Drivers of carbon dioxide emissions: an empirical investigation using hierarchical and non-hierarchical clustering methods

John Inekwe1 · Elizabeth Ann Maharaj2 · Mita Bhattacharya3

Received: 28 June 2019 / Revised: 26 November 2019 / Published online: 18 December 2019
© Springer Science+Business Media, LLC, part of Springer Nature 2019

Abstract
The mitigation of CO₂ emissions requires a global effort with common but differentiated responsibilities. In this paper, we identify clusters of CO₂ emissions across 72 countries. First, using the stochastic version of the IPAT and employing the dynamic common correlated effects technique, we identify three key determinants affecting CO₂ emissions (non-renewables, population, and real GDP). In the second step, both hierarchical and non-hierarchical clustering methods are considered to identify the optimal number of clusters. We identify two to four clusters with different member countries, and in particular establish that in most cases, a 2-cluster solution appears to be optimal. The contents of clusters vary slightly according to the clustering methods for each period. The clustering results from using only the overall CO₂ emissions indicate that the countries we consider form three clusters, with China and the USA each within a single member cluster. The remaining 70 countries form the third cluster. Our findings reflect the prominent roles of China and the USA in overall CO₂ emissions. Analyses with sub-period and largest emitters reflect a different clustering structure. Some policy recommendations in setting emission reductions are made, considering different clusters across countries.

Keywords Cluster analysis · CO₂ emissions · Key determinants · Stochastic version of IPAT

Handling Editor: Pierre Dutilleul.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10651-019-00433-4) contains supplementary material, which is available to authorized users.

* John Inekwe
john.inekwe@mq.edu.au

1 Department of Applied Finance, Macquarie University, Sydney, Australia
2 Department of Econometrics & Business Statistics, Monash University, Melbourne, Australia
3 Department of Economics, Monash University, Melbourne, Australia
1 Introduction

Global warming and the associated environmental degradation are under constant scrutiny. The Intergovernmental Panel on Climate Change (IPCC) has released guidelines for the national greenhouse (GHG) inventory since 1995 and published several updated versions in the last three decades. The United Nations Framework Convention on Climate Change (UNFCCC) follows the IPCC guidelines in setting the national emissions inventory. Implementation depends on ‘common but differentiated responsibility’, based on the economic development, structural composition and emission growth factors of countries. At the same time, countries emitting high CO\textsubscript{2} emissions per capita may be grouped together and coordinate to set a similar reduction target.

Scientific research addresses the problem of climate change as a result of anthropogenic emissions. Since the first IPCC assessment report (Change 1990), there have been active efforts to design and adopt policies for controlling the emissions of pollutants. The effects of CO\textsubscript{2} emissions are evident globally; mitigation requires managing the global commons and maintaining a measure of international coordination among nations. Reducing emissions and maintaining economic activities at the same time need a balance between economic development and environmental policy. Analysing the drivers of greenhouse gas (GHG) emissions provides useful insights for carrying out climate negotiation agreements and mitigation actions.\textsuperscript{1}

Climate change is recognized as a problem on a global scale; however, not all countries have similar structures of production, consumption, and CO\textsubscript{2} emissions per capita. In order to identify similar countries (in a group), key factors that relate to the production of CO\textsubscript{2} emissions must be identified for the countries in groups.\textsuperscript{2}

There is a rich volume of literature identifying key factors causing an increasing level of CO\textsubscript{2} emissions. Recent research can be found in Apergis (2016); Sadorsky (2014) and Wang and Lin (2017).\textsuperscript{3} In a recent survey on applied energy economics research relating energy consumption-CO\textsubscript{2} emissions to economic growth, Smyth and Narayan (2015) emphasised the over-use of time series or panel estimation techniques in identifying key non-energy variables relating to energy variables. Decoupling global emissions and economic growth has been discussed widely in the empirical literature.\textsuperscript{4}

In the world energy arena, there is increasing interest in mitigating CO\textsubscript{2} emissions, either convergence patterns are established across countries and regions, while there is a significant difference across countries or regions, or alternatively clustering, as a result of the disparities in their CO\textsubscript{2} emissions. The recent studies on convergence with CO\textsubscript{2} emissions are discussed in Aldy (2006); Criado and Grether

\textsuperscript{1} CO\textsubscript{2} emissions is considered here as the primary source of GHG.

\textsuperscript{2} We consider CO\textsubscript{2} emissions, as this emission is the primary mitigating factor for climate change in the global context. Other studies include Awaworyi Churchill et al. (2018); Cai et al. (2018); Churchill et al. (2018); Mikayilov et al. (2018); Mrabet et al. (2017); and Wu et al. (2018); Zhou et al. (2018).

\textsuperscript{3} Energy use and the associated pollution has been an important research issue (see for example, Anandalingam and Bhattacharya 1985; Li et al. 2012).

\textsuperscript{4} See Schandl et al. (2016) for further discussion.

\(\text{Springer}\)
(2011); Strazicich and List (2003), and others. The advantage of using clustering analysis can be identified by the fact that cluster analysis aims at sorting different countries into groups in a way that the degree of association between any two countries is maximal if they belong to the same group and minimal otherwise.

In this study, we provide a rational basis for clustering groups from 72 countries following their emissions characteristics over time. While there are studies in the literature that propose clustering of locations based on air-pollution data (D’Urso et al. 2013, 2014, 2015, 2017; Lafuente-Rego et al. 2018; Vilar et al. 2018), there is a scant literature in energy economics in using clustering methods. In particular, Phillips and Sul (2007, 2009) use a club convergence method, where they develop a new algorithm to identify clusters of convergent subgroups. Their algorithm is a data-driven method that avoids ex-ante sample separation. It can be used as a general panel method to cluster objects into groups with similar transition paths. Using this approach, Bhattacharya et al. (2020); Bhattacharya et al. (2018) identify clubs based on energy efficiency across states in Australia and India respectively.

To our knowledge, there is no study based on a clustering approach applied to countries where consideration of the effects of key factors contributing to CO₂ emissions is considered. We fill the research gap in this respect. In our analysis, we also exclude China and the USA who are the largest polluters in the world, in order to identify different clusters for which the influence of these two countries is minimised. Therefore, we believe our analysis is based on a robust econometric approach and some policy suggestions are indicated in the concluding section.

We use an extended version of IPAT identity developed by Ehrlich and Holdren (1971) to establish the major drivers of emissions for the clustering analysis. We follow three steps. Firstly, we analyse the findings with a full version of the CO₂ model, considering CO₂ emissions as features. Secondly, we follow these clusters within the CO₂ trajectory considering significant variables. Finally, we attempt to provide some policy suggestions for reducing emissions where countries can act within a cluster in a coordinated manner. It implies that, as generally accepted by proponents of cluster-based policies, various countries should not try to create clusters starting from scratch.

Section 2 covers the most significant aspects of the theoretical model used and the empirical methodologies we implement to identify key factors; we describe the measures of variables and data sources in this section. In Sect. 3, we define the various aspects of the cluster analysis, while in Sect. 4, we analyse the findings from different clusters and the dynamics of these factors over time with changing CO₂ emissions. The concluding section covers our major findings with possible policy suggestions.

5 Impact (I) = Population (P) × Affluence (A) × Technology (T).
6 The drivers of Chinese CO₂ emissions are explored in Guan et al. (2008). Supply-chain clusters with high CO₂ emissions are presented in Kagawa et al. (2015).
7 Features are the input variables used for clustering.
2 Model, data and estimated parameters

The IPAT identity was developed by Ehrlich and Holdren (1971) to consider the influence of human activity on the environment (Ehrlich and Holdren 1971). A simpler version of the identity is defined as Impact (I) = Population (P) × Affluence (A) × Technology (T). Following IPAT, environmental impact (measured in CO₂ emissions) is influenced by population growth, proxies of affluence, and technology. Since its development, the IPAT Model has been widely used in the literature to analyse the influence of economic and social factors on the environment. We use here the extended version of the model. The conventional IPAT equation is an accounting identity where the coefficients of all factors have fixed units of elasticity, and it is therefore restricted for empirical testing purposes. Dietz and Rosa (1994) relaxed the setting of fixed units of elasticity and modified it with a stochastic version of the IPAT (STIRPAT) model. A convenient calculation could be achieved through logarithmic dimensionless processing of variables and, accordingly, the basic analysis framework of the STIRPAT model can be expressed as follows:

\[
CO_2_{it} = \alpha_i + \delta_1 CO_2_{it-1} + \delta_2 NRE_{it} + \delta_3 RNE_{it} + \delta_4 POP_{it} + \delta_5 RGDP_{it} \\
+ \delta_6 URB_{it} + \delta_7 SER_{it} + \delta_8 IND_{it} + u_{it}
\]  

(1)

\[
u_{it} = \gamma f_{it} + e_{it}
\]

(2)

where CO₂ measures CO₂ emissions per capita in metric tons; \(\alpha_i\) is the constant term; NRE is the non-renewable energy intensity; RNE is the renewable energy intensity; POP is population; RGDP is the real gross domestic product per capita (measured in constant 2010 US dollar); URB is the urban population as a share of total population; SER is the share of service sector value added (as a percentage of GDP); and IND is the share of industrial sector value added (constant US dollar). Parameters \(\delta_s\) are vectors of coefficient, where \(s = 1, ..., 8; i\) is the index of countries; \(t\) is the time index; \(u_{it}\) is the error term, \(f_{it}\) is \(m \times 1\) vector of unobserved common effects with country-specific factor loadings \(\gamma_i\); and \(e_{it}\) are the idiosyncratic errors. We consider the logarithmic version of this model for estimation purposes following Apergis and Payne (2010), and Bhattacharya et al. (2016). These variables are extracted from the World Development Indicators. The period of analysis covers 1970–2015.

2.1 Estimations and the selection of key determinants of CO₂ emissions

Over the last two decades, the popularity of panel time-series data (large N and T) has been enhanced by the availability of international macroeconomic databases and

---

8 Non-renewable energy consumption (NRE) is the sum of coal, gas and oil. Renewable energy consumption (RNE) is the sum of hydro, wind, solar, geothermal, marine, waste, and solid, liquid and gaseous biofuel-derived energy. Renewable (combustibles) and non-renewable (fossil fuel) energy intensity measures are in kilograms of oil equivalent per GDP (in US dollar).
the development of estimation, which extends first-generation panel time series techniques to allow for heterogeneity in the slope coefficients. Notable among these estimation techniques are the mean group OLS (MG), developed by Pesaran and Smith (1995), and the pooled mean group (PMG), developed by Pesaran et al. (1999). The MG estimator assumes that slope coefficients, error variances, and intercepts differ across countries. In contrast, the PMG estimator allows intercepts, error variances, and short-run coefficients to vary across countries while the long-run coefficients are constrained to be the same. The criticism of these panel time-series estimations is that they tend to generate inconsistent estimates in the presence of common shocks or common factors, i.e. in the presence of cross-sectional dependence across the panel.

The correlation of the residuals across panel units may arise from unobserved common factors in the panel, or a correlation between the predictors and the residuals may be present. Biased estimates arising from the unobserved common factors in the panel unit have garnered attention among researchers. In addressing this issue of unobservable common factors, recent research has focused on either slope homogeneity (see, Bai 2009; Moon and Weidner 2015), which yields inconsistent estimates, or slope heterogeneity and cross-sectional dependence. Notable among the latter developments was the introduction of the common correlated effects (CCE) estimator developed by Pesaran (2006). This estimator uses the cross-section averages of the variables to approximate the projection space of the unobserved factors while considering the heterogeneity of the slopes. The extension of the CCE technique to include weakly exogenous (endogenous independent) variables with the lags of the dependent variable is developed by Chudik and Pesaran (2015). They establish the dynamic common corrected effects (DCCE) estimator, which increases the overall efficiency of estimations.

In this paper, we employ the dynamic common correlated effects technique across a panel of countries. The estimation involves a full panel of 72 countries with a sufficient number of observations to conduct estimations. For the period of our study, these countries emit 82% of total global CO₂ emissions (based on all country-level data available in the World Bank dataset). The panel also has a good mixture of developed and developing countries. The advantage of this estimation is that it allows the coefficients to be obtained for the entire sample and for each of the countries within the panel. Refer to Tables B1 and B2 in the Online Supplementary Material for the summary statistics of all the explanatory variables and the summary statistics of CO₂ emissions of the 72 countries, respectively.

Using the dynamic common correlated effects estimation, we account for common shocks across the countries. Two models are examined in this preliminary analysis. The first version of the model involves all the variables as stated in Eq. 1; in the second model, we employ only the significant variables observed from the first model.

We present the results of the regression analysis for the full version of the model as well as for other versions in Table 1. The results establish non-renewable energy

---

9 The data is from the World Bank (2015).
intensity (NRE), population (POP), and the real gross domestic product (RGDP) as significant in driving CO₂ emissions. Non-renewable energy sources causing CO₂ emissions is established as a significant positive factor in empirical literature; for example, in a recent study with OECD countries, Shafiei and Salim (2014) find that non-renewable energy consumption has a positive and statistically significant effect on CO₂ emissions in the long-run. This effect has been a major cause worldwide, which demands that countries change their energy-mix and introduce clean technology. Population has been found to be a significant factor in studies; e.g. Cole and Neumayer (2004) and Liddle (2015), among others. The nexus between CO₂ emissions and economic income has been explored in studies testing the Environmental Kuznets Curve (EKC) Hypothesis (see Ozturk and Acaravci, 2013 and references therein).

In Table 1, for Models (1) to (1’), variables are in levels while for Models (2) to (5), we consider variables in first differences to further account for non-stationarity. In particular,

- Model (1) is the full version of the model in levels for the period 1970–2015 while.
- Model (1’), consists only of significant variables in levels, for this period (except for the dynamic term which is not significant in this case, but is included in all models).

### Table 1  Determinants of CO₂ emissions

| Variables | Model (1) | (1’) | (2) | (3) | (4) | (5) |
|-----------|-----------|------|-----|-----|-----|-----|
| NRE       | 1.168*    | 1.044*** | 1.027*** | 0.993*** | 0.820* | 0.804 |
|           | (0.629)   | (0.359) | (0.369) | (0.358) | (0.469) | (0.519) |
| RNE       | 0.004     |       |      |      |      |      |
|           | (1.599)   |       |      |      |      |      |
| POP       | 1.173***  | 1.180*** | 1.090*** | 1.121*** | 1.002*** | 1.044*** |
|           | (0.242)   | (0.253) | (0.234) | (0.234) | (0.227) | (0.229) |
| RGDP      | 0.328**   | 0.360*** | 0.353*** | 0.338*** | 0.390** | 0.367** |
|           | (0.128)   | (0.134) | (0.131) | (0.130) | (0.153) | (0.151) |
| URB       | −0.103    |      |      |      |      |      |
|           | (1.314)   |      |      |      |      |      |
| SER       | −0.065    |      |      |      |      |      |
|           | (0.271)   |      |      |      |      |      |
| IND       | 0.068     |      |      |      |      |      |
|           | (0.119)   |      |      |      |      |      |
| CO₂t−1    | −0.016    | 0.004 | −0.072*** | −0.070*** | −0.052*** | −0.055*** |
|           | (0.012)   | (0.008) | (0.012) | (0.012) | (0.015) | (0.016) |
| Constant  | −15.504   | −11.987** | 0.000 | 0.000 | 0.001 | 0.001 |
|           | (9.930)   | (5.644) | (0.001) | (0.001) | (0.002) | (0.002) |
| Observations | 3190   | 3240 | 3168 | 3080 | 1800 | 1750 |

Standard errors in parentheses, *, ** and *** where p-value < 0.10, p-value < 0.05, p-value < 0.01, respectively.
• Model (2), consists of significant variables only in their first difference for the period 1970–2015.
• Model (3) consists of significant variables only in their first difference, excluding China and the USA, the largest emitters, for the period 1970–2015.
• Model (4), for which we consider the period of analysis, 1990–2015, consists of significant variables only in their first difference.
• Model (5) consists of the significant variables only in their first difference excluding China and the USA, for the period 1990–2015.

The sub-period (viz., 1990–2015) is chosen because the implementation of renewable energy sources was initiated in most of the countries in the early 1990s. In determining clusters in Sect. 4, we consider only the significant variables (NRE, POP RGDP, and CO$_2$t−1) in identifying the statistical strength of forming clusters. The coefficients at the country level for the first differences are used for the cluster analysis, i.e., coefficients from Models (2)–(5).

In the USA, in recent years, CO$_2$ emissions are on the rise, while in contrast, China has been very proactive in meeting the target in 2015 Paris agreement. In our analysis, we exclude these largest polluters for robustness checks, namely, China and the USA are excluded from Models (3) and (5) so that when comparing the clustering results with those of Models (2) and (4), respectively, we can determine if China and the USA influence the cluster membership of all other countries.

3 Cluster analysis

Cluster analysis is an exploratory tool to identify groupings of similar entities. While different methods can produce difference results, we select the most likely results based on optimality measures. In this paper, we use several well-known clustering methods and assess the compatibility of results between the methods. Our research relates this approach to identify CO$_2$ clusters using economic analysis.

Cluster analysis is used extensively in various fields of study (such as Life Sciences, Behavioural and Social Sciences, Earth Sciences and Engineering Sciences). Caiado et al. (2015); Dai and Kuosmanen (2014); Everitt et al. (2011); Anderberg (2014); Hennig et al. (2015) and Manly and Alberto (2016) are useful references for understanding the nature of and concepts about the cluster analysis. We provide a brief introduction to cluster analysis for non-specialist readers, with particular focus on the methods we use in our analyses in Sect. 4.

Cluster analysis is a multivariate method which aims to group a set of objects or cases on the basis of a set of variables into a number of mutually exclusive groups based on distances between objects. In other words, it partitions the set of objects and cases. Amongst different methods, we focus here on two popular ones: hierarchical and non-hierarchical methods.

The advantage of using clustering analysis in this context can be identified by the fact that cluster analysis aims at sorting different countries into groups, in a way that the degree of association between any two countries is maximal if they belong to the same group and minimal otherwise. While there are other multivariate statistical
methods such as multidimensional scaling that can also identify groupings, the outcomes of cluster analysis are much easier to convey graphically especially when there are more than two clusters. Other statistical methods such as discriminant analysis and the k-nearest neighbours approach require known groupings of existing cases before new cases can be classified.

3.1 Distance measures

The choice of distance measures can have an influence on the clustering results. There are many different distance measures but the most commonly known and used distance measure is the Euclidean distance. In general, if we have \( p \) variables \( X_1, X_2, \ldots, X_p \) measured on a sample of \( n \) objects of cases, the observed data for case \( i \) can be denoted by \( x_{i1}, x_{i2}, \ldots, x_{ip} \) and the observed data for subject \( j \) by \( x_{j1}, x_{j2}, \ldots, x_{jp} \). The Euclidean distance between these two objects is calculated as

\[
d_{ij}^E = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \ldots + (x_{ip} - x_{jp})^2}
\]

While it has been documented that the Euclidean distance has its short-comings (see for example, Shirkhorshidi et al. 2015), the way to overcome them is to ensure that the data is first normalised before determining the distances. Another distance measure is the Manhattan distance which is calculated as

\[
d_{ij}^M = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \ldots + |x_{ip} - x_{jp}|
\]

As well, in this case to overcome shortcomings, data must be first normalised before determining the distances.

3.2 Hierarchical methods

Hierarchical methods involve a series of \((n-1)\) decisions \((n\) equals the number of objects) that combine observations into a hierarchy or a tree-like structure (refer to Everitt et al. 2011). They can be either ‘agglomerative’, when groups are merged, or ‘divisive’, when one or more groups are split at each stage. More specifically, for agglomerative methods, each object starts out as its own cluster and is successively joined based on the distance between the objects. The two most similar clusters are then joined, and this continues until only a single cluster remains. For divisive methods, all objects start in a single cluster and are successively divided (first into two, then three, and so on) until each is a single-member cluster. We apply hierarchical agglomerative clustering, in particular, the minimum variance method of Ward Jr (1963) along with average and complete linkage methods; refer to Everitt et al. (2011) for more details about these methods.

---

10 In implementing our clustering procedures, we use the Euclidean and Manhattan distances on normalised data.
Once hierarchical clustering analysis is carried out, we select the optimum cluster solution in the next step. The literature emphasises both informal (more subjective), and formal methods (based on probability) in selecting the optimum cluster solution. An informal method is to examine a diagram called a dendrogram by which the hierarchical procedure is represented graphically. The dendrogram illustrates which clusters are joined at each stage of the analysis and the distance between clusters at the time of joining. If there is a large jump in the distance between clusters from one stage to another, this implies that at the first stage, clusters that are relatively close together are joined. In the next stage, the clusters that are joined are relatively far apart. This implies that the optimum number of clusters may be the number of clusters present immediately before the large jump in distance occurred. Formal methods include the determination of coefficients based on distances such as the Euclidean or Manhattan distances. One of these methods, the silhouette method will be used. In our discussion, we report the hierarchical clustering results that are based on Euclidean distances.

3.3 Non-hierarchical methods

In contrast to hierarchical methods, non-hierarchical methods do not involve a tree-like construction process. Instead, they assign objects into clusters once the number of cluster is specified. The first task is to identify starting points known as cluster seeds for each cluster. A cluster seed may be pre-specified, or objects may be randomly selected as seeds. We use two popular non-hierarchical methods, viz., the k-means and k-medoids.

The k-means method (Everitt et al. 2011) has been effective in producing good clustering results for many practical applications. The method operates through the following steps. Firstly, initial cluster centres (cluster seeds) are selected from among the objects, each of these objects then forms a cluster of one, and its centre is the value of the variables of that object.11 Secondly, each object is assigned to its ‘nearest’ cluster defined in terms of the distance to the centroid. The centroids which are the means of the clusters are then determined. Finally, the distance from each object to each centroid is recalculated, and objects that are not in the cluster are moved to the clusters they are closest to. This process continues until the centroid remains relatively stable. In practice, this is achieved by repeating the algorithm a large number of times using software programs such as MATLAB and R.

Whereas for the k-mean method, the mean of all the objects serves as the centre or prototype of a cluster, the k-medoids method selects an observed data point as centre of each cluster. This observed data point in each cluster is referred to as medoid which serves as the prototype of the cluster whose average dissimilarity to all the objects in the cluster is minimal, that is, it is the most centrally located point in the cluster. The k-medoids method is more robust to outlier data compared to

---

11 In practice in order to ensure that different seed choices do not produce different results and the k-means procedure converges, the k-means algorithm is run with multiple repetitions. In our implementation in MATLAB, we use 1000 repetitions.
the k-means method. The most common algorithm for implementing the k-medoids method is partitioning around medoids (PAM). Refer to Kaufman and Rousseeuw (2009) for details about the algorithm.

In our discussion, we will report k-means and k-medoids clustering results that are based on Euclidean distances.

### 3.4 Identifying the optimal number of clusters

There are several methods that can be used to identify the optimal number of clusters. We consider the average silhouette coefficient method. The silhouette method is used to find the strength of clusters and is based on distances. The number of clusters that gives the largest average silhouette coefficient is considered to be the optimal cluster solution; however, average silhouette coefficient values of less than 0.5 imply a weak separation of clusters. Kaufman and Rousseeuw (2009) give a detailed discussion on silhouette coefficients and plots.

### 3.5 Compatibility of cluster solutions

Since we obtain cluster solutions using different methods, we would like to assess the compatibility of these solutions. In what follows, we use a measure, viz., the Adjusted Rand Index. The Rand Index developed by Rand (1971) which is a measure of similarity, allows the unique evaluation of hard (so-called crisp or non-fuzzy) clustering partitions. The Rand Index is defined as the number of pairs of objects that are either in the same group or in different groups in both partitions divided by the total number of pairs of objects. The Rand Index lies between 0 and 1. When two partitions agree perfectly, the Rand Index achieves the maximum value 1. A problem with the Rand index is that the expected value of the Rand Index between two random partitions is not a constant. This problem is corrected by the Adjusted Rand Index which assumes the generalized hyper-geometric distribution as the specification of randomness. The Adjusted Rand Index has the maximum value 1, and its expected value is 0 in the case of random clusters. A larger Adjusted Rand Index means a higher agreement between two partitions. Hubert and Arabie (1985) have a detailed discussion on this index.

### 3.6 Determination of clusters

To determine how countries group together based on significant explanatory variables driving CO₂ emissions in the model, we apply cluster analysis to the standardised coefficients (betas) of all significant explanatory variables i.e., non-renewable energy, population, the real GDP, and lagged CO₂ emissions, from various models in Table 1. Comparisons across various model specifications are made to identify if the study period and the largest emitters have any influence in forming different clusters. Comparisons are made between:
• Models (2) and (3) to determine the impact the largest CO$_2$ emitters, China and the USA have on the grouping of the countries for the period 1970 to 2015.
• Models (2) and (4) to determine whether there is a substantial difference with the contents of clusters for the full period (1970–2015) and sub-period (1990–2015), considering the full panel of countries.
• Models (4) and (5) to determine the impact the largest CO$_2$ emitters have on the grouping of the countries for the sub-period (1990–2015).
• Models (3) and (5) to determine if excluding China and the USA, results in a substantial difference with the contents of clusters for the full and sub-period of analysis.

We consider different models to identify if the presence of the largest emitters and various periods may have any influence in determining clusters.

The sub-period (1990–2015) is chosen because the implementation of renewable energy sources was initiated in most of the countries in the early 1990s.

In each case we implement the following steps:

• Apply three hierarchical methods: Ward’s, average and complete linkage methods.
• Apply two non-hierarchical methods: k-means and k-medoids.
• Identify the possible optimal number of clusters for each method using the average silhouette coefficient.
• Assess the compatibility of the solutions across the four methods using the Adjusted Rand Index. Following that, we focus on the k-means method and one hierarchical method that is most compatible with the k-means method.
• Profile the cluster solutions based on summary statistics of the total CO$_2$ emissions per capita; henceforth we will refer to these summary statistics as features. Profiling will be undertaken by comparing the distribution of each of the features across the identified number of clusters. This is identified by examining boxplots.

However, to determine whether significant drivers of CO$_2$ emissions can lead to more meaningful clusters compared to clusters formed from total CO$_2$ emissions only, we cluster only total CO$_2$ emissions based on the summary statistics (features) of the series for each of Models (2) to (5).

4 Analysis of cluster results

4.1 Model 2: All countries, 1970–2015

In this sub-section, we consider the total CO$_2$ emissions as well as the significant explanatory variables for all countries for the full period of our analysis.
4.1.1 Cluster analysis with only CO₂ features

For this purpose, we consider maximum, minimum, mean, variance, skewness, and kurtosis for the CO₂ per capita for each country. From the dendrogram of Ward’s method in Figure A1, we can identify five possible cluster solutions. The average silhouette coefficients for each method are shown in Table 2 from where it is apparent that the 3-cluster solution is optimal for all methods.

On assessing the compatibility of the 2-cluster and 3-cluster solutions from the Adjusted Rand Indices (Table A5) across the three hierarchical methods and two non-hierarchical methods we observe 100% compatibility between all five methods.

For all models, dendrograms are shown and referred to in the Online Supplementary Material (Figures A1–A8).

For all models, tables with the Adjusted Rand Indices are shown in the Online Supplementary Material (Table A5–A12).

Table 2 Average Silhouette Coefficients for CO₂ Features: Model 2

| Number of clusters | k-means | k-medoids | Ward   | Average | Complete |
|--------------------|---------|-----------|--------|---------|----------|
| 2                  | 0.9470  | 0.9470    | 0.9470 | 0.9470  | 0.9470   |
| 3                  | 0.9716  | 0.9716    | 0.9716 | 0.9716  | 0.9716   |
| 4                  | 0.4621  | 0.4967    | 0.7153 | 0.7853  | 0.7853   |
| 5                  | 0.7041  | 0.5893    | 0.7566 | 0.7566  | 0.7543   |

Fig. 1 Boxplots of CO₂ features for 3-cluster solution: Model 2

12 For all models, dendrograms are shown and referred to in the Online Supplementary Material (Figures A1–A8).
13 For all models, tables with the Adjusted Rand Indices are shown in the Online Supplementary Material (Table A5–A12).
for the 2- and 3-cluster solutions. The results for the 3-cluster solution as shown in Table A1. The findings are identical across five methods we consider here. It is noticeable that Cluster 2 consists of only China, Cluster 3 consists only the USA, and the remaining 70 countries belong to Cluster 1.

### 4.1.1.1 Profiling of k-means cluster solution

Based on the above analysis, we profile the 3-cluster solutions on each of the CO2 features that were used to obtain the clusters. They are displayed in boxplots in Fig. 1. A boxplot summarises information about the distribution of a variable. It plots the summary statistics, median (Q2), represented by the horizontal line in box, the first quartile (Q1), third quartile (Q3) which is represented by the lower and upper edges of the box, the maximum and minimum values represented by horizontal lines on either side of the box, commonly referred to as the whiskers. The median indicates the level of the distribution, and the width of the box indicates the spread of the distribution. If the distribution has outliers, i.e., points away from the upper or lower edges of the box, then each such value will be represented by a cross (+).

From the boxplots, it is clear that the distribution of each feature is quite different for each cluster. Clusters 2 and 3 have single members, Cluster 2 is with China only, Cluster 3 is with the USA, while Cluster 1 consists of all other countries. China has the highest maximum emissions over the 46 years between 1970 and 2015 while the USA has the highest minimum and mean emissions over the 46 years. China has the largest variation and relative to the mean this variation is 80% compared to that of 17% for the USA. There are also distinct differences in the distribution for each of these countries following the skewness and kurtosis. Therefore, it is apparent that the overall CO2 features yield meaningful separation for our panel of countries into three clusters.

| Number of clusters | k-means | k-medoids | Ward | Average | Complete |
|--------------------|---------|-----------|------|---------|----------|
| 2                  | 0.580   | 0.606     | 0.556| 0.604   | 0.604    |
| 3                  | 0.454   | 0.420     | 0.467| 0.530   | 0.304    |
| 4                  | 0.458   | 0.428     | 0.421| 0.448   | 0.291    |

For all models, cluster solutions based on CO2 features are shown and referred to in the online Supplementary Material (Tables A1–A4).

For the 1970–2015 period the mean and standard deviation of China’s CO2 emissions are 2574324.35 metric tons and 2053792.29 metric tons, respectively.

For the 1970–2015 period the mean and standard deviation of USA’s CO2 emissions are 4682367.57 metric tons and 790581.60 metric tons, respectively.
4.1.2 Clustering with key significant variables contributing to CO2 emissions

We now determine how the cluster solution based only on the features of CO2 emissions compares to that of the cluster solution based on the coefficients of the key significant variables contributing to CO2 emissions for the period 1970–2015. From the dendrogram of Ward’s method in Figure A2, it seems possible to identify 2, 3, and 4 cluster solutions. The average silhouette coefficients for 2 to 4 clusters are given in Table 3. Since the 2-cluster solution produces the maximum average coefficient for all methods, it may be considered to be the optimal solution.

On assessing the compatibility of the 2-cluster solution from the Adjusted Rand Indices across the five methods (Table A6), we can observe that for the 2-cluster solution, there is 100% compatibility between the average and complete linkage methods, and of the hierarchical methods, Ward’s method is most compatible with the two non-hierarchical methods (68.7% and 68.4%). In addition, it is noticeable that there is very good compatibility between two non-hierarchical methods (78.3%).

The cluster solutions for the Ward and k-means methods are given in Table 4. We notice for the 2-cluster solution, the four countries, viz., Japan, Denmark, Sweden, and Sri Lanka switch clusters between these two methods. It does appear that the combined impact of non–renewables, population, real GDP, and lagged CO2 emissions on CO2 emissions tends to reasonably separate the countries into two clusters. Each cluster contains a mixture of developed and developing countries, but the Cluster 1 solution for each method contains a much higher percentage of developing countries than Cluster 2; for example, Ward’s Method (Cluster 1: 92% of developing countries, Cluster 2: 29% of developing countries), k-means Method (Cluster 1: 90% of developing countries, Cluster 2: 33% of developing countries). In both cases, the largest emitters like China and the USA are in Cluster 1.

While the contents of each cluster in this case are not necessarily consistent across all the methods considered, as indicated in the compatibility percentages, these results are clearly different from just clustering of the features of the CO2 emissions as indicated in Figure A1, indicating the usefulness of including the impact of the significant explanatory variables on CO2 emissions. It is interesting to note that in this case, China and the USA are also in different clusters.

4.1.2.1 Profiling of k‑means cluster solution Based on the above analysis, we profile this 2-cluster solution on each of the CO2 features. Boxplots of the 2-cluster solution are given in Fig. 2. We observe that except of a few outliers, the levels and spreads of distributions of the maximum, minimum, mean and variance of Cluster 1 are similar to that of Cluster 2. However, the level of the distribution of skewness of Cluster 1 is greater than that of Cluster 2 but the spread of this distribution for Cluster 2 is slightly larger than that of Cluster 1, whereas, the level of the distribution of kurtosis of Cluster 2 is greater than that of Cluster 1, but the spread of this distribution for Cluster 2 is slightly larger than that of Cluster 1. Similar observations were made when the 2-cluster solution from Ward’s method was profiled.
| Country                  | Cluster | Country       | Cluster | Country                  | Cluster |
|------------------------|---------|---------------|---------|------------------------|---------|
| 2 Algeria               | 1       | 1 Albania     | 2       | 2 Algeria               | 1       |
| 3 Argentina             | 1       | 4 Australia   | 2       | 3 Argentina             | 1       |
| 7 Benin                 | 1       | 5 Austria     | 2       | 7 Benin                 | 1       |
| 8 Bolivia               | 1       | 6 Bangladesh  | 2       | 8 Bolivia               | 1       |
| 9 Botswana              | 1       | 13 Chile      | 2       | 9 Botswana              | 1       |
| 10 Brazil               | 1       | 16 Congo, Dem. Rep | 2 | 10 Brazil               | 1       |
| 11 Bulgaria             | 1       | 21 Denmark    | 2       | 11 Bulgaria             | 1       |
| 12 Cameroon             | 1       | 26 Finland    | 2       | 12 Cameroon             | 1       |
| 14 China                | 1       | 27 France     | 2       | 14 China                | 1       |
| 15 Colombia             | 1       | 28 Greece     | 2       | 15 Colombia             | 1       |
| 17 Congo, Rep.          | 1       | 33 Italy      | 2       | 17 Congo, Rep.          | 1       |
| 18 Costa Rica           | 1       | 38 Luxembourg | 2       | 18 Costa Rica           | 1       |
| 19 Cuba                 | 1       | 39 Malaysia   | 2       | 19 Cuba                 | 1       |
| 20 Cyprus               | 1       | 44 Mozambique | 2       | 20 Cyprus               | 1       |
| 22 Dominican Republic   | 1       | 46 Netherlands| 2       | 22 Dominican Republic   | 1       |
| 23 Ecuador              | 1       | 48 Norway     | 2       | 23 Ecuador              | 1       |
| 24 Egypt, Arab Rep.     | 1       | 49 Pakistan   | 2       | 24 Egypt, Arab Rep.     | 1       |
| 25 Ethiopia             | 1       | 54 Portugal   | 2       | 25 Ethiopia             | 1       |
| 29 Honduras             | 1       | 58 Spain      | 2       | 29 Honduras             | 1       |
| 30 India                | 1       | 59 Sri Lanka  | 2       | 30 India                | 1       |
| 31 Indonesia            | 1       | 61 Sweden     | 2       | 31 Indonesia            | 1       |
| 32 Iran, Islamic Rep.   | 1       | 68 United States | 2   | 32 Iran, Islamic Rep.   | 1       |
| 34 Japan                | 1       | 69 Uruguay    | 2       | 34 Japan                | 1       |
| 35 Jordan               | 1       | 71 Zambia     | 2       | 35 Jordan               | 1       |
| 36 Kenya                | 1       |               |         | 36 Kenya                | 1       |
| 37 Korea, Rep.          | 1       |               |         | 37 Korea, Rep.          | 1       |

| Country                  | Cluster | Country       | Cluster | Country                  | Cluster |
|------------------------|---------|---------------|---------|------------------------|---------|
| 2 Algeria               | 1       | 1 Albania     | 2       | 2 Albania               | 1       |
| 3 Argentina             | 1       | 4 Australia   | 2       | 3 Argentina             | 1       |
| 7 Benin                 | 1       | 5 Austria     | 2       | 7 Benin                 | 1       |
| 8 Bolivia               | 1       | 6 Bangladesh  | 2       | 8 Bolivia               | 1       |
| 9 Botswana              | 1       | 13 Chile      | 2       | 9 Botswana              | 1       |
| 10 Brazil               | 1       | 16 Congo, Dem. Rep | 2 | 10 Brazil               | 1       |
| 11 Bulgaria             | 1       | 21 Denmark    | 2       | 11 Bulgaria             | 1       |
| 12 Cameroon             | 1       | 26 Finland    | 2       | 12 Cameroon             | 1       |
| 14 China                | 1       | 27 France     | 2       | 14 China                | 1       |
| 15 Colombia             | 1       | 28 Greece     | 2       | 15 Colombia             | 1       |
| 17 Congo, Rep.          | 1       | 33 Italy      | 2       | 17 Congo, Rep.          | 1       |
| 18 Costa Rica           | 1       | 38 Luxembourg | 2       | 18 Costa Rica           | 1       |
| 19 Cuba                 | 1       | 39 Malaysia   | 2       | 19 Cuba                 | 1       |
| 20 Cyprus               | 1       | 44 Mozambique | 2       | 20 Cyprus               | 1       |
| 22 Dominican Republic   | 1       | 46 Netherlands| 2       | 22 Dominican Republic   | 1       |
| 23 Ecuador              | 1       | 48 Norway     | 2       | 23 Ecuador              | 1       |
| 24 Egypt, Arab Rep.     | 1       | 49 Pakistan   | 2       | 24 Egypt, Arab Rep.     | 1       |
| 25 Ethiopia             | 1       | 54 Portugal   | 2       | 25 Ethiopia             | 1       |
| 29 Honduras             | 1       | 58 Spain      | 2       | 29 Honduras             | 1       |
| 30 India                | 1       | 59 Sri Lanka  | 2       | 30 India                | 1       |
| 31 Indonesia            | 1       | 61 Sweden     | 2       | 31 Indonesia            | 1       |
| 32 Iran, Islamic Rep.   | 1       | 68 United States | 2   | 32 Iran, Islamic Rep.   | 1       |
| 34 Japan                | 1       | 69 Uruguay    | 2       | 34 Japan                | 1       |
| 35 Jordan               | 1       | 71 Zambia     | 2       | 35 Jordan               | 1       |
| 36 Kenya                | 1       |               |         | 36 Kenya                | 1       |
| 37 Korea, Rep.          | 1       |               |         | 37 Korea, Rep.          | 1       |
Table 4 (continued)

| Country                  | Cluster | Country        | Cluster |
|--------------------------|---------|----------------|---------|
| 37 Korea, Rep.           | 1       | 40 Mauritius   | 1       |
| 40 Mauritius             | 1       | 42 Mongolia    | 1       |
| 41 Mexico                | 1       | 43 Morocco     | 1       |
| 42 Mongolia              | 1       | 45 Nepal       | 1       |
| 43 Morocco               | 1       | 47 New Zealand | 1       |
| 45 Nepal                 | 1       | 50 Panama      | 1       |
| 47 New Zealand           | 1       | 51 Paraguay    | 1       |
| 50 Panama                | 1       | 52 Peru        | 1       |
| 51 Paraguay              | 1       | 53 Philippines | 1       |
| 52 Peru                  | 1       | 55 Saudi Arabia| 1       |
| 53 Philippines           | 1       | 56 Senegal     | 1       |
| 55 Saudi Arabia          | 1       | 57 South Africa| 1       |
| 56 Senegal               | 1       | 59 Sri Lanka   | 1       |
| 57 South Africa          | 1       | 60 Sudan       | 1       |
| 60 Sudan                 | 1       | 61 Sweden      | 1       |
| 62 Thailand              | 1       | 62 Thailand    | 1       |
| 63 Togo                  | 1       | 63 Togo        | 1       |
| 64 Trinidad and Tobago   | 1       | 64 Trinidad and Tobago | 1 |
| 65 Tunisia               | 1       | 65 Tunisia     | 1       |
| 66 Turkey                | 1       | 66 Turkey      | 1       |
| 67 United Kingdom        | 1       | 67 United Kingdom | 1     |
| 70 Venezuela, RB         | 1       | 70 Venezuela, RB | 1     |
| 72 Zimbabwe              | 1       | 72 Zimbabwe    | 1       |
4.2 Model 3: All countries except China and USA, 1970–2015

In this part of the analysis to determine if China and the USA have an influence on cluster membership, we consider the total CO₂ emissions as well as the significant explanatory variables for all countries, except China and the USA for the period 1970 to 2015.

4.2.1 Cluster analysis with only CO₂ features

We consider maximum, minimum, mean, variance, skewness, and kurtosis for the CO₂ per capita for each country except China and the USA. From the dendrogram of Figure A3 based on Ward’s method, we can identify five possible cluster solutions. The average silhouette coefficients for each method are shown in Table 5 where it is clear that the 2-cluster solution is optimal for all methods.

![Boxplots of CO₂ features for a 2-cluster solution based on all significant explanatory variables: Model 2](image)

![Table 5 Average Silhouette coefficients for CO₂ features: Model 3](image)

| Number of clusters | k-means | k-medoids | Ward | Average | Complete |
|--------------------|---------|-----------|------|---------|----------|
| 2                  | 0.868   | 0.889     | 0.868| 0.918   | 0.918    |
| 3                  | 0.867   | 0.861     | 0.861| 0.861   | 0.861    |
| 4                  | 0.572   | 0.311     | 0.730| 0.870   | 0.581    |
| 5                  | 0.523   | 0.511     | 0.587| 0.740   | 0.594    |

![Fig. 2 Boxplots of CO₂ features for a 2-cluster solution based on all significant explanatory variables: Model 2](image)
On assessing the compatibility of the 2-cluster solution from the Adjusted Rand Indices (Table A7) across the five methods, we observe that for the 2-cluster solution there is 100% compatibility between Ward’s and the k-means methods, and between the average and complete linkage methods. We also observe very good compatibility between Ward’s and the k-medoids methods (83.6%) and between the k-means and k-medoids methods (83.6%).

The results of the 2-cluster solution are presented in Table A2. The findings are identical across Ward’s and the k-means methods. Cluster 2 contains France, India, Japan, and the United Kingdom; surprisingly, these four countries are forefront in the world in implementing nuclear energy resources as a major source of energy mix or until recent years (in the case of Japan). Cluster 1 includes all remaining 68 countries. Previously the inclusion of China and the USA, the two largest emitters of CO₂, would have masked the grouping of the countries in Cluster 2.

4.2.1.1 Profiling of k-means cluster solution We now profile the 2-cluster solutions on each of the CO₂ features that were used to obtain the clusters. They are displayed in boxplots in Fig. 3. The level of the distribution of the maximum, minimum, mean and variance of emissions for Cluster 2 is at a much higher level than that of Cluster 1, with the spread of the distribution of minimum, mean and variance being larger for Cluster 2 than for Cluster 1. The distribution of skewness of Cluster 2 has a slightly lower level than that of Cluster 1, but it has greater spread. The distribution of kurtosis of Cluster 2 has a higher level than that of
Cluster 1, but the spreads are not all that different. Similar observations were made when the 2-cluster solution from Ward’s method was profiled.

### 4.2.2 Clustering with key significant variables contributing to CO₂ emissions

We now determine how the cluster solution based only on the features of CO₂ emissions compares to that of the cluster solution based on the coefficients of the key significant variables contributing to CO₂ emissions for the period 1970–2015, when China and the USA are excluded from the analysis. From the dendrogram of Ward’s method in Figure A4, it seems possible to identify 2, 3, and 4 cluster solutions. The average silhouette coefficients for 2 to 4 clusters are given in Table 6. Since the 2-cluster solution produces the maximum average coefficient for all methods, it may be considered to be the optimal solution. We now assess the compatibility of the 2-cluster solutions across the four methods.

From the Adjusted Rand Indices (Table A8), we observe that for the 2-cluster solution, there is good compatibility between the average and complete linkage methods (85.7%), and of the hierarchical methods, Ward’s method is the most compatibility with non-hierarchical methods (63% and 68%). There is also good compatibility between the k-means and k-medoids methods (83.3%).

The cluster solutions for the Ward and k-means methods are given in Table 7 from where we observe for that four countries switch clusters between these two methods. It does appear as was the case Model 2, that the combined impact of lag 1 of CO₂ emissions, non–renewables, population, and real GDP on CO₂ emissions for Model 3 tends to reasonably separate the countries into two clusters. Each cluster contains a mixture of developed and developing countries, but the Cluster 1 solution for each method contains a much higher percentage of developing countries than Cluster 2; Ward’s Method (Cluster 1: 92%, Cluster 2: 24%), K-means Method (Cluster 1: 91%, Cluster 2: 29%). Cluster 1 has more countries with high emitters than Cluster 2. The sources of increasing CO₂ emissions in recent years occurred in developing countries, where population, urbanization, and economic growth have been the main driving forces as suggested in Nejat et al. (2015).

While the contents of each cluster in this case are not necessarily consistent across all the methods considered, these results are clearly different from just clustering of the features of the CO₂ emissions as observed in Table A3, indicating again the usefulness of including the impact of the significant explanatory variables on CO₂ emissions.

| Number of clusters | k-means | k-medoids | Ward | Average | Complete |
|--------------------|---------|-----------|------|---------|----------|
| 2                  | 0.569   | 0.538     | 0.540| 0.601   | 0.589    |
| 3                  | 0.403   | 0.409     | 0.399| 0.545   | 0.261    |
| 4                  | 0.426   | 0.412     | 0.388| 0.507   | 0.341    |
| Country                  | Cluster | Country                  | Cluster | Country                  | Cluster |
|-------------------------|---------|-------------------------|---------|-------------------------|---------|
| Algeria                 | 1       | 1 Albania               | 2       | 2 Algeria               | 1       |
| Argentina               | 1       | 4 Australia             | 2       | 3 Argentina             | 1       |
| Benin                   | 1       | 5 Austria               | 2       | 7 Benin                 | 1       |
| Bolivia                 | 1       | 6 Bangladesh            | 2       | 8 Bolivia               | 1       |
| Botswana                | 1       | 15 Congo, Dem. Rep.     | 2       | 9 Botswana              | 1       |
| Brazil                  | 1       | 20 Denmark              | 2       | 10 Brazil               | 1       |
| Bulgaria                | 1       | 25 Finland              | 2       | 11 Bulgaria             | 1       |
| Cameroon                | 1       | 26 France               | 2       | 12 Cameroon             | 1       |
| Chile                   | 1       | 27 Greece               | 2       | 13 Chile                | 1       |
| Colombia                | 1       | 32 Italy                | 2       | 14 Colombia             | 1       |
| Congo, Rep.             | 1       | 37 Luxembourg           | 2       | 16 Congo, Rep.          | 1       |
| Costa Rica              | 1       | 38 Malaysia             | 2       | 17 Costa Rica           | 1       |
| Cuba                    | 1       | 43 Mozambique           | 2       | 18 Cuba                 | 1       |
| Cyprus                  | 1       | 45 Netherlands          | 2       | 19 Cyprus               | 1       |
| Dominican Republic      | 1       | 47 Norway               | 2       | 20 Dominican Republic   | 1       |
| Ecuador                 | 1       | 48 Pakistan             | 2       | 21 Ecuador              | 1       |
| Egypt, Arab Rep.        | 1       | 53 Portugal             | 2       | 22 Egypt, Arab Rep.     | 1       |
| Ethiopia                | 1       | 57 Spain                | 2       | 23 Ethiopia             | 1       |
| Honduras                | 1       | 66 United Kingdom       | 2       | 24 Ethiopia             | 1       |
| India                   | 1       | 67 Uruguay              | 2       | 25 India                | 1       |
| Indonesia               | 1       | 69 Zambia               | 2       | 26 Indonesia            | 1       |
| Iran, Islamic Rep.      | 1       | 31 Iran, Islamic Rep.   | 2       | 32 Iran, Islamic Rep.   | 1       |
| Japan                   | 1       | 33 Japan                | 1       | 33 Japan                | 1       |
| Jordan                  | 1       | 34 Jordan               | 1       | 34 Jordan               | 1       |
| Kenya                   | 1       | 35 Kenya                | 1       | 35 Kenya                | 1       |
| Korea, Rep.             | 1       | 36 Korea, Rep.          | 1       | 36 Korea, Rep.          | 1       |
| Mauritius               | 1       | 39 Mauritius            | 1       | 39 Mauritius            | 1       |
| Mongolia                | 1       | 41 Mongolia             | 1       | 41 Mongolia             | 1       |
| Country         | Cluster | Country         | Cluster |
|-----------------|---------|-----------------|---------|
| 39 Mauritius    | 1       | 42 Morocco      | 1       |
| 40 Mexico       | 1       | 44 Nepal        | 1       |
| 41 Mongolia     | 1       | 46 New Zealand  | 1       |
| 42 Morocco      | 1       | 49 Panama       | 1       |
| 44 Nepal        | 1       | 50 Paraguay     | 1       |
| 46 New Zealand  | 1       | 51 Peru         | 1       |
| 49 Panama       | 1       | 52 Philippines  | 1       |
| 50 Paraguay     | 1       | 54 Saudi Arabia | 1       |
| 51 Peru         | 1       | 55 Senegal      | 1       |
| 52 Philippines  | 1       | 56 South Africa | 1       |
| 54 Saudi Arabia | 1       | 58 Sri Lanka    | 1       |
| 55 Senegal      | 1       | 59 Sudan        | 1       |
| 56 South Africa | 1       | 61 Thailand     | 1       |
| 58 Sri Lanka    | 1       | 62 Togo         | 1       |
| 59 Sudan        | 1       | 63 Trinidad and Tobago | 1 |
| 60 Sweden       | 1       | 64 Tunisia      | 1       |
| 61 Thailand     | 1       | 65 Turkey       | 1       |
| 62 Togo         | 1       | 66 United Kingdom | 1 |
| 63 Trinidad and Tobago | 1 | 68 Venezuela, RB | 1 |
| 64 Tunisia      | 1       | 70 Zimbabwe     | 1       |
| 65 Turkey       | 1       |                 |         |
| 68 Venezuela, RB | 1    |                 |         |
| 70 Zimbabwe     | 1       |                 |         |
4.2.2.1 Profiling of k-means cluster solution  Boxplots of the 2-cluster solution are given in Fig. 4. We observe that the levels of distributions of the maximum, minimum and mean of Cluster 2 are slightly higher than that of Cluster 1, whereas the spread of these distribution is slightly wider for Cluster 2 than for Cluster 1. The level and spread of the distribution of the variance of both clusters are approximately the same. The level of the distribution of skewness of Cluster 1 is higher than that of Cluster 2 with the spread of this distribution being approximately the same for both clusters, whereas the level of the distribution of kurtosis of Cluster 2 is higher than that of Cluster 2 with the spread of this distribution being approximately the same for both clusters. Similar observations were made when the 2-cluster solution from Ward’s method was profiled.

4.2.2.2 Comparison of Models 2 and 3  On comparing the Model 3 cluster solution (Table 7) with that of Model 2 (Table 4), we observe that removing China and the USA from the analysis for the period 1970–2015, has not had much of an impact on the composition of the clusters. Two of the four countries that appear together in the 2-cluster solution of the CO2 emissions appear together in each of the clusters in this Model 3 cluster solution (Table 7). However, the profiling of cluster solution for Model 3 indicates slight differences in the CO2 feature levels and spreads between the clusters as is evident as compared to that of Model 2; refer to the boxplots for Models 2 and 3. When China and the USA were included in the analysis in Model 2 (members of Cluster 1 and 2 respectively), it appears that their extremely high CO2 emissions have masked any difference between the levels of the distributions of the minimum,
maximum and means CO$_2$ emissions between the two clusters (Fig. 2). This result implies that in formulating coordinated energy policies in combating emissions, the largest emitters should act together with other countries.

4.3 Model 4: all countries, 1990–2015

In this part of analysis, we consider the total CO$_2$ emissions as well as the significant explanatory variables according to Model 4 for the period 1990 to 2015. We wish to determine if there is any change in cluster composition for this later period compared to the entire period (1970 to 2015) we considered in previous section.

4.3.1 Cluster analysis with only CO$_2$ features

We consider maximum, minimum, mean, variance, skewness, and kurtosis for the total CO$_2$ emissions for each country for the period 1990–2015. From the dendrogram of Ward’s method in Figure A5, we can identify five possible cluster solutions. The average silhouette coefficients for each method are shown in Table 8 from where it is clear that the 2-cluster solution is optimal for the k-means and Ward’s methods, but the 3-cluster solution is optimal for the average, complete and k-medoids methods.

For the 2-cluster solution, we observe from Adjusted Rand Indices (Table A9) that there is 100% compatibility between every pair of methods whereas for the 3-cluster solution, there is 100% compatibility between Ward and k-means methods, Average and complete linkage methods, average linkage and k-medoids methods, complete linkage and k-medoids methods.

We will proceed with the 2-cluster solution which is identical for all methods. We observe from the 2-cluster solution Table A3 that China and the USA group together in one cluster while all the other countries do so in the other cluster. Also, if we consider a 3-cluster solution, China and USA still appear together in one cluster. It is interesting to note here that for this period 1990–2015, China and USA appear together in a cluster while for the period 1970–2015, they appear in separate clusters. Our findings appear to be consistent with the recent initiatives both countries have implemented both changing energy-mix and improving energy technology in combating emissions. Bhattacharya et al. (2015) and Ahmed et al. (2016) discuss

| Number of clusters | k-means | k-medoids | Ward | Average | Complete |
|--------------------|---------|-----------|------|---------|----------|
| 2                  | 0.947   | 0.947     | 0.947| 0.947   | 0.947    |
| 3                  | 0.717   | 0.966     | 0.717| 0.966   | 0.966    |
| 4                  | 0.740   | 0.685     | 0.740| 0.782   | 0.649    |
| 5                  | 0.749   | 0.751     | 0.760| 0.745   | 0.760    |
recent policy changes after the reform period in case of China; while De Gouw et al. (2014) discuss for the U.S.A.

4.3.1.1 Profiling of k-means cluster solution  We now profile the 2-cluster solutions on each of the CO₂ features that were used to obtain the clusters. They are displayed in boxplots in Fig. 5. The level of the distribution of the maximum, minimum, mean and variance of emissions for Cluster 2 which contains China and the USA is at a much higher level than that of Cluster 1, with the spread of these distributions being much wider for Cluster 2 than for Cluster 1. The distribution of skewness of Cluster 2 is approximately at the same level as that of Cluster 1, but it has a greater spread. The distribution of kurtosis of Cluster 2 has a higher level than that of Cluster 1, but the spreads are not all that different. Similar observations were made when the 2-cluster solution from Ward’s method was profiled.

4.3.2 Clustering with key significant variables contributing to CO₂ emissions

We now determine how the cluster solution based only on the features of CO₂ emissions compares to that of the cluster solution based on the coefficients of the key significant variables contributing to CO₂ emissions for the period 1990–2015. From the dendrogram of Ward’s method in Figure A6, it seems possible to identify 2, 3, and 4 cluster solutions. The average silhouette coefficients for 2 to 4 clusters are given in Table 9.
Since the 2-cluster solution produces the maximum average coefficient for all methods, it may be considered to be the optimal solution. We now assess the compatibility of the 2-cluster solutions across the five methods. We observe from the Adjusted Rand Indices (Table A10) there is good compatibility between Ward’s and complete linkage methods (0.831) and of the hierarchical methods, the complete linkage method is most compatibility with the k-means and k-medoids methods (78%, 77.1%).

The cluster solutions for the complete linkage and k-means methods are given in Table 10 from where we observe for the 2-cluster solution, the four countries, viz., Bangladesh, Italy, Spain, and Uruguay, switch clusters between these two methods. It does appear that the combined impact of lag 1 of CO2 emissions, non-renewables, population, and real GDP on CO2 emissions tends to reasonably separate the countries into two clusters. Each cluster contains a mixture of developed and developing countries, but the Cluster 1 solution for each method contains a much higher percentage of developing countries than Cluster 2; Complete linkage Method (Cluster 1: 89%, Cluster 2: 33%), k-means method (Cluster 1: 92%, Cluster 2: 27%). Our findings are consistent across various models in finding Cluster 1 with a high percentage of developing countries compared to Cluster 2.

While the contents of each cluster in this case are not necessarily consistent across all the methods considered, as indicated in the compatibility percentages, these results are clearly different from just clustering of the features of the CO2 emissions as observed in Table A5 indicating the usefulness of including the impact of the significant explanatory variables on CO2 emissions. It is interesting to note that in this case, China and the USA are in different clusters.

### 4.3.2.1 Profiling of k-means cluster solution

We now profile these 2-cluster solutions on each of the CO2 features. Boxplots of the 2-cluster solution are given in Fig. 6. We observe that except of a few outliers, the levels and spreads of distributions of each of the maximum, minimum, mean and variance of Cluster 1 are similar to that of that of Cluster 2. However, the level of the distribution of skewness of Cluster 1 is greater than that of Cluster 2 but the spread of this distribution fare similar for both clusters. Except for a few outliers, the level and spread of the distribution of kurtosis are quite similar. Similar observations were made when the 2-cluster solution from the complete linkage method was profiled.

### 4.3.2.2 Comparison of Models 2 and 4

Comparing the findings from Tables 10 and 4 allows us to determine if differences in period have significantly caused changes in the structure of the clusters. Comparing complete linkage 2-cluster solution, we...
| Country               | Cluster | Country      | Cluster |
|-----------------------|---------|--------------|---------|
| Algeria               | 1       | Albania      | 2       |
| Argentina             | 1       | 4 Australia  | 2       |
| Bangladesh            | 1       | 5 Austria    | 2       |
| Benin                 | 1       | 13 Chile     | 2       |
| Bolivia               | 1       | 16 Congo, Dem. Rep. | 2 |
| Botswana              | 1       | 26 Finland   | 2       |
| Brazil                | 1       | 27 France    | 2       |
| Bulgaria              | 1       | 28 Greece    | 2       |
| Cameroon              | 1       | 35 Jordan    | 2       |
| China                 | 1       | 38 Luxembourg| 2       |
| Colombia              | 1       | 39 Malaysia  | 2       |
| Congo, Rep.           | 1       | 44 Mozambique| 2       |
| Costa Rica            | 1       | 46 Netherlands| 2  |
| Cuba                  | 1       | 48 Norway    | 2       |
| Cyprus                | 1       | 54 Portugal  | 2       |
| Denmark               | 1       | 61 Sweden    | 2       |
| Dominican Republic    | 1       | 68 United States | 2  |
| Ecuador               | 1       | 71 Zambia    | 2       |
| Egypt, Arab Rep.      | 1       | 24 Egypt, Arab Rep. | 1  |
| Ethiopia              | 1       | 25 Ethiopia  | 1       |
| Honduras              | 1       | 29 Honduras  | 1       |
| India                 | 1       | 30 India     | 1       |
| Indonesia             | 1       | 31 Indonesia | 1       |
| Iran, Islamic Rep.    | 1       | 32 Iran, Islamic Rep. | 1   |
| Japan                 | 1       | 34 Japan     | 1       |

Table 10 2-Cluster solutions with all explanatory variables: Model 4

Complete linkage 2-cluster solution

| Country               | Cluster | Country      | Cluster |
|-----------------------|---------|--------------|---------|
| Algeria               | 1       | Albania      | 2       |
| Argentina             | 1       | 4 Australia  | 2       |
| Bangladesh            | 1       | 5 Austria    | 2       |
| Benin                 | 1       | 13 Chile     | 2       |
| Bolivia               | 1       | 16 Congo, Dem. Rep. | 2 |
| Botswana              | 1       | 26 Finland   | 2       |
| Brazil                | 1       | 27 France    | 2       |
| Bulgaria              | 1       | 28 Greece    | 2       |
| Cameroon              | 1       | 35 Jordan    | 2       |
| China                 | 1       | 38 Luxembourg| 2       |
| Colombia              | 1       | 39 Malaysia  | 2       |
| Congo, Rep.           | 1       | 44 Mozambique| 2       |
| Costa Rica            | 1       | 46 Netherlands| 2  |
| Cuba                  | 1       | 48 Norway    | 2       |
| Cyprus                | 1       | 54 Portugal  | 2       |
| Denmark               | 1       | 61 Sweden    | 2       |
| Dominican Republic    | 1       | 68 United States | 2  |
| Ecuador               | 1       | 71 Zambia    | 2       |
| Egypt, Arab Rep.      | 1       | 24 Egypt, Arab Rep. | 1  |
| Ethiopia              | 1       | 25 Ethiopia  | 1       |
| Honduras              | 1       | 29 Honduras  | 1       |
| India                 | 1       | 30 India     | 1       |
| Indonesia             | 1       | 31 Indonesia | 1       |
| Iran, Islamic Rep.    | 1       | 32 Iran, Islamic Rep. | 1   |
| Japan                 | 1       | 34 Japan     | 1       |

K-means 2-cluster solution

| Country               | Cluster | Country      | Cluster |
|-----------------------|---------|--------------|---------|
| Algeria               | 1       | 1 Albania    | 2       |
| Argentina             | 1       | 4 Australia  | 2       |
| Bangladesh            | 1       | 5 Austria    | 2       |
| Benin                 | 1       | 6 Bangladesh | 2       |
| Bolivia               | 1       | 13 Chile     | 2       |
| Botswana              | 1       | 16 Congo, Dem. Rep. | 2 |
| Brazil                | 1       | 16 Congo, Dem. Rep. | 2  |
| Bulgaria              | 1       | 26 Finland   | 2       |
| Cameroon              | 1       | 27 France    | 2       |
| China                 | 1       | 28 Greece    | 2       |
| Colombia              | 1       | 35 Jordan    | 2       |
| Congo, Rep.           | 1       | 38 Luxembourg| 2       |
| Costa Rica            | 1       | 39 Malaysia  | 2       |
| Cuba                  | 1       | 44 Mozambique| 2       |
| Cyprus                | 1       | 46 Netherlands| 2  |
| Denmark               | 1       | 48 Norway    | 2       |
| Dominican Republic    | 1       | 54 Portugal  | 2       |
| Ecuador               | 1       | 61 Sweden    | 2       |
| Egypt, Arab Rep.      | 1       | 68 United States | 2  |
| Ethiopia              | 1       | 71 Zambia    | 2       |
| Honduras              | 1       | 24 Egypt, Arab Rep. | 1  |
| India                 | 1       | 25 Ethiopia  | 1       |
| Indonesia             | 1       | 29 Honduras  | 1       |
| Iran, Islamic Rep.    | 1       | 30 India     | 1       |
| Japan                 | 1       | 31 Indonesia | 1       |
| Japan                 | 1       | 32 Iran, Islamic Rep. | 1   |

Environmental and Ecological Statistics (2020) 27:1–40
Table 10 (continued)

| Country  | Cluster | Country       | Cluster |
|----------|---------|---------------|---------|
| Italy    | 1       | Kenya         | 1       |
| Japan    | 1       | Korea, Rep.   | 1       |
| Kenya    | 1       | Mauritius     | 1       |
| Korea, Rep. | 1    | Mexico        | 1       |
| Mauritius| 1       | Mongolia      | 1       |
| Morocco  | 1       | Morocco       | 1       |
| Nepal    | 1       | Nepal         | 1       |
| New Zealand | 1    | New Zealand   | 1       |
| Pakistan | 1       | Pakistan      | 1       |
| Panama   | 1       | Panama        | 1       |
| Paraguay | 1       | Philippines   | 1       |
| Peru     | 1       | Saudi Arabia  | 1       |
| Philippines | 1    | Senegal       | 1       |
| Saudi Arabia | 1   | South Africa  | 1       |
| Senegal  | 1       | Sri Lanka     | 1       |
| South Africa | 1   | Sudan         | 1       |
| Spain    | 1       | Thailand      | 1       |
| Sri Lanka| 1       | Togo          | 1       |
| Sudan    | 1       | Trinidad and Tobago | 1 |
| Thailand | 1       | Tunisia       | 1       |
| Togo     | 1       | Turkey        | 1       |
| Country               | Cluster | Country               | Cluster |
|-----------------------|---------|-----------------------|---------|
| Trinidad and Tobago   | 1       | United Kingdom        | 1       |
| Tunisia               | 1       | Venezuela, RB         | 1       |
| Turkey                | 1       | Zimbabwe              | 1       |
| United Kingdom        | 1       |                       |         |
| Uruguay               | 1       |                       |         |
| Venezuela, RB         | 1       |                       |         |
| Zimbabwe              | 1       |                       |         |

Table 10 (continued)
notice countries like Bangladesh, Denmark, Italy, Spain, Sri Lanka, and Uruguay have shifted from Cluster 2 to Cluster 1 for the later period (1990–2015), while only one country Cameroon has disappeared from Cluster 1 when we shift the time from 1970–2015 to 1990–2015. We notice, the changes over time, has had some impact on the structure of the clusters.

### 4.4 Model 5: all countries except China and USA, 1990–2015

In this sub-section, we consider the total CO₂ emissions as well as the significant explanatory variables for all countries, except China and the USA for the period 1990 to 2015.

#### 4.4.1 Cluster analysis with only CO₂ features

We now consider maximum, minimum, mean, variance, skewness, and kurtosis for the CO₂ for each country except China and the USA, for the period 1990–2015. From the dendrogram using Ward’s method in Figure A7, we can identify four possible cluster solutions. The average silhouette coefficients for each method are shown in Table 11 from where we observe that the 2-cluster solution is optimal for the k-means, Ward and average linkage methods while the 3-cluster solution is optimal for the complete linkage and the k-medoids methods.
The results of the compatibility analysis from the Adjusted Rand Indices (Table A11) between the three hierarchical methods and the two non-hierarchical methods show that there is 100% compatibility between Ward’s and complete linkage method, Ward’s and the k-means method, and between Complete linkage and k-means methods for the 2-cluster solution. For the 3-cluster solution there is an 81.8% compatibility between Ward’s and the k-means methods and 100% compatibility between complete linkage and the k-medoids methods. For all other pairs, there is very poor compatibility. We will proceed with the 2-cluster solution which has better compatibility outcomes than the 3-cluster solution.

The results of the 2-cluster solution are presented in Table A4. The findings are identical across Ward’s and the k-means methods. Cluster 2 contains France, India, Japan, and the United Kingdom, and Cluster 1 includes all remaining 68 countries. This solution is exactly the same cluster solution for the 1970–2015 period when China and the USA were excluded. As before it appears that the inclusion of China and the USA, the two highest emitters of CO₂, has masked the grouping of the countries in Cluster 2. All of these four countries viz. France, India, Japan, and the United Kingdom have taken significant initiatives to move from non-renewable energy resources (particularly in Nuclear energy) during this time.

### 4.4.1.1 Profiling of k-means cluster solution

We now profile the 2-cluster solutions on each of the CO₂ features that were used to obtain the clusters. They are displayed in boxplots in Fig. 7. The level of the distribution the of maximum, minimum, mean and variance of emissions for Cluster 2 is at a much higher level than that of Cluster 1, with the spread of the distribution of minimum, mean and variance being larger for Cluster 2 than for Cluster 1. The distribution of skewness of Cluster 2 has a slightly lower level than that of Cluster 1, but it has greater spread. The distribution of kurtosis of Cluster 2 has a higher level and wider spread than that of Cluster 1. Similar patterns were observed when the 2-cluster solution from Ward’s method was profiled.

### 4.4.2 Clustering with key significant variables contributing to CO₂ emissions

We will now determine how this cluster solution based only on the features of CO₂ emissions compares when the clusters are based on the coefficients of the key significant variables contributing to CO₂ emissions. We exclude China and USA from the analysis and the considered time period is 1990–2015.
From the dendrogram of Ward’s method in Figure A8 it seems possible to identify 2, 3, and 4 cluster solutions. The average silhouette coefficients for 2 to 4 clusters are given in Table 12. The 2-cluster solution produces the maximum average silhouette coefficient for all methods except the k-medoids method. The k-medoids method produces the maximum average silhouette coefficient for the 3-cluster solution but given all the other methods produces much lower average silhouette coefficient, we will proceed with the 2-cluster solution as the optimal solution. We now assess the compatibility of the 2-cluster solutions across the five methods.

From the Adjusted Rand Indices (Table A12) from observe that for the 2-cluster solution, there is poor compatibility between all of the pairs except for that between Ward’s and the k-means methods (71.8%).

The cluster solutions for the Ward and k-means methods are given in Table 13. We notice that three countries switch clusters between these two methods. It does appear that the combined impact of lagged CO₂ emissions, non-renewables, population, and real GDP on CO₂ emissions for Model 5 tends to reasonably separate the countries into two clusters. Each cluster contains a mixture of developed and
Table 13 2-Cluster solutions with all explanatory variables: Model 5

| Country                  | Cluster | Country     | Cluster |
|--------------------------|---------|-------------|---------|
| 2 Algeria                | 1       | 1 Albania   | 2       |
| 3 Argentina              | 1       | 4 Australia | 2       |
| 6 Bangladesh             | 1       | 5 Austria   | 2       |
| 7 Benin                  | 1       | 13 Chile    | 2       |
| 8 Bolivia                | 1       | 19 Cyprus   | 2       |
| 9 Botswana               | 1       | 25 Finland  | 2       |
| 10 Brazil                | 1       | 26 France   | 2       |
| 11 Bulgaria              | 1       | 27 Greece   | 2       |
| 12 Cameroon              | 1       | 32 Italy    | 2       |
| 14 Colombia              | 1       | 37 Luxembourg | 2     |
| 15 Congo, Dem. Rep.      | 1       | 38 Malaysia | 2       |
| 16 Congo, Rep.           | 1       | 43 Mozambique | 2   |
| 17 Costa Rica            | 1       | 45 Netherlands | 2   |
| 18 Cuba                  | 1       | 47 Norway   | 2       |
| 20 Denmark               | 1       | 53 Portugal | 2       |
| 21 Dominican Republic    | 1       | 58 Sri Lanka | 2    |
| 22 Ecuador               | 1       | 60 Sweden   | 2       |
| 23 Egypt, Arab Rep.      | 1       | 69 Zambia   | 2       |
| 24 Ethiopia              | 1       | 28 Honduras | 1       |
| 29 India                 | 1       | 29 India    | 1       |
| 30 Indonesia             | 1       | 30 Indonesia | 1    |
| 31 Iran, Islamic Rep.    | 1       | 31 Iran, Islamic Rep. | 1 |
| 33 Japan                 | 1       | 33 Japan    | 1       |

K-means 2-cluster solution

| Country                  | Cluster | Country     | Cluster |
|--------------------------|---------|-------------|---------|
| 2 Algeria                | 1       | 1 Albania   | 2       |
| 3 Argentina              | 1       | 4 Australia | 2       |
| 7 Benin                  | 1       | 5 Austria   | 2       |
| 8 Bolivia                | 1       | 13 Chile    | 2       |
| 9 Botswana               | 1       | 19 Cyprus   | 2       |
| 10 Brazil                | 1       | 25 Finland  | 2       |
| 11 Bulgaria              | 1       | 26 France   | 2       |
| 12 Cameroon              | 1       | 27 Greece   | 2       |
| 14 Colombia              | 1       | 32 Italy    | 2       |
| 15 Congo, Dem. Rep.      | 1       | 37 Luxembourg | 2    |
| 16 Congo, Rep.           | 1       | 38 Malaysia | 2       |
| 17 Costa Rica            | 1       | 43 Mozambique | 2   |
| 18 Cuba                  | 1       | 45 Netherlands | 2   |
| 19 Cyprus                | 1       | 47 Norway   | 2       |
| 20 Denmark               | 1       | 53 Portugal | 2       |
| 21 Dominican Republic    | 1       | 58 Sri Lanka | 2    |
| 22 Ecuador               | 1       | 60 Sweden   | 2       |
| 23 Egypt, Arab Rep.      | 1       | 69 Zambia   | 2       |
| 24 Ethiopia              | 1       | 28 Honduras | 1       |
| 29 India                 | 1       | 29 India    | 1       |
| 30 Indonesia             | 1       | 30 Indonesia | 1    |
| 31 Iran, Islamic Rep.    | 1       | 31 Iran, Islamic Rep. | 1 |
| 33 Japan                 | 1       | 33 Japan    | 1       |
Table 13 (continued)

| Country          | Cluster | Country          | Cluster |
|------------------|---------|------------------|---------|
| 34 Jordan        | 1       | 34 Jordan        | 1       |
| 35 Kenya         | 1       | 35 Kenya         | 1       |
| 36 Korea, Rep.   | 1       | 36 Korea, Rep.   | 1       |
| 39 Mauritius     | 1       | 39 Mauritius     | 1       |
| 40 Mexico        | 1       | 40 Mexico        | 1       |
| 41 Mongolia      | 1       | 41 Mongolia      | 1       |
| 42 Morocco       | 1       | 42 Morocco       | 1       |
| 44 Nepal         | 1       | 44 Nepal         | 1       |
| 46 New Zealand   | 1       | 46 New Zealand   | 1       |
| 48 Pakistan      | 1       | 48 Pakistan      | 1       |
| 49 Panama        | 1       | 49 Panama        | 1       |
| 50 Paraguay      | 1       | 50 Paraguay      | 1       |
| 51 Peru          | 1       | 51 Peru          | 1       |
| 52 Philippines   | 1       | 52 Philippines   | 1       |
| 54 Saudi Arabia  | 1       | 54 Saudi Arabia  | 1       |
| 55 Senegal       | 1       | 55 Senegal       | 1       |
| 56 South Africa  | 1       | 56 South Africa  | 1       |
| 57 Spain         | 1       | 57 Spain         | 1       |
| 59 Sudan         | 1       | 59 Sudan         | 1       |
| 61 Thailand      | 1       | 61 Thailand      | 1       |
| 62 Togo          | 1       | 62 Togo          | 1       |
| 63 Trinidad and Tobago | 1    | 63 Trinidad and Tobago | 1    |
| 64 Tunisia       | 1       | 64 Tunisia       | 1       |
| Country         | Cluster | Country         | Cluster |
|-----------------|---------|-----------------|---------|
| 65 Turkey       | 1       | 65 Turkey       | 1       |
| 66 United Kingdom| 1       | 66 United Kingdom| 1       |
| 67 Uruguay      | 1       | 68 Venezuela, RB| 1       |
| 68 Venezuela, RB| 1       | 70 Zimbabwe     | 1       |
| 70 Zimbabwe     | 1       |                 |         |
developing countries, but the Cluster 1 solution for each method contains a much higher percentage of developing countries than Cluster 2; Ward’s Method (Cluster 1: 90%, Cluster 2: 22%), K-means Method (Cluster 1: 92%, Cluster 2: 26%). These findings are similar to our previous observations.

While the contents of each cluster in this case are not necessarily consistent across all the methods considered, as indicated in the compatibility percentages these results are clearly different from just clustering of the features of the CO2 emissions as observed in Table A4, indicating again the usefulness of including the impact of the significant explanatory variables on CO2 emissions.

### 4.4.2.1 Profiling of k-means cluster solution

We now profile the 2-cluster solutions on each of the CO2 features. Boxplots of the 2-cluster solution are given in Fig. 8. We observe that the levels of distributions of the maximum, minimum, and mean of Cluster 2 are slightly higher than that of Cluster 1. The spread of the distribution the minimum and mean are slightly wider for Cluster 2 than for Cluster 1. Except for a few outliers, the level and spread of the distribution of the variance of both clusters are approximately the same. The level of the distribution of skewness of Cluster 1 is higher than that of Cluster 2 with the spread of this distribution being a bit narrower for Cluster. The level of the distribution of kurtosis of Cluster 2 is higher than that of Cluster 2 with the spread of this distribution being approximately the same for both clusters. Similar observations were made when the 2-cluster solution from Ward’s method was profiled.
4.4.2.2 Comparison of Models 4 and 5 Here we compare the Model 5 cluster solution (Table 13) with that of Model 4 (Table 10), we observe that removing China and the USA from the analysis for the period 1970–2015, has not had much of an impact on the composition of the clusters. Three of the four countries that appear together in the 2-cluster solution of the CO2 emissions appear together in Cluster 1 for this Model 5 cluster solution (Table 13). However, the profiling of this cluster solutions indicates slight differences in the CO2 feature levels and spreads between the clusters as is evident when comparing the boxplots in Figs. 6 and 8. Clearly, when China and the USA were included in the analysis in Model 4, it appears that inclusion of China and the USA as the largest emitters may have masked any difference between the levels of the distributions of the minimum, maximum and means CO2 emissions between the two clusters (in Figure A3).

4.4.2.3 Comparison of Models 3 and 5 Finally, we compare the Model 3 cluster solution (Table 7) with that of Model 5 (Table 13). Model 3 which covers the full period, while Model 5 is for 1990–2015. In both cases, we exclude China and the U.S.A. for analysis. For complete linkage 2-cluster solution, Chile, Cyprus, Sri Lanka, Sweden shifted from Cluster 1 to Cluster 2, while Pakistan, Uruguay, Bangladesh, Spain, Denmark, UK, and the Congo Democratic Republic shifted from Cluster 2 to Cluster 1. For the k-means 2-cluster solution, Congo Democratic Republic, Congo Republic, Mexico, Pakistan shifted from Cluster 2 to Cluster 1. These changes reflect the selection of period has a significant influence in cluster formation. The findings reflect the presence of China and the U.S.A has an influence on the structure of the cluster. Here, there is a significant shift of countries from Cluster 1 to Cluster 2 and vice versa when you exclude the largest emitters from the sample.

4.5 Clustering and de-carbonization

It seems reasonable to summarise that over the period 1970 to 2015, there are differences in the levels of CO2 emissions between the identified clusters of countries considering combined effects of the three key determinants (non-renewables, population, and real GDP). While we have more than one solution in each case, the findings can guide us in forming clusters based on the relationship between CO2 emissions and each of these key determinants influencing CO2 emissions.

The findings of the cluster analysis considering total CO2 emissions alone revealed that the 72 countries separated into three distinct clusters, two single-membered countries with China and the USA and the other containing the remaining 70 countries within our panel. This is to be expected given that China and the USA are the largest CO2 emitters, with differing maximum, minimum, and mean CO2 emissions over the 46 years from 1970 to 2015. One clear observation from our analysis is that clustering separation is different when we consider key determinants of CO2 emissions compared to the clustering separation we obtain only with CO2 features.

Overall, each of these three key determinants—non-renewables, population, and real GDP play different roles across countries in forming clusters and their combined
effects on CO₂ emissions are different across countries; therefore, clusters form with different members. Findings from our full panel with entire time reflect that China and the USA in isolation should take the lead role in reducing overall CO₂ emissions while the other countries in our panel follow suit. To test the validation of this claim we have excluded China and the U.S. and considered sample period as 1990–2015. Various specifications result in a different structure of clusters, albeit an optimum number of clusters remain the same.

5 Conclusions and policy implications

Our primary objective of this research is to apply cluster analysis in examining a stochastic version of the IPAT model to decompose per capita emissions with three major components: population, affluence, and technology. We establish stable groups of countries forming clusters; based on the factors we identify to influence CO₂ emissions. The three key factors we identified are non-renewables, population, and real GDP. We summarise the key messages from this research in the next few paragraphs.

Firstly, our findings reflect that these factors can influence the polluting behaviour of individual countries differently. Secondly, each cluster is characterized by the fact that its CO₂ emissions are, to a significant extent, linked to a particular explanatory variable. This also links the differentiation between each cluster from the others. Each cluster therefore, may have a distinct role in contributing towards overall CO₂ emissions, identified for specific clusters of countries. This finding indicates a need for distinct policies and actions for the various clusters.

In future work, we will consider the use of the Mahalanobis distance measure which takes into account the correlation between the clustering variables. We will also consider fuzzy clustering to determine overlapping clusters. This may give us some insight into which countries could belong to more than one cluster based on the CO₂ features and based on the variables contributing to CO₂ levels.

We have identified three key determinants, non-renewable energy, population, and real GDP, which play a significant role in increasing CO₂ emissions and forming the clusters. Economic growth will increase in many countries, particularly the Asian, African, and South American continents, in the coming decades. Therefore, economic policies combating population, and simultaneous reduction of the use of non-renewable energy sources, can mitigate CO₂ emissions both for individual countries and for the rest of the world. In this respect, countries within similar clustering may coordinate to implement policies in reducing emissions and enhance economic growth. With changes in economic, demographic, and energy policies, along with other countries, the role of China and the USA will be significant in forecasting the overall CO₂ emissions trajectory. Our analysis also emphasises the roles of time and the largest emitters in controlling future emissions. Future research should include the role of trade and other variables such as political and social factors in explaining CO₂ clusters. The detail input–output data are needed across a large panel of countries to incorporate this analysis.
Author contributions  All authors contributed to the development of ideas and methods. JI collected the data and wrote the code in STATA for the generation output of the models; EAM wrote the code in MATLAB for the generation of the cluster analysis output; all authors contributed to the writing the paper and editing it to arrive at the final draft.

References

Ahmed K, Bhattacharya M, Qazi AQ, Long W (2016) Energy consumption in China and underlying factors in a changing landscape: empirical evidence since the reform period. Renew Sust Energ Rev 58:224–234

Aldy JE (2006) Per capita carbon dioxide emissions: convergence or divergence? Environ Resour Econ 33(4):533–555

Anandalingam G, Bhattacharya D (1985) Process modelling and industrial energy use in developing countries—the steel industry in India. Omega 13(4):295–306

Anderberg MR (2014) Cluster analysis for applications: probability and mathematical statistics: a series of monographs and textbooks. Academic press, Cambridge

Apergis N (2016) Environmental Kuznets curves: new evidence on both panel and country- level CO₂ emissions. Energy Econ 54:263–271

Apergis N, Payne JE (2010) Renewable energy consumption and growth in Eurasia. Energy Econ 32(6):1392–1397

Awaworyi Churchill S, Inekwe J, Ivanovski K (2018) Conditional convergence in per capita carbon emissions since 1900. Appl Energy 228:916–927

Bai J (2009) Panel data models with interactive fixed effects. Econometrica 77(4):1229–1279

Bhattacharya M, Rafiq S, Bhattacharya S (2015) The role of technology on the dynamics of coal consumption—economic growth: new evidence from China. Appl Energy 154:686–695

Bhattacharya M, Paramati SR, Ozturk I, Bhattacharya S (2016) The effect of renewable energy consumption on economic growth: evidence from top 38 countries. Appl Energy 162:733–741

Bhattacharya M, Inekwe JN, Sadorsky P, Saha A (2018) Convergence of energy productivity across Indian states and territories. Energy Econ 74:427–440

Bhattacharya M, Inekwe JN, Sadorsky P (2020) Convergence of energy productivity in Australian states and territories: determinants and forecasts. Energy Econ 85:104538

Cai Y, Sam CY, Chang T (2018) Nexus between clean energy consumption, economic growth and CO₂ emissions. J Clean Prod 182:916–1011

Caiado J, Maharaj E, D’urso P (2015) Time series clustering. Handbook of cluster analysis. Chapman and Hall/CRC. Boca Raton

Change IPOC (1990) Climate change: the IPCC scientific assessment. Mass, Cambridge

Chudik A, Pesaran MH (2015) Common correlated effects estimation of heterogeneous dynamic panel data models with weakly exogenous regressors. J Economet 188(2):393–420

Churchill SA, Inekwe J, Ivanovski K, Smyth R (2018) The environmental Kuznets curve in the OECD: 1870–2014. Energy Econ 75:389–399

Cole MA, Neumayer E (2004) Examining the impact of demographic factors on air pollution. Popul Environ 26(1):5–21

Criado CO, Grether JM (2011) Convergence in per capita CO₂ emissions: a robust distributional approach. Resour Energy Econ 33(3):637–665

D’Urso P, Di Lallo D, Maharaj EA (2013) Autoregressive model-based fuzzy clustering and its application for detecting information redundancy in air pollution monitoring networks. Soft Comput 17(1):83–131

Dai X, Kuosmanen T (2014) Best-practice benchmarking using clustering methods: application to energy regulation. Omega 42(1):179–188

De Gouw J, Parrish D, Frost G, Trainer M (2014) Reduced emissions of CO₂, NOₓ, and SO₂ from US power plants owing to switch from coal to natural gas with combined cycle technology. Earth’s Fut 2(2):75–82

Dietz T, Rosa EA (1994) Rethinking the environmental impacts of population, affluence and technology. Hum Ecol Rev 1(2):277–300
Strazicich MC, List JA (2003) Are CO₂ emission levels converging among industrial countries? Environ Resour Econ 24(3):263–271
Vilar JA, Lafuente-Rego B, D’Urso P (2018) Quantile autocovariances: a powerful tool for hard and soft partitional clustering of time series. Fuzzy Sets Syst 340:38–72
Wang A, Lin B (2017) Assessing CO₂ emissions in China’s commercial sector: determinants and reduction strategies. J Clean Prod 164:1542–1552
Ward JH Jr (1963) Hierarchical grouping to optimize an objective function. J Am Stat Assoc 58(301):236–244
World Bank (2015) World development indicators (WDI) database
Wu TH, Chen YS, Shang W, Wu JT (2018) Measuring energy use and CO₂ emission performances for APEC economies. J Clean Prod 183:590–601
Zhou WY, Wi Yang, Wan WX, Zhang J, Zhou W, Yang HS, Yang H, Xiao H, Deng SH, Shen F, Wang YJ (2018) The influences of industrial gross domestic product, urbanization rate, environmental investment, and coal consumption on industrial air pollutant emission in China. Environ Ecol Stat 25(4):429–442

**John Inekwe** is Research Fellow at the Centre for Risk Analytics, and the Department of Applied Finance, Macquarie University, Australia. John holds a Ph.D. from Monash University and his key areas of research include Financial Econometrics, Monetary Policy, Financial Networks, and Energy Markets. John was awarded the prestigious Macquarie University Research Fellowship. He has published papers in Macroeconomic Dynamics, Economics Letters, Energy Economics, Scottish Journal of Political Economy, Journal of Small Business Management, International Review of Economics and Finance, Social Indicators Research, Emerging Markets Review, Economic Modelling, Applied Energy, Empirical Economics, Australian Journal of Management, Applied Economics and The World Economy. For more information, visit [https://researchers.mq.edu.au/en/persons/john-inekwe](https://researchers.mq.edu.au/en/persons/john-inekwe).

**Elizabeth Ann Maharaj** is Associate Professor in the Department of Econometrics and Business Statistics at the Caulfield Campus of Monash University. She is an elected member of the International Statistical Institute (ISI), a member of the International Association of Statistical Computing (IASC) and of the Statistical Society of Australia (SSA). She is also an accredited statistician with the SSA. Elizabeth’s main research interests are in time series classification, wavelets analysis, fuzzy classification and interval time series analysis. As well as co-authoring a book on time series clustering and classification, she has published many papers on these topics. She has also published papers in climatology, environmental science, labour markets, human mobility and finance. Further details can be found at [https://research.monash.edu/en/persons/ann-maharaj](https://research.monash.edu/en/persons/ann-maharaj).

**Mita Bhattacharya** is Senior Lecturer in Economics at Monash University, Australia. Mita has published widely in Applied Economics with a recent focus on energy-related topics. Amongst other journals, her publications appear in Energy Economics, Applied Energy, Renewable Energy, Renewable and Sustainable Energy Reviews and Energy journals. Mita is the recipient of the ‘Applied Energy 2016–2017 highly-cited research paper award’. Mita is the Director of the Masters Program in Economics at Monash. More information can be found at [https://research.monash.edu/en/persons/mita-bhattacharya](https://research.monash.edu/en/persons/mita-bhattacharya).