3 + \gamma Z + \varepsilon_{it} \tag{1}$$

E stands for CO_2 emissions, Y for GDP per capita and Z the vector of control variables.

The panel estimation modifies (1) to include country v_i and time-specific τ_t fixed effects:

$$E_{it} = \beta_0 + \beta_1 Y_{it} + \beta_2 Y_{it}^2 + \beta_3 Y_{it}^3 + \gamma Z + v_i + \tau_t + \varepsilon_{it} \tag{2}$$

The semi-parametric approach follows a basic specification:

$$E_{it} = f(Y_{it}) + \lambda(\tau_t) + \gamma Z + \varepsilon_{it} \tag{3}$$

Where $f(Y_{it})$ and $\lambda(\tau_t)$ are unknown functions of income and time; Z are the control variables with an assumed linear effect on emissions.

For the final step of applying a regression tree technique, no functional form can be defined a priory nor can be inferred by the data structure.

For ease of presentation, we will refer to equation (1) as Levels, equation (2) as Panel and (3) as Semi-parametric. To compare the results of the tests, these will be adjusted across sections to either include or exclude control variables and different income polynomials. In order to select the best transformation of variables, several versions of CO_2 and income will be tested on equation (1). These includes levels (as reported in the dataset), logarithmic transformation and a first difference approaches.

4 Estimation results

4.1 Analysis of the linear specification

We start by testing the implications of using different transformations of CO_2 emissions and GDP for equation (1). The different transformations are levels, first differences, logarithmic and the first difference of the logarithm. The first difference polynomials for levels and logarithms are computed as $\hat{x}_t^q = (x_t - x_{t-1})^q$. All the specifications were estimated by using robust standard errors unless a specific test required the contrary.

As Table 8 in Appendix A.3 shows, including high order polynomials induces collinearity in the linear estimation of the EKC. Only perfect multicollinearity is a concern for modern econometrics. However, we note that including higher-order polynomials causes great increase in the VIF indicator. Thus, we are conservative in deciding to restrict the linear specification of equations (1) and (2) as a third-degree polynomial in GDP.

Second, we test the presence of autocorrelation in order to avoid spurious inference. Only the levels and logarithmic specifications were tested because the first difference approach already corrects this problem to some extent. We use the Wooldridge test for autocorrelation in panel data under the H_0 of no serial autocorrelation.[16] We find that both specifications suffer from panel autocorrelation, being more pronounced in the logarithmic model with a higher F-statistic.

Third, by using equation (2), we perform a cointegration analysis to determine whether the actual effect can be attributed to cross-sectional cointegration or time. Authors who apply a panel cointegration approach seek to find if a long-run income-emission relationship can arise [28, 40, 53, 61, 74, see]. To have a first glimpse of the homogeneity of the parameters is valid, it is necessary to examine if the cointegration is country-specific or can be observed across cross-sectional units. We apply the Westerlund test applicable for unbalanced panels, including (or not) time trends. The results are presented in Annex A.3, 10. We find that when time trends are not considered, the null hypothesis of no cointegration is rejected. However, by including time trends, the null hypothesis cannot be rejected. According to Hsiao [33], first differencing

[16] Results presented in Appendix A.3, Table 9.

the variables in a model leads to removal of the long-run relationships. Thus, autocorrelation is difficult to correct within our panel. We interpret these results as an indicator that there might be a long-run process governing the income-emission. However, it can be different for different types of countries.

Table 2: Panel estimation results.

Variables	Levels	Logarithmic	Differences	Logarithmic differences
GDP	0.000496***	-2.713*	0.000257***	0.796***
	(3.61e-05)	(1.433)	(2.94e-05)	(0.0916)
GDP^2	-9.81e-09***	0.493***	-6.80e-09	0.724***
	(1.08e-09)	(0.161)	(8.76e-09)	(0.259)
GDP^3	7.09e-14***	-0.0225***	-1.17e-11***	-2.126***
	(1.25e-14)	(0.00600)	(2.77e-12)	(0.453)
Share of S1	0.0207***	-0.00102	-0.0126***	-0.000274
	(0.00747)	(0.00246)	(0.00440)	(0.00160)
Share of S2	0.0559***	0.00742***	-0.00621	0.000531
	(0.00851)	(0.00172)	(0.00465)	(0.00130)
Share of S3	-0.00398	-0.00617***	-0.00564	-0.000301
	(0.00800)	(0.00162)	(0.00346)	(0.00123)
Exports	-0.0137***	0.000408	0.000424	0.000530
	(0.00331)	(0.000537)	(0.00143)	(0.000330)
Urban population	0.0824***	0.0162***	0.000719	-0.00167*
	(0.00750)	(0.00172)	(0.00425)	(0.000938)
Population growth	0.121***	0.0108	0.0215	-0.000245
	(0.0389)	(0.0114)	(0.0227)	(0.00396)
Renewable electricity	-0.0179***	-0.00654***	-0.00245**	-0.00115***
	(0.00194)	(0.000636)	(0.000954)	(0.000280)
Constant	-4.862***	1.115	1.023***	0.217*
	(0.771)	(4.214)	(0.382)	(0.119)
Country FE	YES	YES	YES	YES
Year FE	YES	YES	YES	YES
Observations	2,101	2,101	2,032	2,032
R-squared	0.981	0.993	0.186	0.184

Notes: Linear: $GDP = x$; Logarithmic: $GDP = \ln x$; Differences: $GDP = x_t - x_{t-1}$;

Logarithmic differences: $GDP = \ln x_t - \ln x_{t-1}$

Robust standard errors in parentheses.

p<0.01, ** p<0.05, * p<0.1

After testing the properties of the linear regressions, we conducted a panel estimation with country and year Fixed Effects for the different transformations of emissions and GDP. By estimating the model in differences, the long-run relationship indeed vanishes and the resulting R^2 reflects this situation. Although, both the Levels and Logarithmic estimators seem to fit the data, we decide to keep the estimation in levels as the preferred specification. The autocorrelation found on the Logarithmic was too high to be ignored, while it also suffered from a higher degree of collinearity.

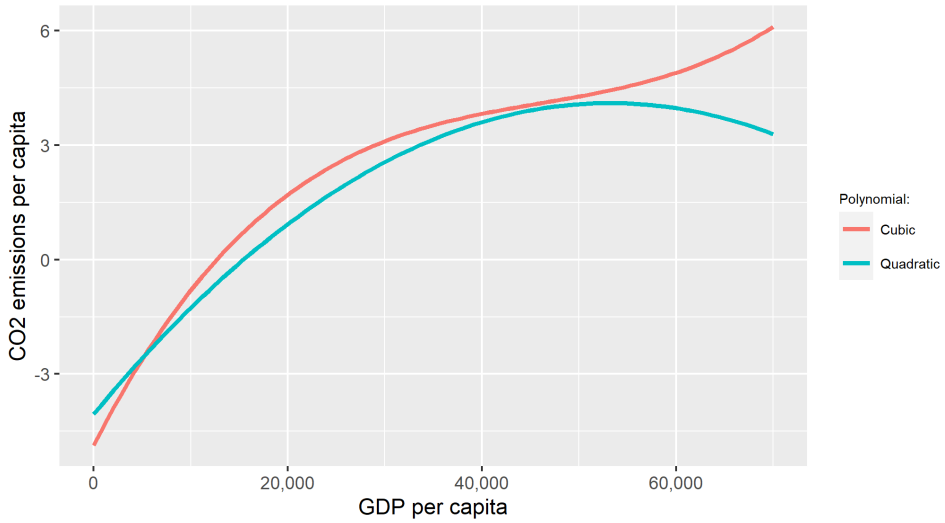


Figure 8: Estimated functions for the quadratic and cubic specifications.

The linear estimation of the EKC computes the turning point as the value of GDP that satisfies the first order condition $\frac{\partial E}{\partial GDP} = 0$. The derivative of our estimated function in Levels does not have a real number solution. As Figure 8 shows, there is no local minimum in our cubic estimation. For comparison, we estimated a quadratic specification of the model for which the turning point is found at 53,125.00 per capita units of GDP. Around the quadratic income level that achieves the turning point, increase in emissions of the cubic specification seems to slow down. However, they increase after that. Although this could be analytically interpreted as evidence for the linking and delinking hypothesis, results should be interpreted with caution. Less than 5% of the sample observations exceed the threshold value.[17].

[17]Some authors have even found out-of sample turning points.For instance [14, 57]

4.2 Non-parametric estimation

The non-parametric approach was carried out by estimating three versions of equation (3). The first one was a Kernel regression using the improved Alkaike Information Criterion and bootstrapped standard errors. That is $E_{it} = f(Y_{it}) + \varepsilon_{it}$. Unfortunately, this approach is cannot separate the function in more than one component, or include linear effects in the specification. The second version estimated follows the general specification proposed by Vollebergh et al. [71]. The equation is $E_{it} = f(Y_{it}) + \lambda(\tau) + \varepsilon_{it}$, Finally, we estimate (3) without any further alteration. The graphical results are shown in Figure 9. Although the non-parametric approach generates an inverted U-shape EKC similar to the one estimated in the quadratic linear specification, we can observe that the income-emission trajectory predicts higher emission levels after GDP reaches 10,000.

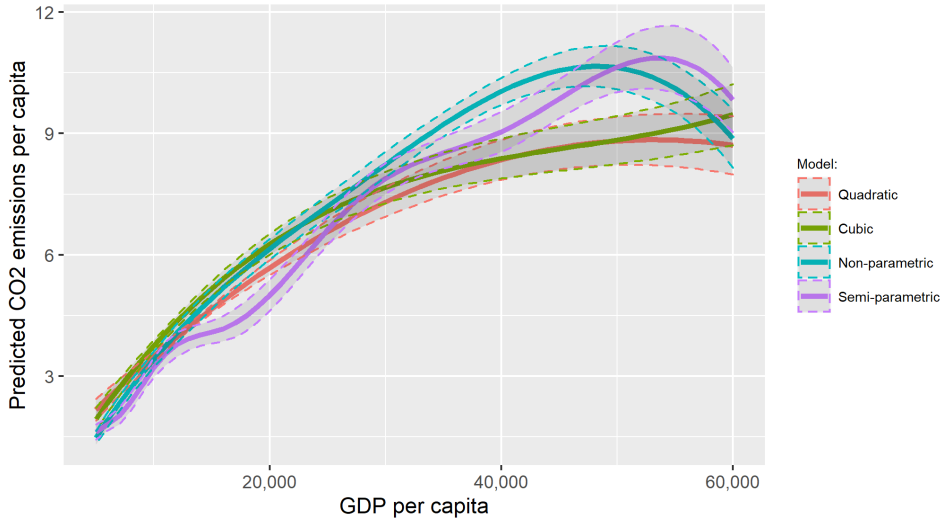


Figure 9: Predicted emissions for the linear and non-parametric specifications. 95% confidence interval in shadows.

The turning point of the non-parametric approaches is approximately near the one in the quadratic estimation at 50,000 GDP., coinciding with the quadratic estimation. The standard errors of the model also increase around this value. Again, we attribute this result to the few observations that surpass this level of income and the heterogeneity in the emissions per unit of GDP. For instance, during 2014, Switzerland's GDP per capita was 66,038.73 and emitted 4.31 metric tons per capita of CO_2. In the same year, with a lower GDP per capita (57,313.85), the US emissions were almost four times larger.

Table 3: Polynomial and semi-parametric estimation results.

Variables	Linear		Semi-parametric
	Quadratic	Cubic	
GDP	0.000306***	0.000496***	
	(1.99e-05)	(3.61e-05)	
GDP^2	-2.88e-09***	-9.81e-09***	
	(2.43e-10)	(1.08e-09)	
GDP^3		7.09e-14***	
		(1.06e-14)	
Share of S1	0.0121	0.0207***	0.00916
	(0.00805)	(0.00747)	(0.0134)
Share of S2	0.0597***	0.0559***	0.0860***
	(0.00919)	(0.00851)	(0.0123)
Share of S3	-0.00418	-0.00398	-0.0327**
	(0.00835)	(0.00800)	(0.0130)
Exports	-0.0141***	-0.0137***	-0.0162***
	(0.00341)	(0.00331)	(0.00330)
Urban population	0.0935***	0.0824***	0.0206***
	(0.00825)	(0.00750)	(0.00364)
Population growth	0.137***	0.121***	-0.226***
	(0.0428)	(0.0389)	(0.0491)
Renewable electricity	-0.0190***	-0.0179***	-0.0296***
	(0.00199)	(0.00194)	(0.00166)
Constant	-4.045***	-4.862***	
	(0.799)	(0.771)	
Country FE	YES	YES	NO
Year FE	YES	YES	Non-parametric
Observations	2,101	2,101	2,101
R-squared	0.980	0.981	

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

The specifications that omit control variables yield similar results.[18], regardless if time effects are considered or not. By including control variables, the shape of the curve changes its slope for different income levels.

The significant coefficients for control variables vary only within the standard errors when the linear specification is augmented from a second to a third-degree polynomial. When the model is non-parametric in income and time, all the control coefficients vary significantly except for the share of services and exports.

4.3 Heterogeneity analysis

If imposing linear restrictions in the EKC functional form leads to biased results as argued by Stern [64] or identification problems, a potential solution would be the estimation of non-parametric models [5, 71] as we did on the previous subsections. However, pairs of data on income and emissions for the wealthiest countries exhibit differentiated trends. In order to identify them, we attempt to explain the sources of heterogeneity considering the control variables included in our model as plausible candidates. First, we conduct the Hsiao test for heterogeneity [33].[19]

Table 4: Estimated F-statistics for the Hsiao test.

Hypothesis	Estimated F-statistic	p-value
H_1: Identical slopes, heterogeneous intercepts	9.7547	0.0000
H_3: Identical coefficients	152.0062	0.0000
H_4: Identical intercepts conditional on identical slopes	274.1231	0.0000

Table 4 presents the estimated F-statistic and the associated p-value for each of the hypotheses of the test. Following the procedure proposed by Hsiao, the hypothesis H_3 of full parameter homogeneity is rejected. The hypothesis H_1 of identical slopes is also rejected, confirming the full heterogeneity of the model. Usually, the tests halt when H_1 is rejectd. H_4 is presented for illustrative purposes. Since the test analyses the covariance across different assumptions on the parameter's behavior is important to note that the smallest RSS is found for the individual country regression estimates. Imposing a homogeneity restriction on the parameter values and

[18]Thus, only report the specification including non-parametric time effects.

[19]The test procedure and our results can be consulted in Appendix A.2.

turning points for the EKC in our model will result in coefficients unable to predict individual country behavior. Results are consistent with those found by de Bruyn [20] and confirm heterogeneity in the income-emission relationship found by other authors [35, 53, 55, 70].

Using the regression tree approach to classify the data and estimating under the regression tree approach, we obtain the results summarized in Figure 10. Although it was not reported, the previous linear and semi-parametric estimates for the time variable were found to be statistically not significant. With a few exceptions, most of the countries were classified within the same category over the years. For this reason, the year variable did not generate meaningful splits to categorize the data. Subsequently, we will generalize the results as depicting characteristics of representative economies.

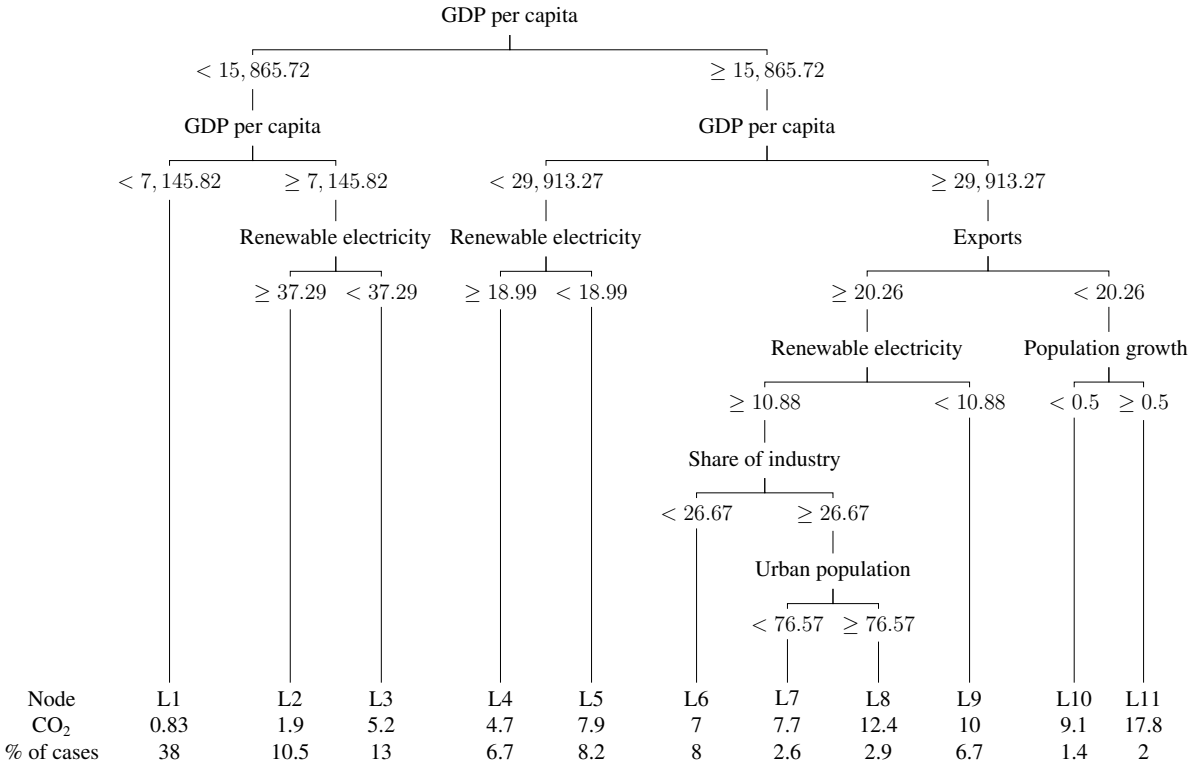


Figure 10: Regression tree model results

The essential variable to generate splits in our model is GDP per capita, producing at least four country categories. Despite this values do not correspond to the World Bank Classification Income Classification, we will adopt the same naming.[20] Emissions for Low income countries

[20]The current World Bank thresholds [66] of income are Low income up to 1,026 GDP; Lower-middle income up to 3,995; Upper-middle income up to 12,375; and High income all countries above. Our lowest GDP edge classified Low income countries amidst middle income countries.

(GDP<7,145.82) are the lowest across the sample. No further variable can explain differences for this group. Lower-middle income countries with a GDP per capita between 7,145.8 and 15,865.7 emit less CO_2 gases if the renewable electricity generation accounts for more than 37% of the total output. Renewable electricity generation is also relevant for Upper-middle income countries. However, their average electricity generation is 14.78% lower than Lower-middle income countries. Thus, overall emissions are larger.

Figure 10 portrays the increasing complexity of the income-emission relationship as the income levels increase, which can also be appreciated in Figure 11, where the resulting regions of three regression tree are mapped on top of the scatter plot. For the High income countries, the first relevant distinction is the value of their exports represented as a share of the GDP. For countries with the lower level of exports, emissions are linked to population growth. We must remark the scarcity of countries classified on the nodes L10 and L11 with low levels of exports. Only Australia and the US belong to the node L11 (high income and high population growth), while node L10 only classifies Japan during all periods, and three years for Italy and Greece.

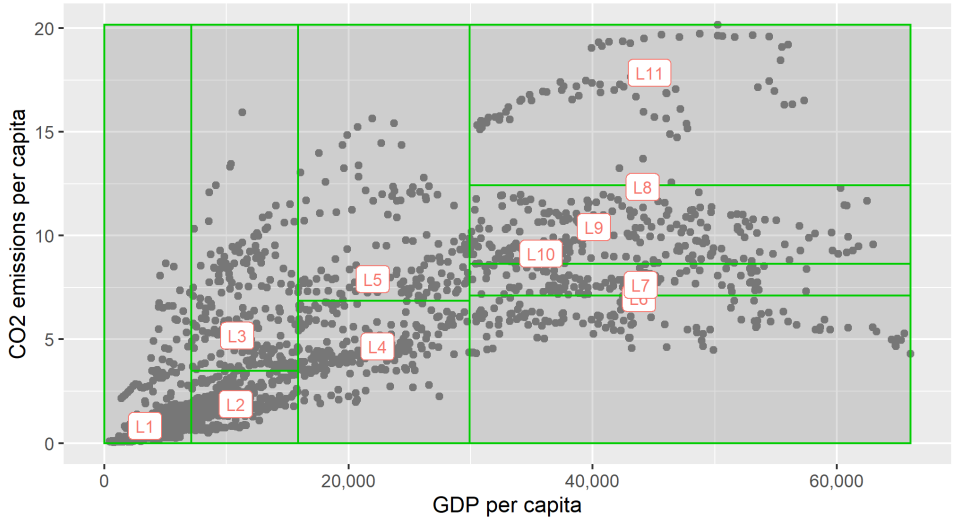


Figure 11: Income-emission classification according to the regression tree results.

For countries with high exports (exceeding 20.26%), the lowest CO_2 emission levels are found when either the industrial activity or the share of urban population is lower than the rest. Secondly, emissions are higher when renewable electricity accounts for less than 10.88% of the electricity output. Finally, when both industrial activity and urban population is high, emissions are the second largest among all the nodes, despite renewable electricity generation outpacing the threshold.

Table 5: Marginal effects of the variables according to the mean values classified by the regression tree.

Node	Mean values						Emissions			
	GDP per capita	Renewable electricity	Exports	Share of industry	Urban population	Population growth	Average	Semi-parametric		
								Predicted	95% interval	
L1	3350.47						0.83	1.01	0.78	1.25
L2	10775.74	73.57					1.88	2.42	2.17	2.68
L3	10891.92	10.09					5.17	4.34	4.01	4.67
L4	22375.23	43.13					4.65	5.58	5.22	5.94
L5	21981.1	7.82					7.91	6.49	6.1	6.89
L6	43796.27	36.16	40.07	22.78			6.99	9.26	8.79	9.72
L7	43939.91	58.83	43.86	28.8	66		7.63	9.25	8.76	9.74
L8	44097.94	54.07	37.07	30.19	79.6		12.34	9.93	9.41	10.44
L9	40102.27	4.13	48.14				10.42	9.82	9.35	10.3
L10	35777.03		13.53			0.14	9.13	9.18	8.79	9.57
L11	44654.46		14.16			1.13	17.85	10.15	9.6	10.7

Note: Predicted emissions are estimated only with the values shown in each node.

As a final step in our strategy, we computed the average values for each node's relevant variables and plugging the values in the mean function of income and the linear coefficients associated with the control variables of the semi-parametric estimation of the previous subsection. Table 5 presents the input values and the prediction results. Out of the 11 nodes, the semi-parametric function only achieved to estimate the average CO_2 emissions within the 95% confidence interval at nodes L1 and L10. Unfortunately, there is no proper metric to contrast a regression tree with a semi-parametric estimation. Since the regression tree method computes the actual average of CO_2 emissions for each terminal node, we might expect that the results of an adequate linear prediction resemble the mean values of each node. This was not the case.

5 Discussion and conclusions

Our main conclusion is confirming that the heterogeneity of the EKC shape for individual countries found by previous literature [20, 35, 53, 55, 70]. As observed in Figures 11 and 10, the income-emission relationship cannot be explained only by differences in income levels. Even countries within similar income groups can emit different amounts of CO_2 conditionally on the value of other variables relevant to economic growth. In other words, the *scale effect* is

not sufficient to portray a single path of emission development.

The generation of renewable electricity can alter the income-emission relationship between countries. A surprising fact found in our data is that renewable electricity generation is proportionally higher in Low and Lower-middle income countries than the rest. Hence, it is directly related to lower CO_2 emissions. Following Smulders et al. [62], enhanced and environmentally friendly technology should be used in more developed countries. Since renewable electricity can be considered as one, we find the opposite.

Energy is a relevant variable to promote economic development. For instance, Wolfram et al. [73] perform a country-level empirical examination. Controlling for simultaneity and endogeneity issues, they expose that energy demand, poverty reduction, and pollution are associated. Our results for High income countries seem to contradict those of Glaeser and Kahn [30] who identify that cities tend to have lower CO_2 emissions than suburban areas. In contrast, we find that higher concentrations of urban populations yield higher per capita emissions. This finding coincides with the recent papers of Borck and Schrauth [6], and Carozzi and Sefi Roth [12]. A potential explanation is pointed out by Moreno-Cruz and Taylor [52], who develop a model in which cheap and available energy is required to foster urban growth. Further research is required on this subject.

Since the exports variable was only meaningful for high income countries, we do not detect evidence in favor of the *Pollution Heaven* or the *Race to the Bottom* hypotheses. It would have been the case if different levels of exports generated splits for the lower income countries. Cherniwchan et al. [15] reaches a similar conclusion.

The *composition effect* might be relevant for high income countries with larger shares of urban population. Our regression tree results only generated splits for the industry sector, probably because of the hump-shaped pattern shown in Figure 6. We cannot say more about the interactions of different sectors.

Potentially, incorporating technological development in the model could modify these results. However, we could not use a suitable variable to incorporate in our models. Hopefully, in the future, more appropriate measurements will be available.

All the variables' implications described above illustrate not only the CO_2 emissions gen-

erated by an economy but also the underlying characteristics that shape growth. We consider that theoretical and empirical research on the environmental linking and de-linking process of development should be done, as different income-emission relationships might arise. For instance, if we imagine a country in transit from node L1 to L6, going through L2 and L4, the development process would be accompanied by initial investments on renewable energy and a less industrialized economy. On the other hand, we would characterize the opposite from L1 to L11, through L3 and L5. Other meaningful variables can operate in-between, but they are yet to be recognized.

We contribute to the existing literature by introducing the estimation of regression trees as a data analysis methodology seldom used in economics to examine differentiated implications of variables. With this approach, further research can improve the application of structural econometric analysis to face econometric identification challenges [37]. Second, our analysis suggest that biased estimations can arise in both linear and semi-parametric estimations. Our results for linear specifications approximated a turning point similar to those of the semi-parametric approach. However, the predicted CO_2 emissions were lower, suggesting probable bias. Likewise, the analytical examination of the data, after applying the regression tree method showed that the semi-parametric estimations could also yield inaccurate estimates. Subsequent research must focus on determining whether dynamic interaction of variables such as industrial activity, population growth, urban population, or others have linear or dynamic implications in emissions.

Some of these results could be of interest to policy makers. Estimations of the urban population growth in the world place developing countries as faster-growing than developed countries. We have provided evidence that sustainable energy generation is related to lower CO_2 emissions, while the concentration of urban population can be associated. Thus, further efforts should be made to jointly reduce urban and energy interrelated emissions.

References

[1] Philippe Aghion et al. “Carbon Taxes, Path Dependency, and Directed Technical Change: Evidence from the Auto Industry.” In: *Journal of Political Economy* 124.1 (Feb. 2016), pp. 1–51. ISSN: 00223808.

[2] Halil Altıntaş and Yacouba Kassouri. “Is the Environmental Kuznets Curve in Europe Related to the Per-Capita Ecological Footprint or CO2 Emissions?” In: *Ecological Indicators* 113 (June 2020), p. 106187. ISSN: 1470160X. DOI: 10.1016/j.ecolind.2020.106187.

[3] Bo Pieter Johannes Andrée et al. “Revisiting the Relation between Economic Growth and the Environment; a Global Assessment of Deforestation, Pollution and Carbon Emission”. In: *Renewable and Sustainable Energy Reviews* 114 (2019), p. 109221. ISSN: 1364-0321. DOI: 10.1016/j.rser.2019.06.028.

[4] James Andreoni and Arik Levinson. *The Simple Analytics of the Environmental Kuznets Curve*. Working Paper 6739. National Bureau of Economic Research, Sept. 1998. DOI: 10.3386/w6739.

[5] Jean-Thomas Bernard et al. “Environmental Kuznets Curve: Tipping Points, Uncertainty and Weak Identification”. In: *Environ Resource Econ* 60.2 (2015), pp. 285–315. ISSN: 0924-6460, 1573-1502. DOI: 10.1007/s10640-014-9767-y.

[6] Rainald Borck and Philipp Schrauth. *Population Density and Urban Air Quality*. Tech. rep. 08. Center for Economic Policy Analysis, May 2019.

[7] Alain Bousquet and Pascal Favard. “Does S. Kuznets’s Belief Question the Environmental Kuznets Curves?” In: *The Canadian Journal of Economics / Revue canadienne d’Economique* 38.2 (2005), pp. 604–614. ISSN: 00084085, 15405982.

[8] Leo Breiman et al. *Classification and Regression Trees*. OCLC: 1007134524. 2017. ISBN: 978-1-315-13947-0.

[9] William A Brock and M Scott Taylor. “The Green Solow Model”. In: *Journal of Economic Growth* 15.2 (2010), pp. 127–153. ISSN: 13814338. DOI: 10.1007/s10887-010-9051-0.

[10] Gro Harlem. Brundtland and World Commission on Environment and Development. *Our Common Future*. Oxford: Oxford University Press, 1987. Chap. XV, 400 p. ; 21 cm. ISBN: 0-19-282080-X 978-0-19-282080-8.

[11] Annegrete Bruvoll and Hege Medin. “Factors Behind the Environmental Kuznets Curve. A Decomposition of the Changes in Air Pollution”. In: *Environmental and Resource Economics* 24.1 (Jan. 2003), pp. 27–48. ISSN: 1573-1502. DOI: 10.1023/A:1022881928158.

[12] Felipe Carozzi and Sefi Roth. *Dirty Density: Air Quality and the Density of American Cities*. Tech. rep. dp1635. Centre for Economic Performance, LSE, July 2019.

[13] R. T. Carson. "The Environmental Kuznets Curve: Seeking Empirical Regularity and Theoretical Structure". In: *Review of Environmental Economics and Policy* 4.1 (Dec. 2010), pp. 3–23. ISSN: 1750-6816, 1750-6824. DOI: 10.1093/reep/rep021.

[14] Therese A. Cavlovic et al. "A Meta-Analysis of Environmental Kuznets Curve Studies". In: *Agricultural and Resource Economics Review* 29.1 (2000), pp. 32–42. ISSN: 1068-2805. DOI: 10.1017/S1068280500001416.

[15] Jevan Cherniwchan, Brian R. Copeland, and M. Scott Taylor. "Trade and the Environment: New Methods, Measurements, and Results". In: *Annu. Rev. Econ.* 9.1 (Aug. 2017), pp. 59–85. ISSN: 1941-1383, 1941-1391. DOI: 10.1146/annurev-economics-063016-103756.

[16] Ariaster B. Chimeli. "Growth and the Environment: Are We Looking at the Right Data?" In: *Economics Letters* 96.1 (July 2007), pp. 89–96. ISSN: 01651765. DOI: 10.1016/j.econlet.2006.12.016.

[17] Brian R. Copeland and M. Scott Taylor. "Trade, Growth, and the Environment". In: *Journal of Economic Literature* 42.1 (2004), pp. 7–71. DOI: 10.1257/002205104773558047.

[18] Graeme S. Cumming and Stephan von Cramon-Taubadel. "Linking Economic Growth Pathways and Environmental Sustainability by Understanding Development as Alternate Social–Ecological Regimes". In: *Proc Natl Acad Sci USA* 115.38 (Sept. 2018), p. 9533. DOI: 10.1073/pnas.1807026115.

[19] Susmita Dasgupta et al. "Confronting the Environmental Kuznets Curve". In: *Journal of Economic Perspectives* 16.1 (2002), pp. 147–168. DOI: 10.1257/0895330027157.

[20] Sander M. de Bruyn. *Economic Growth and the Environment*. Vol. 18. Economy & Environment. Dordrecht: Springer Netherlands, 2000. ISBN: 978-94-010-5789-9 978-94-011-4068-3. DOI: 10.1007/978-94-011-4068-3.

[21] Henri L. F. de Groot, Cees A. Withagen, and Zhou Minliang. "Dynamics of China's Regional Development and Pollution: An Investigation into the Environmental Kuznets Curve". In: *Environment and Development Economics* 9.4 (2004), pp. 507–537. ISSN: 1355-770X. DOI: 10.1017/S1355770X0300113X.

[22] Henri L.F. de Groot. *Structural Change, Economic Growth and the Environmental Kuznets Curve : A Theoretical Perspective*. OCFEB Research Memorandum 9911, 'Environmental Policy, Economic Reform and Endogenous Technology' 1. Rotterdam: Research Centre for Economic Policy, 1999. ISBN: 90-5539-085-2 978-90-5539-085-4.

[23] Mehmet Akif Destek et al. “The Relationship between Economic Growth and Carbon Emissions in G-7 Countries: Evidence from Time-Varying Parameters with a Long History”. In: *Environ Sci Pollut Res* (May 2020). ISSN: 1614-7499. DOI: 10.1007/s11356-020-09189-y.

[24] Giuseppe Di Vita. “Is the Discount Rate Relevant in Explaining the Environmental Kuznets Curve?” In: *Journal of Policy Modeling* 30.2 (Mar. 2008), pp. 191–207. ISSN: 01618938. DOI: 10.1016/j.jpolmod.2007.04.012.

[25] Soumyananda Dinda. “A Theoretical Basis for the Environmental Kuznets Curve”. In: *Ecological Economics* 53.3 (May 2005), pp. 403–413. ISSN: 09218009. DOI: 10.1016/j.ecolecon.2004.10.007.

[26] P Ekins. “The Kuznets Curve for the Environment and Economic Growth: Examining the Evidence”. In: *Environ Plan A* 29.5 (May 1997), pp. 805–830. ISSN: 0308-518X, 1472-3409. DOI: 10.1068/a290805.

[27] Clas Eriksson and Joakim Persson. “Economic Growth, Inequality, Democratization, and the Environment”. In: *Environmental and Resource Economics* 25.1 (May 2003), pp. 1–16. ISSN: 1573-1502. DOI: 10.1023/A:1023658725021.

[28] Sahbi Farhani et al. “The Environmental Kuznets Curve and Sustainability: A Panel Data Analysis”. In: *Energy Policy* 71 (Aug. 2014), pp. 189–198. ISSN: 0301-4215. DOI: 10.1016/j.enpol.2014.04.030.

[29] Eugenio Figueroa and Roberto Pastén. “Beyond Additive Preferences: Economic Behavior and the Income Pollution Path”. In: *Resource and Energy Economics* 41 (Aug. 2015), pp. 91–102. ISSN: 09287655. DOI: 10.1016/j.reseneeco.2015.04.004.

[30] Edward L. Glaeser and Matthew E. Kahn. “The Greenness of Cities: Carbon Dioxide Emissions and Urban Development”. In: *Journal of Urban Economics* 67.3 (2010), pp. 404–418. ISSN: 0094-1190. DOI: 10.1016/j.jue.2009.11.006.

[31] Gene Grossman. *Pollution and Growth: What Do We Know?* Tech. rep. 848. C.E.P.R. Discussion Papers, Oct. 1993.

[32] Gene Grossman and Alan Krueger. *Environmental Impacts of a North American Free Trade Agreement*. Tech. rep. w3914. Cambridge, MA: National Bureau of Economic Research, Nov. 1991, w3914. DOI: 10.3386/w3914.

[33] Cheng Hsiao. *Analysis of Panel Data*. Third edition. Econometric Society Monographs. New York, NY: Cambridge University Press, 2014. ISBN: 978-1-107-03869-1 978-1-107-65763-2.

[34] "Tree-Based Methods". In: *An Introduction to Statistical Learning: With Applications in R*. Ed. by Gareth James et al. Springer Texts in Statistics 103. OCLC: ocn828488009. New York: Springer, 2013, pp. 303–335. ISBN: 978-1-4614-7137-0.

[35] Thomas Jobert, Fatih Karanfil, and Anna Tykhonenko. "Estimating Country-Specific Environmental Kuznets Curves from Panel Data: A Bayesian Shrinkage Approach." In: *Applied Economics* 46.13 (May 2014), pp. 1449–1464. ISSN: 00036846.

[36] Yoichi. Kaya and Keiichi. Yokobori. *Environment, Energy, and Economy : Strategies for Sustainability*. Ed. by Energy Tokyo Conference on "Global Environment and Economic Development (25-10-1993 - 27-10-1993 : Tokyo) and Energy Global Environment and Economic Development (1993 : Tokio). Tokyo: United Nations University Press, 1997. Chap. x, 381 p. : ill. ; 24 cm. ISBN: 92-808-0911-3 978-92-808-0911-4.

[37] Michael P. Keane. "Structural vs. Atheoretic Approaches to Econometrics". In: *Journal of Econometrics* 156.1 (May 2010), pp. 3–20. ISSN: 0304-4076. DOI: 10.1016/j.jeconom.2009.09.003.

[38] Masaaki Kijima, Katsumasa Nishide, and Atsuyuki Ohyama. "Economic Models for the Environmental Kuznets Curve: A Survey". In: *Journal of Economic Dynamics and Control* 34.7 (July 2010), pp. 1187–1201. ISSN: 01651889. DOI: 10.1016/j.jedc.2010.03.010.

[39] Simon Kuznets. "Economic Growth and Income Inequality". In: *The American Economic Review* 45.1 (1955), pp. 1–28. ISSN: 00028282.

[40] Chien-Chiang Lee, Yi-Bin Chiu, and Chia-Hung Sun. "Does One Size Fit All? A Reexamination of the Environmental Kuznets Curve Using the Dynamic Panel Data Approach". In: *Review of Agricultural Economics* 31.4 (Dec. 2009), pp. 751–778. ISSN: 10587195, 14679353. DOI: 10.1111/j.1467-9353.2009.01465.x.

[41] Christoph Martin Lieb. *The Environmental Kuznets Curve: A Survey of the Empirical Evidence and of Possible Causes*. Tech. rep. 391. University of Heidelberg, Department of Economics, 2003.

[42] John A. List and Craig A. Gallet. "The Environmental Kuznets Curve: Does One Size Fit All?" In: *Ecological Economics* 31.3 (Dec. 1999), pp. 409–423. ISSN: 0921-8009. DOI: 10.1016/S0921-8009(99)00064-6.

[43] Ramón E. López. "The Environment as a Factor of Production: The Effects of Economic Growth and Trade Liberalization". In: *Journal of Environmental Economics and Management* 27.2 (Sept. 1994), pp. 163–184. ISSN: 0095-0696. DOI: 10.1006/jeem.1994.1032.

[44] Ramón E. López and Sang Yoon. "Pollution–Income Dynamics". In: *Economics Letters* 124.3 (Sept. 2014), pp. 504–507. ISSN: 01651765. DOI: 10.1016/j.econlet.2014.07.024.

[45] Tommaso Luzzati, Marco Orsini, and Gianluca Gucciardi. "A Multiscale Reassessment of the Environmental Kuznets Curve for Energy and CO2 Emissions". In: *Energy Policy* 122 (Nov. 2018), pp. 612–621. ISSN: 03014215. DOI: 10.1016/j.enpol.2018.07.019.

[46] Shiguo Ma and Lei Shi. "The Micro-Foundations of the Environmental Kuznets Curve". In: *Fudan J. Hum. Soc. Sci.* 7.3 (Sept. 2014), pp. 471–482. ISSN: 1674-0750, 2198-2600. DOI: 10.1007/s40647-014-0036-9.

[47] N. Gregory Mankiw, David Romer, and David N Weil. *A Contribution to the Empirics of Economic Growth*. Working Paper 3541. National Bureau of Economic Research, Dec. 1990. DOI: 10.3386/w3541.

[48] Giovanni Marin and Massimiliano Mazzanti. "The dynamics of delinking in industrial emissions: The role of productivity, trade and R&D". In: *Journal of Innovation Economics & Management* 3.1 (2009), pp. 91–117. DOI: 10.3917/jie.003.0091.

[49] Kenneth E. McConnell. "Income and the Demand for Environmental Quality". In: *Envir. Dev. Econ.* 2.4 (July 1997), pp. 383–399. ISSN: 1355-770X, 1469-4395. DOI: 10.1017/S1355770X9700020X.

[50] Donella H. Meadows and Club of Rome. *The Limits to Growth; a Report for the Club of Rome's Project on the Predicament of Mankind*. Potomac Associates Books. New York: Universe Books, 1972. Chap. 205 pages illustrations 21 cm. ISBN: 0-87663-165-0 978-0-87663-165-2 0-85644-008-6 978-0-85644-008-3 0-87663-918-X 978-0-87663-918-4.

[51] Petar Mitić, Milena Kresoja, and Jelena Minović. "A Literature Survey of the Environmental Kuznets Curve". In: *Economic Analysis* 52.1 (2019), pp. 109–127. ISSN: 1821-2573, 2560-3949. DOI: 10.28934/ea.19.52.12.pp109\%0010127.

[52] Juan Moreno-Cruz and M. Scott Taylor. "An Energy-Centric Theory of Agglomeration". In: *Journal of Environmental Economics and Management* 84 (2017), pp. 153–172. ISSN: 0095-0696. DOI: 10.1016/j.jeem.2017.02.006.

[53] Antonio Musolesi, Massimiliano Mazzanti, and Roberto Zoboli. "A Panel Data Heterogeneous Bayesian Estimation of Environmental Kuznets Curves for CO2 Emissions." In: *Applied Economics* 42.18 (July 2010), pp. 2275–2287. ISSN: 00036846.

[54] Theodore Panayotou. *Empirical Tests and Policy Analysis of Environmental Degradation at Different Stages of Economic Development*. Tech. rep. 992927783402676. International Labour Organization, 1993.

[55] Matías Piaggio and Emilio Padilla. "CO2 Emissions and Economic Activity: Heterogeneity across Countries and Non-Stationary Series". In: *Energy Policy* 46 (July 2012), pp. 370–381. ISSN: 0301-4215. DOI: 10.1016/j.enpol.2012.03.074.

[56] David Popp. "Induced Innovation and Energy Prices". In: *American Economic Review* 92.1 (2002), pp. 160–180. DOI: 10.1257/000282802760015658.

[57] Alexandra-Anca Purcel. "New Insights into the Environmental Kuznets Curve Hypothesis in Developing and Transition Economies: A Literature Survey". In: *Environ Econ Policy Stud* (Mar. 2020). ISSN: 1432-847X, 1867-383X. DOI: 10.1007/s10018-020-00272-9.

[58] Amran Md. Rasli et al. "New Toxics, Race to the Bottom and Revised Environmental Kuznets Curve: The Case of Local and Global Pollutants". In: *Renewable and Sustainable Energy Reviews* 81 (Jan. 2018), pp. 3120–3130. ISSN: 1364-0321. DOI: 10.1016/j.rser.2017.08.092.

[59] Haroon Rasool, Mushtaq Ahmad Malik, and Md. Tarique. "The Curvilinear Relationship between Environmental Pollution and Economic Growth: Evidence from India". In: *IJESM* ahead-of-print.ahead-of-print (Mar. 2020). ISSN: 1750-6220, 1750-6220. DOI: 10.1108/IJESM-04-2019-0017.

[60] Sulhi Ridzuan. "Inequality and the Environmental Kuznets Curve". In: *Journal of Cleaner Production* 228 (Aug. 2019), pp. 1472–1481. ISSN: 09596526. DOI: 10.1016/j.jclepro.2019.04.284.

[61] Burak Sencer Atasoy. "Testing the Environmental Kuznets Curve Hypothesis across the U.S.: Evidence from Panel Mean Group Estimators". In: *Renewable and Sustainable Energy Reviews* 77 (Sept. 2017), pp. 731–747. ISSN: 13640321. DOI: 10.1016/j.rser.2017.04.050.

[62] Sjak Smulders, Lucas Bretschger, and Hannes Egli. "Economic Growth and the Diffusion of Clean Technologies: Explaining Environmental Kuznets Curves". In: *Environ Resource Econ* 49.1 (May 2011), pp. 79–99. ISSN: 0924-6460, 1573-1502. DOI: 10.1007/s10640-010-9425-y.

[63] Sigrid Stagl. *Delinking Economic Growth from Environmental Degradation? A Literature Survey on the Environmental Kuznets Curve Hypothesis*. SSRN Scholarly Paper ID 223869. Rochester, NY: Social Science Research Network, Aug. 1999. DOI: 10.2139/ssrn.223869.

[64] David I Stern. "The Rise and Fall of the Environmental Kuznets Curve". In: *World Development* 32.8 (Aug. 2004), pp. 1419–1439. ISSN: 0305750X. DOI: 10.1016/j.worlddev.2004.03.004.

[65] Nancy L. Stokey. “Are There Limits to Growth?” In: *International Economic Review* 39.1 (Feb. 1998), p. 1. ISSN: 00206598.

[66] World Bank Data Team. *New Country Classifications by Income Level: 2019-2020*. https://blogs.worldbank.org/op country-classifications-income-level-2019-2020. Jan. 2019.

[67] Terry M Therneau, Elizabeth J Atkinson, and Mayo Foundation. “An Introduction to Recursive Partitioning Using the RPART Routines”. In: (), p. 60.

[68] Tetsuya Tsurumi and Shunsuke Managi. “Decomposition of the Environmental Kuznets Curve: Scale, Technique, and Composition Effects”. In: *Environ. Econ. Policy Stud.* 11.1-4 (Feb. 2010), pp. 19–36. ISSN: 1432-847X, 1867-383X. DOI: 10.1007/s10018-009-0159-4.

[69] OAR US EPA. *Global Greenhouse Gas Emissions Data*. https://www.epa.gov/ghgemissions/global-greenhouse-gas-emissions-data. Overviews and Factsheets. Jan. 2016.

[70] Herman R.J. Vollebergh, Elbert Dijkgraaf, and Bertrand Melenberg. “Environmental Kuznets Curves for CO2: Heterogeneity versus Homogeneity”. In: *SSRN Journal* (2005), p. 39. ISSN: 1556-5068. DOI: 10.2139/ssrn.683109.

[71] Herman R.J. Vollebergh, Bertrand Melenberg, and Elbert Dijkgraaf. “Identifying Reduced-Form Relations with Panel Data: The Case of Pollution and Income”. In: *Journal of Environmental Economics and Management* 58.1 (2009), pp. 27–42. ISSN: 0095-0696. DOI: 10.1016/j.jeem.2008.12.005.

[72] Martin Wagner. “The Carbon Kuznets Curve: A Cloudy Picture Emitted by Bad Econometrics?” In: *Resource and Energy Economics* 30.3 (Aug. 2008), pp. 388–408. ISSN: 09287655. DOI: 10.1016/j.reseneeco.2007.11.001.

[73] Catherine Wolfram, Orie Shelef, and Paul Gertler. “How Will Energy Demand Develop in the Developing World?” In: *Journal of Economic Perspectives* 26.1 (2012), pp. 119–38. DOI: 10.1257/jep.26.1.119.

[74] Zakaria Zoundi. “CO2 Emissions, Renewable Energy and the Environmental Kuznets Curve, a Panel Cointegration Approach”. In: *Renewable and Sustainable Energy Reviews* 72 (May 2017), pp. 1067–1075. ISSN: 13640321. DOI: 10.1016/j.rser.2016.10.018.

A Appendix

A.1 Selection criteria

Table 6: Removal reasons.

Code	Reason
A	Small states
B	Mainly oil producers (share of oil in GDP >80%)
C	Industry not reported
D	Share of industry <10%
E	Share of services >60%
F	GDP >70,000
G	Observations <20 or emissions not reported
H	Armed conflicts interrupting the data set from 1990 - 2014
I	Presence of severe outliers
J	Low statistical score (<60) according to the Statistical Capacity Indicator in 2010

Note: Criteria A, B, and J are similar to the sample definition by Mankiw, Romer and Weil [47].

The included countries by income level are:

Low income: Burkina Faso, Madagascar, Malawi, Mali, Mozambique, Niger, Rwanda, Tanzania, Uganda.

Lower middle income: Bangladesh, Bolivia, Cambodia, Cameroon, Egypt, El Salvador, Ghana, Honduras, India, Kenya, Kyrgyz Republic, Lao PDR, Mauritania, Moldova, Mongolia, Morocco, Nicaragua, Nigeria, Pakistan, Philippines, Senegal, Tunisia, Ukraine, Vietnam.

Upper middle income: Albania, Argentina, Belarus, Bosnia and Herzegovina, Brazil, China, Colombia, Dominican Republic, Ecuador, Georgia, Guatemala, Jordan, Kazakhstan, Malaysia, Mexico, North Macedonia, Paraguay, Peru, Romania, Russian Federation, South Africa, Sri Lanka, Thailand, Turkey.

High income: Australia, Austria, Belgium, Canada, Chile, Croatia, Czech Republic, Den-

mark, Finland, France, Germany, Greece, Hungary, Ireland, Israel, Italy, Japan, Korea, Rep., Latvia, Lithuania, Netherlands, New Zealand, Norway, Panama, Poland, Portugal, Slovak Republic, Slovenia, Spain, Sweden, Switzerland, United Kingdom, United States, Uruguay.

A.2 Hsiao test

The test is built upon an analysis of covariance with restrictions imposed on the individual regression for cross-sectional units:

$$y_{it} = \alpha_i + \beta_i' X_{it} + u_{it} \tag{4}$$

Where $i = 1, \ldots, N$, $t = 1, \ldots, T$, β_i' is a vector of $1 \times K$ constant parameters, X_{it} is the $1 \times K$ vector of exogenous variables, and u_{it} is the error term. (4) can be slightly modified to test heterogeneity across time. Since this thesis focuses on heterogeneity across countries, we will not test for this.

The restrictions are formulated considering a system of four hypotheses.

H_1 : Slope coefficients are identical, intercepts are not.

$\alpha_1 \neq \alpha_2 \neq \ldots \neq \alpha_N$ and $\beta_1 = \beta_2 = \ldots = \beta_N$

H_2 : Intercepts are identical, slope coefficients are not.

$\alpha_1 = \alpha_2 = \ldots = \alpha_N$ and $\beta_1 \neq \beta_2 \neq \ldots \neq \beta_N$

H_3 : Both slope and intercept coefficients are identical.

$\alpha_1 = \alpha_2 = \ldots = \alpha_N$ and $\beta_1 = \beta_2 = \ldots = \beta_N$

H_4 : Provides a more thorough examination if H_1 is accepted: Both slope and intercept coefficients are identical.

$\alpha_1 = \alpha_2 = \ldots = \alpha_N$, given $\beta_1 = \beta_2 = \ldots = \beta_N$.

The test requires computing the Residual Sum of Squares of the unrestricted equation (4) and of regressions consistent with fixed effects and a pooled models. The unrestricted sum of

squares $S_1 = \sum_{i=1}^{N} RSS_i$ is equivalent to compute individual country regressions and adding all the resulting sum of squares. $S_2 = RSS_{FE}$ and $S_3 = RSS_{Pooled}$. Then, the procedure of the test is carried out by the process of hypotheses elimination by computing F-statistics.

First, H_3 must be tested. The associated F-statistic is:

$$F_3 = \frac{(S_3 - S_1)/[(N-1)(K+1)]}{S_1/[NT - N(K+1)]}$$

If the test is not significant, H_3 cannot be rejected and the model can be estimated by a pooled panel method. In the contrary case, the researcher must investigate whether the source of heterogeneity relies on the intercept or slopes.

The F-statistic to test H_1 is:

$$F_1 = \frac{(S_2 - S_1)[(N-1)K]}{S_1[NT - N(K+1)]}$$

In this step of the test, we have rejected the assumption of full homogeneity in the model. Although H_2 is formulated, the interpretation of a common intercept with heterogeneous slopes lacks any useful meaning. The logical conclusion, if F_1 is significant, will be to keep the assumption of heterogeneous slope and intercept parameters. However, in the cases when H_1 is accepted, Hsiao suggest examining if intercepts are heterogeneous conditional on homogeneous slopes. H_4 should be tested.

$$F_4 = \frac{(S_3 - S_2)/(N-1)}{S_2/[N(T-1) - K]}$$

If H_4 is accepted, then the pooled model is sufficient. If rejected, the estimation model should include individual effects.

Table 7: Covariance of Hsiaos' tests for homogeneity.

Residual sum of squares		Degrees of freedom	
Source of variation	Value	Formula	Value
S_1: Within group with heterogeneous intercept and slope	460.6923	$N(T-K-1)$	1,857
S_2: Constant slope and heterogeneous intercept	1,114.089	$N(T-1)-K$	2,127
S_3: Common intercept and slope	1,4036.41	$NT-(K+1)$	2,217

Note: Adapted from Hsiao [33, p. 22]. Own estimation results.

A.3 Results of testing the linear specification

Table 8: VIF test for different polynomials under different transformations of GDP.

Polynomial	Levels			Differences			Logarithmic			Logarithmic differences		
GDP	12.314	58.487	195.662	5.229	17.171	43.022	211.955	24,064.643	2,038,979	56.011	2,765.962	116,911.32
GDP^2	12.314	295.283	2,961.136	5.229	102.197	731.081	211.955	102,045.008	20,008,402	56.011	11,255.443	1,111,999.625
GDP^3		111.012	5,615.075		48.704	1,727.396		27,349.932	22,143,938		2,991.033	1,226,692.125
GDP^4			1,192.139			443.008			2,756,837.25			156,000.078
Mean VIF	12.314	154.927	2,491.003	5.229	56.024	736.127	211.955	51,153.194	11,737,039.063	56.011	5,670.813	652,900.787

Table 9: Results of the Wooldridge test for autocorrelation in panel data.

Polynomial	Levels		Logarithmic	
	F-statistic	p-value	F-statistic	p-value
GDP	28.591	0.0000	132.218	0.0000
GDP^2	24.262	0.0000	134.871	0.0000
GDP^3	23.663	0.0000	132.841	0.0000

Table 10: Results of the the Westerlunnd test for cointegration.

	Levels		Differences		Logarithmic		Logarithmic differences	
	Trend	No trend	Trend	No trend	Trend	No trend	Trend	No trend
Variance ratio	-6.1302	-1.5071	-9.6888	-11.9406	-5.9161	0.0460	-9.5185	-12.0879
p-value	0.0000	0.0659	-	-	0.0000	0.4817	-	-

Notes: Linear: $GDP = x$; Logarithmic: $GDP = \ln x$; Differences: $GDP = x_t - x_{t-1}$;

Logarithmic differences: $GDP = \ln x_t - \ln x_{t-1}$

A.4 Regression tree method

This section is based on Chapter 8 of James et al. [34] and Chapter 8 of Breiman et al. [8].

A regression tree is a method based developed for machine learning that provides tools for classifying and generating regressions by splitting the predictor space into simpler regions based on a least-square condition. The notion is to divide the predictors $X_1, X_2, \ldots, X_p$ into J distinct and non-overlaping regions $R_1, R_2, \ldots, R_J$. For each region, the mean of the objective variable y is computed as the predicted value $\hat{y}_{Rj}$. Then, the residual sum of squares is computed:

$$RSS = \sum_{j=i}^{J} \sum_{i \in R_j} (y_i - \hat{y}_{Rj})^2 \tag{5}$$

Due to the computational limitations of every possible interaction, the absolute minimum value of equation (5) is unfeasible to estimate. For this reason, the algorithm follows a recursive binary splitting approach. Any predictor X_p is selected and a cutpoint s that divides the sample in two spaces is defined:

$$R_1(j, s) = \{X | X_j < s\} \text{ and } R_2(j, s) = \{X | X_j \geq s\}$$

By recursive itineration, the algorithm stops when the values of j and s minimize

$$\sum_{i: x_i \in R_1(j,s)} (y_i - \hat{y}_{R_1})^2 + \sum_{i: x_i \in R_2(j,s)} (y_i - \hat{y}_{R_2})^2 \tag{6}$$

That is, for each p variable, the algorithm sets different values of s until (6) is minimized. Then, the minimum residual values for each X_p is compared and the first split of the data is generated for the predictor with the lowest result of (6). The process is repeated for each region until a stop criterion is reached.

A tree diagram similar to the one in Figure 10 illustrates the results of the model. The splits (or branches) are based on the criteria exposed in the previous paragraph. The reader shall move to the left if the criteria is meet and to the right if not. Each terminal node (or leave) indicates when further splitting of the data does not lead to minimize (6) and the $\hat{y}_{R_j}$ average is computed for of all the observations that meet the criteria established by the branches above.

The regression tree technique is prone to overfitting the data. In an extreme case, the ideal number of regions will be equal to the number of observations and the residual sum of squares would be 0. To avoid this problem, adjusting the algorithm parameters to halt the process is necessary. This process is called tree pruning. It requires that the researcher stet threshold values, after which the algorithm fully or partially stops. The usual approach to prune the model is by starting with the largest possible tree and adjusting the error complexity measure

$$\sum_{m=1}^{|T|} \sum_{i:x_i \in R_j} (y_i - \hat{y}_{R_2})^2 + \alpha |T| \tag{7}$$

Where $|T|$ represents the number of terminal nodes of the tree and α is a tuning parameter chosen from a validation set or by cross-validation. α is a function of $|T|$. Thus, it allows choosing the number of branches that yield no additional gains in further splits. Additionally, the researcher can set the minimum number of observations to either be included in each terminal node or to generate a split. Subsequently, we will follow both approaches. Another important measure to evaluate the model performance is the variable importance that summarizes the goodness of fit of each split on which it was the primary variable defining a split plus the goodness of fit of all splits on which it was a surrogate. The rpart package in R scales the values to sum 100 [67].

Like many machine learning tools, the regression tree requires to split the sample into two subsamples containing about 80% and 20% of the data. The sample containing 80% of the variables is used to train the tree model and adjust the algorithm's parameters. The model structure is evaluated using the remaining 20% of the sample to compare the prediction errors in both models. Unfortunately, reducing the sample size to estimate the goodness of fit of the model leads to RSS uncomparable to those obtained in linear or non-parametric estimations.

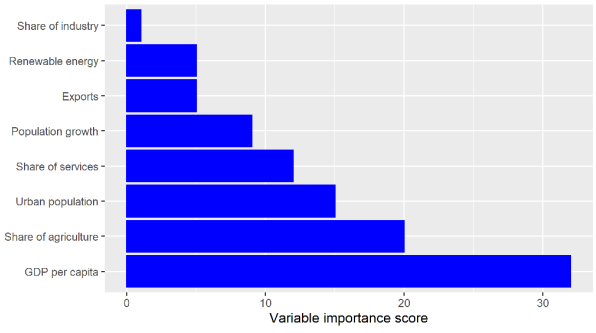


Figure 12: Variable importance values.

In the first model we identified that the variable for years was the least relevant for the model. It was omitted in the final estimation as it failed to generate any branches in the tree. Although the Share of industry was scored lower than services and agriculture, the latter did not generate any splits in our data. The higher relevance of the variables is that if omitted, the classification would be different and generate splits that increase the RSS.

The minimum predicted error generated by the number of branches was found at 11 splits. However, after 10 splits, the increase of the tree complexity adds little information to the predicted values. A second model was generated by limiting the complexity parameter to match the 10 splits, removing the least essential variables and setting the minimum number of observations in a branch to 20.

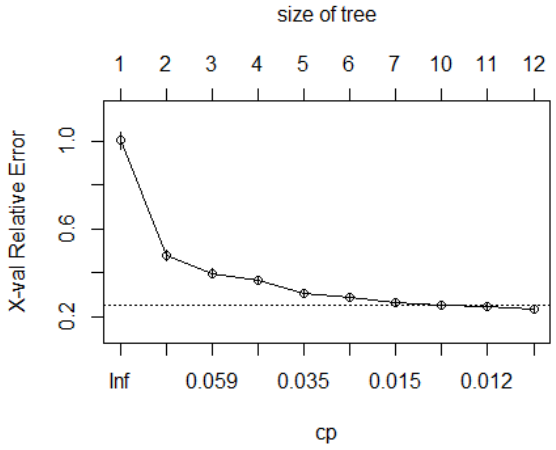


Figure 13: Complexity parameter, tree size and relative error.

As the second model proved to not substantially increase the RSS in the test subsample (Going from 2.01 in the original model to 2.14 in the adjusted), we applied the tuning parameters to the full dataset. The resulting tree model is the one described in the main text. For reproducibility the code is presented in the following Annex.

A.5 Regression tree code in R

The model requires the packages: tree, MASS, tidiverse, rpart, rpart.plot, Metrics and gbm. The full code and dataset are available upon request.

```
Tree_data <- WB_data3 %>%
  dplyr::select(year, co2_pc, gdp_pc, exports_perc, pop_growth,
```

```
                s1_perc, s2_perc, s3_perc, exports_perc,
                urban_perc, renew_out_perc)

set.seed(73)
assignment <- sample(1:2, size = nrow(Tree_data),
                     prob = c(80,20), replace = TRUE)

Tree_train <- Tree_data[assignment == 1,]
Tree_test <- Tree_data[assignment == 2,]

Tree_model <- rpart(formula = co2_pc ~. ,
                    data = Tree_train,
                    method = "anova")

print(Tree_model$cptable)
opt_index <- which.min(Tree_model$cptable[, "xerror"])
cp_opt <- Tree_model$cptable[opt_index,"CP"]

summary(Tree_model)
print(Tree_model)
rpart.plot(x = Tree_model, yesno = 2, type = 0, extra = 1)
plotcp(Tree_model)
cp_table_model <- data.frame(Tree_model$cptable)
row_best_model <- which(cp_table_model$nsplit == 10)
best_cp_model <- cp_table_model$CP[row_best_model]

Tree_model_adjust <- rpart(formula = co2_pc ~ gdp_pc +
                    urban_perc + s1_perc + s2_perc + s3_perc +
                    renew_out_perc + pop_growth + exports_perc,
                    data = Tree_train,
                    method = "anova",
                    control = rpart.control(minbucket = 20,
```

```
                                cp = best_cp_model))

summary(Tree_model_adjust)
print(Tree_model_adjust)
rpart.plot(x = Tree_model_adjust, yesno = 2, type = 0, extra = 1)
plotcp(Tree_model_adjust)

pred_model <- predict(object = Tree_model, newdata = Tree_test)
pred_adjusted <- predict(object = Tree_model_adjust,
                         newdata = Tree_test)

RMSE_model <- rmse(actual = Tree_test$co2_pc,
                   predicted = pred_model)
RMSE_adjusted <- rmse(actual = Tree_test$co2_pc,
                      predicted = pred_adjusted)

Tree_FINAL <- rpart(formula = co2_pc ~ gdp_pc + urban_perc +
                    s1_perc + s2_perc + s3_perc +
                    renew_out_perc + pop_growth + exports_perc,
                    data = Tree_data,
                    method = "anova",
                    control = rpart.control(minbucket = 20,
                              cp = best_cp_model))
summary(Tree_FINAL)
print(Tree_FINAL)
rpart.plot(x = Tree_FINAL, yesno = 2, type = 0, extra = 1)
```