EPJ Nuclear Sci. Technol. **8**, 24 (2022)
© D. Perez, Published by EDP Sciences, 2022
https://doi.org/10.1051/epjn/2022017

EPJ Nuclear Sciences & Technologies

Available online at:
https://www.epj-n.org

Regular Article

Open Access

An attempt of reproduction of Sovacool et al.'s "Differences in carbon emissions reduction between countries pursuing renewable electricity versus nuclear power"

Daniel Perez[1,2,3]*

[1] Département de mathématiques et applications, École normale supérieure, CNRS, PSL Université, 75005 Paris, France
[2] Laboratoire de mathématiques d'Orsay, Université Paris-Saclay, CNRS, 91405 Orsay, France
[3] DataShape, Centre Inria Saclay, 91120 Palaiseau, France

Received: 13 January 2022 / Received in final form: 5 June 2022 / Accepted: 20 July 2022

Abstract. In this paper, we attempt to reproduce the results obtained by Sovacool et al. in their recent paper that focuses on the differences in carbon emissions reduction between countries pursuing renewable electricity versus nuclear power. We have found several flaws in the models and the statistical analysis performed theirein, notably the correlations performed between the fractions of renewable power and of nuclear power and greenhouse gas emissions per capita and the lack of consideration for natural bias between the variables examined.

1 Introduction

Lowering Greenhouse Gas (GHG) emissions, foremost amongst which are carbon dioxide emissions, has been established as a priority in order to mitigate the effects of anthropogenic climate change. While it is clear that abandoning fossil fuels is imperative, there is still some debate about the details of the transition to decarbonized sources of energy. As it was reported in Chapter 2 of the 2018 IPCC report [1], the role of nuclear energy increases along most pathways to decarbonation, although the variance in the share of nuclear energy is quite large across the spectrum of the different models and paths considered in the literature [2,3]. For instance, there are scenarios of 100% renewable energy which have been considered by some authors [4,5], although the validity of the assumptions of these high-renewable models have been contested [6]. By contrast, there are also examples where the role of nuclear power is greatly increased, such as in [7–10].

In the context of this debate, Sovacool et al. performed a study concluding that "the implication for electricity planning is that diverse renewables are generally proving in the real world to be significantly more effective than nuclear power at reducing climate disruption" [11]. Fell et al. have since published a response [12], criticizing the methodology and other aspects of their paper. In particular, the authors find that "nuclear power and renewable energy are both associated with lower per capita CO_2 emissions with effects of similar magnitude and statistical significance".

Similarly, Wagner [13] has recently obtained results in direct contradiction to the conclusions drawn by Sovacool et al. This paper provides supplementary criticism for the validity of the conclusions of [11] and concludes that "both, nuclear and renewable power allow a reduction of national CO_2 emission levels. [...] The analysis of the databases employed in this study did not yield evidence for any further hidden variables, national CO_2 emission might possibly depend on. Specifically, no evidence for "crowding-out" can be detected".

Sovacool et al.'s paper relies on the statistical analysis of historical data available for different variables across a variety of countries. In particular, it relies on establishing correlations amongst the share of nuclear energy (henceforth denoted N) versus renewable energy (henceforth denoted R) as a fraction of the electrical mix and CO_2eq emissions per capita, while taking Gross Domestic Product (GDP) per capita as a confounding variable.

After attempting to reproduce the results of Sovacool et al. [11], we have found that the analysis performed is considerably flawed both because there were mistakes in the statistical analysis and because there were inconsistencies in the logic of the authors, in particular concerning:

- the "crowding out" hypothesis, i.e. that renewables and nuclear power are structurally incompatible, so there is an anticorrelation between them;

* e-mail: daniel.perez@ens.fr

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

- the rejection of the "climate mitigation" hypothesis, which states that "the relative scale of national attachments to nuclear electricity production will vary negatively with carbon emissions".

Both of these elements involved regressions of non-carbonated sources of electricity with GHG emissions, despite the fact that decarbonated energy sources are not good predictors of GHG emissions.

The rest of this paper is separated as follows. First, we will give a more detailed account of each of the arguments above. Then, we will give some complementary technical details regarding the data set and our analysis of the complementary data provided by the authors.

2 Criticism

2.1 Fossil fuels as the real predictor and the "crowding out" hypothesis

Both renewable and nuclear energy emit little to no GHGs, but energy stemmming from fossil fuels does. With respect to the GHG emissions per capita, the only relevant variable is the fraction of fossil fuels in the electricity production of each country (which we will denote F). It follows that the fraction of nuclear energy or renewables in the electrical mix is not a good predictor of GHG emissions, independent of statistical treatment of the data. As shown by our in depth analysis in the appendix, the rejection of the authors of the "climate mitigation" hypothesis arises from an inadequate statistical analysis (cf. Sect. 2.2) and from the following fact.

The fraction of electricity produced with renewable, nuclear and fossil fuels satisfies the tautological relation:

$$R + N + F = 1. \tag{1}$$

This relation implies that these three variables are necessarily correlated with one another. In light of this reasoning, no matter the statistical treatment of the data the predictive power of R or N for the GHG emissions can only stem from that of F (cf. Proposition A.1. and Appendix A.4). In particular, the analysis of the authors of [11] on the effect of renewables and nuclear energy as well as their rejection of the "climate mitigation" hypothesis reflects nothing other than relation (1).

Moreover, the reasoning behind the "crowding out" hypothesis is flawed. Indeed, the authors of [11] motivate the proposal of the "crowding out" hypothesis as follows. Intermittent renewables require a decentralized electrical infrastructure as soon as they occupy a significant fraction of the electricity produced. By contrast, the optimal electrical infrastructure of non-intermittent power sources, such as fossil fuels, hydroelectricity and nuclear power is centralized [14]. The authors then suggest that, for these reasons, there should be an anticorrelation between R and N, which is the statement of the so-called "crowding out" hypothesis. They back this statement by verifying that R and N are indeed anticorrelated and use this to justify their statements.

However, this explanation is inconsistent with the data studied, since most of the electric production considered to be "renewable" from 1990 to 2015 was hydroelectricity – with intermittent power sources such as wind and solar contributing only negligible amounts to the statistic according to the BP Statistical Review of World Energy (2019) [15]. Furthermore, given any three positive random variables satisfying relation (1), one can always find at least two pairs of variables that are negatively correlated (cf. Proposition A.1.). It is thus little to no surprise to find that R and N are negatively correlated, but this has nothing to do with the causality relation of the "crowding out" hypothesis suggested by Sovacool et al. It is simply the consequence of the simple mathematical relation between the variables studied. This statement is backed by our in-depth analysis in the appendix of this paper.

2.2 Flaws in the statistical analysis and the rejection of the "climate mitigation" hypothesis

Sovacool et al. propose two timeframes (1990–2004 and 2000–2014) along which the data is are split and averaged and justify this by claiming this is "an optimal use of the data", because "renewable energy figures were only recorded since the nineties". However, the averaging procedure of the authors is not justified from a time series analysis (TSA) perspective nor does it exploit the data in any sense of optimality (from a statistical standpoint). Furthermore, that this averaging procedure does not affect the conclusions of the paper should have at the very least been shown by the authors by demonstrating the stability of this procedure, i.e. whether a change in the time step and number of timeframes considered changes the conclusions of the regression analysis or not. However, this was never made explicit in the paper. In general, disregarding TSA considerations may lead to modifications in the results of any subsequent analysis, as many potential time series complications could arise, in particular non-stationarity [16–18]. Note also that such concerns could have been easily foreseen, as it is not surprising that the data studied is are non-stationary, since many countries underwent rapid industrialization during the studied time period. This a priori arbitrary treatment of the data questions the integrity of the data set used for the subsequent analysis, and by extension, the entire analysis itself and its conclusions.

However, even if the averaging procedure of the authors turns out to be stable and assuming there are no TSA complications in the study of this data set, there are many inconsistencies and flaws in the subsequent statistical analysis performed in [11]. These will be treated in more detail in the appendix, but a non-exhaustive list includes:

- given the nature of the conclusions and the context of Sovacool et al.'s study, the forward selection performed is inadequate (cf. Appendix A.4): a more appropriate approach would be to consider bidirectional selection, as it also excludes independent variables which do not play a significant role in the predictive power of the model [19].

In other words, the strength of the conclusions drawn by the authors and their policy recommendations cannot be backed by the analysis performed in the paper, since one cannot draw strong conclusions about the relative importance of the variables in the model using forward selection. If the objective was to draw these conclusions, bidirectional selection is more appropriate.

- The poor study of the data set before the start of the regression analysis (for instance, there was no check for heteroskedascity), which inevitably led to a suboptimal model, i.e. one with too many variables (or inappropriate ones) – some of which turn out not to be significative – without an increase in goodness of fit, or predictive power (cf. Appendix A.4).
- The failure to take into account concentration along the fraction of nuclear power axis of the data set (most countries have no nuclear power, hence most of the data set lies exactly at zero with respect to this variable, which is a huge bias of the statistics regarding this variable), which biases the regressions performed (cf. Appendix A.3).
- The interpretation of correlation coefficients as importance measures of the random variables used in the regression analysis (which are correlated) is not justified. While for independent random variables the regression coefficients may be relatively good metrics of importance, this is no longer the case as soon as the variables become correlated. Indeed, even standardizing the correlated set of regression variables, one can obtain results which are misleading. The data set considered presents exactly this problem [11]: we have already pointed out relations between the regression variables considered, which induce correlations between them. If a metric of importance is to be considered, there are appropriate statistical tools to treat the question of the importance of regressed variables in the multicollinear case. We refer the reader to [20–23] and to the appendix of this article (Appendix A.2) for a more in depth description of these methods. The latter have been implemented in `R` as the packages `relaimpo` [21] and `sensitivity` [20]. An in depth analysis of the data using these tools would be enlightening in future works in this direction.

2.3 Other elements previously highlighted by other authors

As previously noted, Sovacool et al.'s paper [11] has been discussed by different authors [12,13]. Let us briefly state some of the main arguments these authors have made regarding this matter.

Fell et al. [12] noted that, among other points,

- Sovacool et al.'s paper [11] does not find a positive correlation between N and emissions, but instead finds a negative correlation, which is non-significant. However, due to the small sample size (30) of nuclear countries, that this turns out not to be significant is not surprising;
- the crowding out hypothesis does not say anything about the ability of nuclear power to avoid emissions;
- the cross-sectional approach of Sovacool et al. with respect to the time frames is also criticized, Fell et al. note that cross-sectional analyses with low sample sizes are sensitive to outliers and sampling choices. Moreover, the choice to study a lagged effect without statistical motivation is criticized by the authors;
- the paper's analysis includes the complete set of countries with low GDP per capita. These countries have low emissions per capita and little to no nuclear power. This choice "appears to establish a weaker correlation between nuclear and low per capita emissions, but likely reveals only that many poorer nations have lower emissions per capita due to greater reliance on agriculture and informal economic activities".

Similarly, Wagner [13] attempted to reproduce the results of Sovacool et al. on different data sets, namely a compilation of datasets of 26 European countries (we refer the reader to ([13], Appendix B) for the detailed sources and reliability controls performed on these databases) and two world-wide databases extracted from the IEA data bank [24]. In his paper, Wagner points out that

- by redoing the study on different publicly avalaible data sets, one fails to reproduce the results of Sovacool et al. Most notably, the fraction of nuclear power in the energy mix correlates negatively with GES emissions and "the analysis of both European and more global data shows that both renewable and nuclear technologies allow a reduction of CO_2 emissions with comparable efficacy";
- "no evidence was found that countries using nuclear power systematically employ more fossil fuels preferentially coal to the extent that the emission-free nature of nuclear energy is not only offset thereby but even overcompensated";
- N and R are anti-correlated, but when performing the study over the smaller set of European countries statistical criteria do not attribute significance to the correlation. As previously noted, given the small sample of countries this lack of significance is not in itself surprising;
- finally, "the regression of the total CO_2 emissions with [the total amount of energy produced, the amount of renewable energy and the amount of nuclear energy (in absolute value)] as independent variables does not provide any new or deeper insight; specifically, it does not [suggest] inner relationships and correlations with cultural and sociological factors as searched for in [11]".

3 Conclusion

The analysis of Sovacool et al. does not back their concluding statements. As demonstrated in this paper and its appendix, all the conclusions of their paper do not follow from the data or from proper statistical treatment of it – in particular, the failure to recognize that the predictive power of their model came from the fraction of fossil fuels in the electrical mix, and to take into account the basic relation between the fraction of renewables and nuclear in the electrical mix is fatal to their conclusions.

Additionally, there are many mistakes in the regression analysis performed and important considerations were not addressed in [11], thus also undermining the validity of their results.

Conflict of interests

The author declares that they have no competing interests to report.

Funding

This research did not receive any specific funding.

Data availability statement

The data as well as the numerical analysis performed is publically available here https://doi.org/10.5281/zenodo.4624748 and is the same as that consired by Sovacool et al. in their paper.

Appendix A Statistical analysis

A.1. Correlation of fractions of the same whole are inherently biased

Let R, N and F be positive random variables (which can be interpreted as fractions), such that the following relation holds

$$R + N + F = 1. \tag{A.1}$$

Linear regression consists in minimizing

$$\mathbb{E}\left[(aN + b - R)^2\right]. \tag{A.2}$$

The minimizer is unique and is exactly equal to the orthogonal projection of R onto the hyperplane spanned by N and the constant random variable 1 in the Hilbert space where these random variables are defined (in the case of this paper, $L^2([0,1])$), which is simply

$$\frac{\mathrm{Cov}(N,R)}{\mathrm{Var}(N)}N + \mathbb{E}[R] - \frac{\mathrm{Cov}(N,R)}{\mathrm{Var}(N)}\mathbb{E}[N], \tag{A.3}$$

from which we deduce that $a = \frac{\mathrm{Cov}(N,R)}{\mathrm{Var}(N)}$. However, the relation between the three variables above immediately implies the following proposition.

Proposition A.1. Suppose that N, R and F are three random variables on $[0,1]$ such that $N+R+F=1$. Then at least two out of the three off-diagonal entries of the covariance matrix are negative, i.e. at least two out of $\mathrm{Cov}(N,R), \mathrm{Cov}(N,F)$ and $\mathrm{Cov}(F,R)$ are negative. Furthermore, the condition to have $\mathrm{Cov}(N,R) > 0$ is that $\mathrm{Var}(F) > \mathrm{Var}(N) + \mathrm{Var}(R)$.

Proof. Suppose otherwise that $\mathrm{Cov}(N,R) > 0$. Take the relation and take the covariance of both sides of equation (1) with respect to N, R and F, respectively, to obtain

$$\mathrm{Cov}(N,R) + \mathrm{Cov}(N,F) = -\mathrm{Var}(N) \tag{A.4}$$
$$\mathrm{Cov}(R,N) + \mathrm{Cov}(R,F) = -\mathrm{Var}(R) \tag{A.5}$$
$$\mathrm{Cov}(F,N) + \mathrm{Cov}(F,R) = -\mathrm{Var}(F). \tag{A.6}$$

By the positivity of the variance, as soon as one of the covariances is positive, the other two are immediately negative. Inverting the above relation, one can write:

$$\mathrm{Cov}(N,R) = \frac{1}{2}(\mathrm{Var}(F) - \mathrm{Var}(N) - \mathrm{Var}(R)) \tag{A.7}$$

which is positive if the condition of the proposition is satisfied. Note also that we retrieve the relationship $\mathrm{Var}(R+N) = \mathrm{Var}(F)$ entailed by relation (1).

This means that, given any situation, one can expect to find negative correlation of fractions of the same whole more than two thirds of the time.

A.2. Some remarks on regression analysis

We will now demonstrate why the use of regression coefficients as an importance metric is misleading for regression variables which are correlated. To do this, let us consider a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, a random variable $Y \in L^2(\Omega)$ which we wish to regress using the set of random variables $\{X_1, \ldots, X_n\} \subset L^2(\Omega)$, which we will assume to be linearly independent, but not necessarily independent (or orthogonal: if $L^2(\Omega)$ is a centered gaussian space these concepts coincide). As previously noted, linear regression in this context is nothing other than the orthogonal projection of Y onto the hyperplane spanned by the so-called explanatory variables $X_1, \ldots, X_n$. We can write

$$Y = \mathrm{Proj}_{\mathrm{Span}\,(X_1,\ldots,X_n)}(Y) + \varepsilon, \tag{A.8}$$

where $\mathrm{Proj}_{\mathcal{H}}(Y)$ denotes the projection of Y onto the vector space $\mathcal{H}$ and ε is the orthogonal component of Y to $\mathrm{Span}(X_1, \ldots, X_n)$. We may express $\mathrm{Proj}_{\mathrm{Span}(X_1,\ldots,X_n)}(Y)$ in the coordinates of the basis $X_1, \ldots, X_n$. This yields the classical regression analysis equation

$$Y = \sum_{k=1}^{n} \beta_k X_k + \varepsilon. \tag{A.9}$$

If X_k are orthogonal in $L^2(\Omega)$, the β_k can be simply expressed in terms of the inner product of Y with X_k,

$$\beta_k = \frac{\langle Y, X_k \rangle}{\langle X_k, X_k \rangle}. \tag{A.10}$$

However, if X_k are not orthogonal (and therefore not independent), the coefficients can be found by virtue of orthogonalizing the $(X_k)_k$ basis to perform the projection and finally changing back the result to $(X_k)_k$ coordinates.

With this said, as depicted in Figure A.1 it is easy to geometrically see why the regression coefficients β_k stemming from correlated variables might be misleading.

From Figure A.1 we see that any metric of "importance" considered for correlated variables should avoid using a coordinate dependent framework. Instead, a coordinate-free approach should be considered. Such a coordinate-free coefficient of importance of variable X_k could be a (weighted) average over all subsets

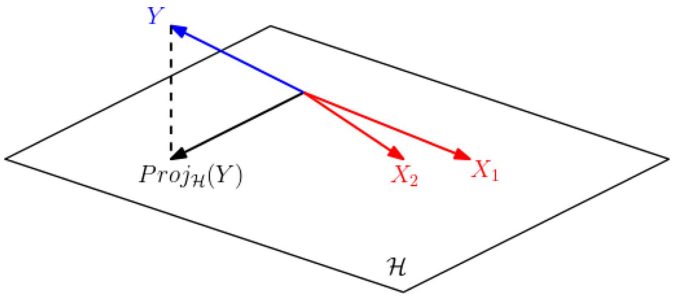

Fig. A.1. Two correlated random variables X_1 and X_2 spanning plane $\mathcal{H}$ in $L^2(\Omega)$. The projection of Y onto plane $\mathcal{H}$ has strictly larger R^2 than the that of the projection of Y onto either X_1 or X_2 alone. However, when expressed in (X_1, X_2) coordinates, the expression of $\mathrm{Proj}_{\mathcal{H}}(Y)$ is misleading, since the coordinates of this projection are large due to the fact variables X_1 and X_2 are almost collinear (in fact they are larger than those of the projection of Y onto X_1 or X_2 individually).

$A \subset \{X_1, \ldots, X_n\} \setminus \{X_k\}$ of ΔR^2_A, the improvement in R^2 (or, up to a sign, difference in variance unexplained) when including variable X_k in a linear model spanned by the variables in A. Geometrically, going back to Figure A.1, this is similar (up to a sign and normalization) to taking the average over all such subsets A of the difference in distance between hyperplane $\mathrm{Span}(A \cup \{X_k\})$ to Y and the hyperplane $\mathrm{Span}(A)$ to Y. This measure of importance is not coordinate dependent, as the measure depends only on the vector spaces spanned by the different variables. As a side note, we remark that this interpretation provides the link between the different formulæ typically used for the so-called LMG. On one hand, notice there are exactly $\binom{n-1}{j}$ hyperplanes spanned by j vectors chosen from a set of $n-1$ vectors. It follows that the average of ΔR^2_A over all possible linear models A not including X_k can be written as

$$\mathrm{LMG}(X_k) := \frac{1}{n} \sum_{j=0}^{n-1} \sum_{A_j} \frac{\Delta R^2_{A_j}}{\binom{n-1}{j}}, \tag{A.11}$$

where the second sum is carried over the subsets $A_j \subset \{X_1, \ldots, X_n\} \setminus \{X_k\}$ having cardinality j. This is the formula for LMG first discovered by Christensen [25]. On the other hand, it is also easy to see that this average can also be written down as

$$LMG(X_k) := \frac{1}{n!} \sum_{A \subset \{X_1, \ldots, X_n\} \setminus \{X_k\}} (\#A)!(n - \#A - 1)! \Delta R^2_A, \tag{A.12}$$

yielding the classical result of the equivalence between all the formulæ from the literature [21].

This is exactly the the approach taken in [20–23], which we suggest should be used to further verify the validity of the conclusions of [11]. These methods should be understood as weighting the contribution of X_2 in reducing the residual variance to linear models spanned by the X_ks. Going back to Figure A.1, it is clear that including X_2 yields non-trivial information, as it considerably increases the variance explained by the linear model, but that this importance is not less than that of X_1. This is not

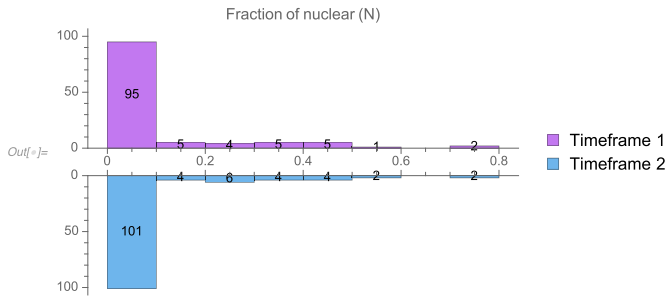

Fig. A.2. Number of countries as a function of N.

as easily seen when considering the (X_1, X_2) coordinates of $\mathrm{Proj}_{\mathcal{H}}(Y)$, but is clear when looking at the geometric layout of the vectors Y, X_1 and X_2. As previously stated, these methods have been implemented in `R` as `relaimpo` and `sensitivity`. Preliminary results suggest that this metric of importance for correlated variables yields different results than that of [11]. It would also be of similar interest to attempt to reproduce the results of Fell et al. [12] and Wagner [13] using this metric.

A.3. Covariances and correlations of N, R and F

If we now let N, R and F denote the fraction of nuclear, renewable and other sources of the electrical production, Proposition A.1. applies. Taking a look at the data from the study, Figures A.2–A.4 show the distributions for each of these variables (the fractions are along the x axis and the y axis is the count of the histogram). Note that F is dominated by fossil fuel contributions. The covariances of these variables can be found in Table A.1 for Timeframe 1 and in Table A.2 for Timeframe 2.

Upon examining the distribution of F, one sees that it is approximately uniform, which should set its variance to be close to $\frac{1}{12}$. By contrast, nuclear power tends to play a small role in the electrical mix of most countries, which tells us that $\mathrm{Var}(N)$ should be negligible with respect to $\mathrm{Var}(R)$ and $\mathrm{Var}(F)$, as the observed values of N concentrate around 0. In particular, this immediately implies that $\mathrm{Cov}(R, F)$ is relatively large in absolute value and negative (independent of interpretation). This is of capital importance when we examine stepwise selection models, which will yield significance for R, but which we will see actually stem from the greater predictive power of variable F of GHG emissions per capita. In particular, the conclusions of Sovacool et al. about the efficacy of renewables to decarbonate do not follow from any statistical analysis, as this covariance is only large and negative because $\mathrm{Var}(N)$ is negligible.

Finally, we must compare $\mathrm{Var}(F)$ with $\mathrm{Var}(R)$. Here, $\mathrm{Var}(R) > \mathrm{Var}(F)$, and so we have the negative correlation between N and R mentioned in the paper. Looking at the distributions of R and F, one finds that this is due to the fact that most countries seem to either focus on renewables (mainly hydroelectric power in the timeframes considered) or not have any at all, whereas the distribution of other (fossil) sources is more or less uniform. This

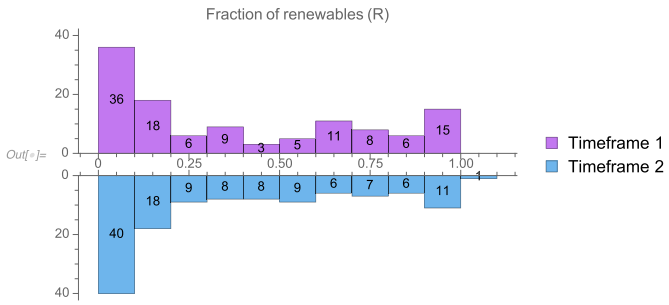

Fig. A.3. Number of countries as a function of R.

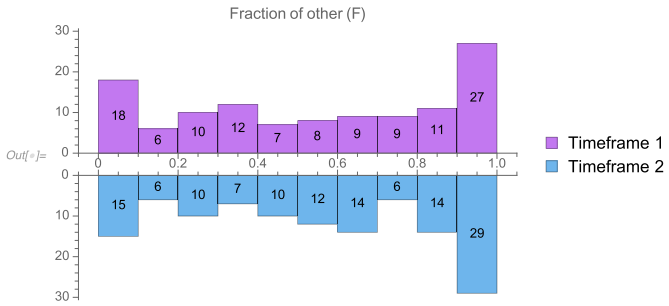

Fig. A.4. Number of countries as a function of the fraction of F.

Table A.1. Covariances between different variables of timeframe 1 for renewable countries (nuclear countries included).

Timeframe 1	N	R	F
N	0.0252887	−0.0161068	−0.00918191
R		0.122027	−0.10592
F			0.115102

negative covariance between N and R is thus explained solely by the latter and the mathematical relation linking the three variables.

The negative nature of the correlations can further be emphasized by examining the rank correlation matrix (here, the Spearman ρ coefficient) between each of the variables. Unsurprisingly, we find that the variables all have negative rank correlation. The exact values of Spearman ρ coefficients between the variables for time frames 1 and 2 are tabulated in Tables A.3 and A.4, respectively.

A.4. Stepwise selection

"Hierachical regression" is more commonly known as stepwise selection in statistics. Stepwise selection can be done in two different directions: forwards or backwards. In forward stepwise selection, one starts with the null model and progressively adds variables while evaluating the significance of each addition, and so, at step n, if variable X_n does not yield a significant improvement in the predictions of the model, this variable is discarded. In backward stepwise selection, the opposite is done. That is, we start with a family of variables, and taking out variables by examining which loss gives the most statistical insignificant detereoration of the model fit. Finally, one can do both steps simultaneously, that is, go backwards and forwards to provide an extra check that the choice of variables is optimal.

Table A.2. Covariances between different variables of timeframe 2 for renewable countries (nuclear countries included).

Timeframe 2	N	R	F
N	0.0252823	−0.0128735	−0.0124088
R		0.107508	−0.0946345
F			0.107043

Table A.3. Spearman ρ between different variables of timeframe 1 for nuclear countries.

Timeframe 1	N	R	F
N	1	−0.336634	−0.580868
R		1	−0.4639
F			1

Table A.4. Spearman ρ between different variables of timeframe 2 for nuclear countries.

Timeframe 2	N	R	F
N	1	−0.23693	−0.682314
R		1	−0.445161
F			1

Beyond this choice of approach, trying to maximize predictive power via improvement of the goodness of fit (R^2) while intending to study causation is wrong. Relying on R^2 alone can induce into error for two main reasons:

- R^2 increases monotonously in the number of parameters added into the model.
- The data span multiple orders of magnitude. This renders small relative variations of the points at large scales to have a considerable effect on the significance of the increase in R^2, despite there being no real meaning behind this significance.

Adapted statistical tools should have been used, such as adjusted R^2 of the fit which takes into account the number of parameters in the model. As for the second point it is more delicate to address so we will do it stepwise, by attempting to reproduce and correct at each step the steps taken in [11]. The data set studied will be that of timeframe 1.

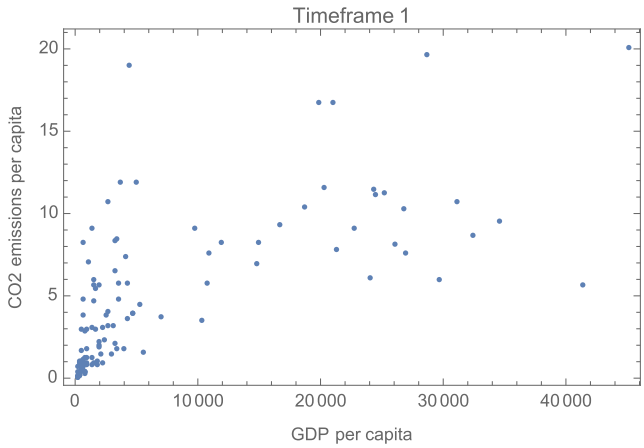

Fig. A.5. CO_2 as a function of GDP in time frame 1 for all countries.

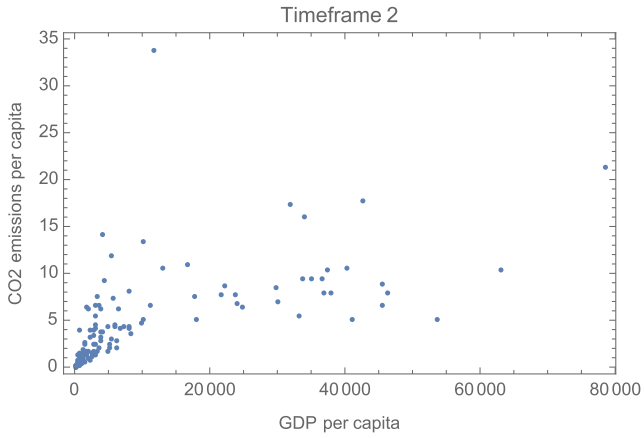

Fig. A.6. CO_2 as a function of GDP in time frame 2 for all countries.

A.4.1. GDP and GHG emissions per capita

Plotting the GDP per capita and the CO_2eq emissions per capita (henceforth denoted GDP and CO_2, respectively, for simplicity) for the countries considered in [11] yields the results in Figures A.5 and A.6.

Following Sovacool et al. the regression model looks like:

$$CO_2 = \beta_0 + \beta_1 \mathrm{GDP}. \tag{A.13}$$

However, after performing a regression analysis in this model, we notice that β_0 is not significative (although we do retrieve their result an R^2 of 0.48 for this model). Applying the principles of bidirectional selection, we exclude β_0 and examine instead:

$$CO_2 = \beta_0 \mathrm{GDP}. \tag{A.14}$$

For timeframe 1, the estimates for the parameters of the model given by the regression are given in Table A.5 and an adjusted R^2 of 0.64, a result which already rivals the (non-adjusted) R^2 they obtain at the end of their forward selection (0.66).

Remark A.1. The reported P-values are grossly underestimated, since the underlying distribution of the residuals is not exactly normal as shown by the Kolmogorov–Smirnov test. A more reliable statistic in this setting is the t-statistic and the standard error. We report nonetheless the P-value for the sake of completeness.

Table A.5. Regression results for model A.14.

	Estimate	Standard error	t-statistic	P-value
β_0	0.0000931758	8.90×10^{-6}	10.4719	1.99×10^{-18}

However, this data set can spans multiple orders of magnitude and is very clearly heteroskedastic in both timeframes, which means that the typical assumptions behind linear regression are not at all satisfied. Failing to take this into account in a linear regression – and particularly one where the data cover such large magnitudes – is catastrophic, as one can have significant increases in R^2 without this reflecting anything other than a couple of points with high GDP getting closer to the regression plane.

Heteroskedastic data spanning many orders of magnitude are often a sign of an underlying Pareto distribution (or power law). This hypothesis can be checked by looking at the data on a $\log - \log$ plot (cf. Figs. A.7 and A.8). By inspection we can see that this hypothesis seems to be confirmed.

The simplest model we can postulate is given by

$$\log CO_2 = \beta_0 + \beta_1 \log \mathrm{GDP}. \tag{A.15}$$

This simple model has an adjusted R^2 of rougly 0.69 and the following regression table for timeframe 1. The full regression results of this model can be found in Table A.6

This shows that poor a priori inspection of the data from the part of the authors of [11] ultimately led to a suboptimal model. In particular, we notice that this adjusted R^2 is already higher than any of the R^2 values obtained by the authors at the end of their forward selection (0.66), despite being penalized for taking into account the number of variables in the model and only having two predictors.

Since the goal of this paper is an attempt to reproduce the results of Sovacool et al. we will keep model of equation A.14 in what will follow, despite the fact that going forwards we should consider accounting for the confounding variable with a power law and not just a linear model.

From Figures A.7 and A.8, it seems clear that a saturation phenomenon occurs and that we enter a different regime as GDP becomes larger. We could think here of postulating a more complicated non-linear regression model to account for this phenomenon (for instance by imposing a quadratic regression, or non-parametric smoother model), taking into account that more complex importance measures for should be then applied [20]. This is, however, beyond the scope of this paper.

A.4.2. Nuclear, renewables, GDP and CO_2eq emissions

We discard the N variable after performing bidirectional selection, as the variable does not prove to be significant or

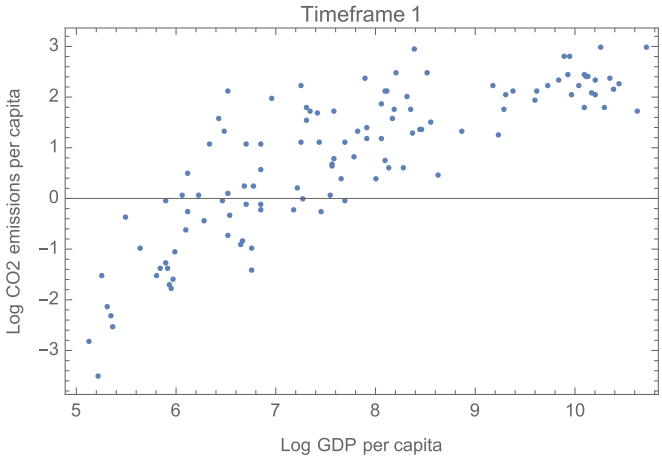

Fig. A.7. CO_2 as a function of log(GDP) in time frame 1 for all countries.

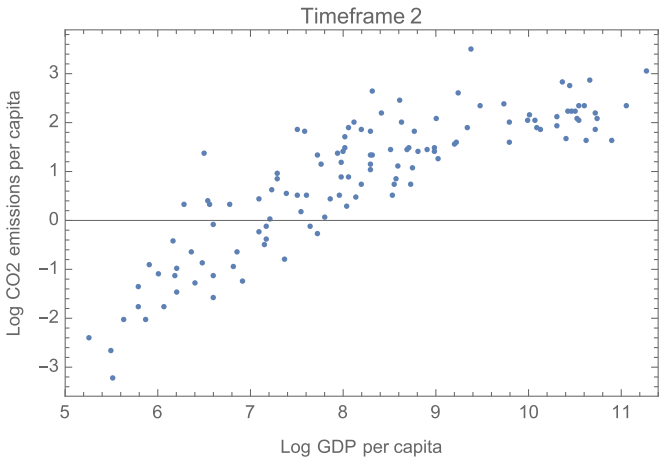

Fig. A.8. CO_2 as a function of log(GDP) in time frame 2 for all countries.

to provide considerable improvement to the adjusted R^2. Other than the obvious reason that nuclear power emits little to no GHGs, there are other explanations of why this is not a significative explanatory variable in our model. The addition of N as a variable only affects 30 of the data points, many of which lie close to 0% nuclear energy, which does not add much information to the model (around half of them are below the 20% mark). On timeframe 2, one can speculate that there are two trendlines, one before the 30% mark, which is increasing, and the other afterwards, which decreases. Of course, this may purely be an artefact of the data given the low sampling. The data have been depicted in Figures A.9 and A.10.

Still following Sovacool et al., let us now look at what happens when we add variable R into the model of equation A.14, which becomes:

$$CO_2 = \beta_0 \mathrm{GDP} + \beta_1 R. \tag{A.16}$$

We can look at the regression analysis of this model, whose details are given in Table A.7.

The adjusted R^2 value for this iteration of the model is 0.64, which does not improve the previous model.

Table A.6. Regression results for model A.15.

	Estimate	Standard error	t-statistic	P-value
β_0	−5.2705	0.389961	−13.5154	1.99×10^{-25}
β_1	0.781987	0.0493198	15.8554	1.31×10^{-30}

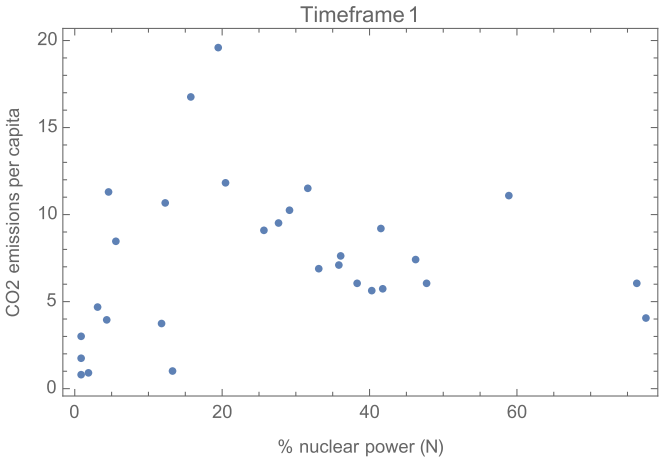

Fig. A.9. CO_2 as a function of N for the nuclear countries in time frame 1.

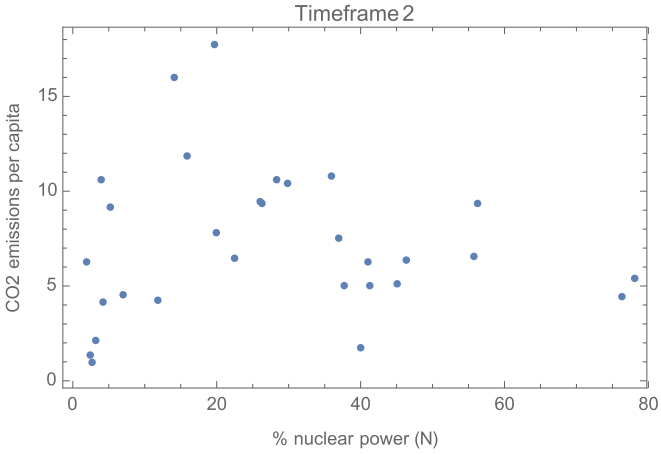

Fig. A.10. CO_2 as a function of N for the nuclear countries in time frame 2.

Furthermore, β_1 is not deemed significatively different from 0, meaning that R plays no role in predicting the GHG emissions per capita. This is of course, obvious from the fact that renewable energy emits little to no GHGs.

By contrast, F mostly carries information about the fraction of fossil fuels in the electrical mix, since other sources of energy are negligible once we have excluded fossil fuels, renewables and nuclear power. It follows that a more reasonable model is simply

$$CO_2 = \beta_0 + \beta_1 \mathrm{GDP} + \beta_2 F. \tag{A.17}$$

As before, parameter β_0 was found to be non-significant. Following the principles of bidirectional selection, we

Table A.7. Regression results for model A.16.

	Estimate	Standard error	t-statistic	P-value
β_0	0.000414812	0.0000307419	13.4934	2.24×10^{-25}
β_1	0.00635995	0.00720374	0.882869	0.379165

Table A.8. Regression results for model A.18.

	Estimate	Standard error	t-statistic	P-value
β_1	0.000314618	0.000025023	12.5731	2.84×10^{-23}
β_2	0.0447815	0.00485628	9.22136	1.81×10^{-15}

Table A.9. Regression results for model A.19.

Model A.19	Estimate	Standard error	t-statistic	P-value
β_0	4.41294	0.47065	9.37627	8.47×10^{-16}
β_1	0.000288171	0.0000268078	10.7495	5.43×10^{-19}
β_2	−0.0468907	0.00785489	−5.96962	2.80×10^{-8}

Table A.10. Regression results for model A.20.

Model A.20	Estimate	Standard error	t-statistic	P-value
β_0	−0.083787	0.546209	−0.153397	0.878359
β_1	0.000316026	0.0000267566	11.8112	1.87×10^{-21}
β_2	0.0457721	0.00809241	5.65617	1.18×10^{-7}

exclude β_0, and instead consider:

$$\mathrm{CO_2} = \beta_1 \mathrm{GDP} + \beta_2 F, \tag{A.18}$$

whose regression table can be found in Table A.8.

Both variables are significant predictors and the adjusted R^2 of this model is 0.80, and the standard error of the predictors decreased.

Finally, let us show that the predictive power of R in Sovacool et al.'s suboptimal model came from F. To do this, we compare their model

$$\mathrm{CO_2} = \beta_0 + \beta_1 \mathrm{GDP} + \beta_2 R, \tag{A.19}$$

to the following (also suboptimal) model

$$\mathrm{CO_2} = \beta_0 + \beta_1 \mathrm{GDP} + \beta_2 F. \tag{A.20}$$

These models have respective regression tables given in Tables A.9 and A.10.

There are couple of things to note. First, the suboptimality of Model A.20 is reflected by the fact that β_0 is evidently not significant. More importantly, β_2 is almost exactly the same in absolute value as it was in the previous model. This, in conjunction with the large anti-correlation between F and R allows us to conclude that the predictive power of R in Model A.19 was in fact inherited from that of F. Of course, there is a tautological causal link behind this correlation given that F mostly consists of the fraction of fossil fuels in the electrical mix.

References

1. J. Rogelj, D. Shindell, S.F.K. Jiang, P. Forster, V. Ginzburg, C. Handa, H. Kheshgi, S. Kobayashi, E. Kriegler, L. Mundaca, R. Séférian, M. Vilariño, Mitigation pathways compatible with 1.5°C in the context of sustainable development, Global Warming of 1.5°C. *An IPCC Special Report on the Impacts of Global Warming of 1.5° C above Pre-industrial Levels and Related Global Greenhouse Gas Emission Pathways, in the Context of Strengthening the Global Response to the Threat of Climate Change, Sustainable Development, and Efforts to Eradicate Poverty* (2018)
2. S.H. Kim, K. Wada, A. Kurosawa, M. Roberts, Nuclear energy response in the EMF27 study, Clim. Change **123**, 443 (2014)
3. J. Rogelj, A. Popp, K.V. Calvin, G. Luderer, J. Emmerling, D. Gernaat, S. Fujimori, J. Strefler, T. Hasegawa, G. Marangoni, V. Krey, E. Kriegler, K. Riahi, D.P. van Vuuren, J. Doelman, L. Drouet, J. Edmonds, O. Fricko, M. Harmsen, P. Havlík, F. Humpenöder, E. Stehfest, M. Tavoni, Scenarios towards limiting global mean temperature increase below 1.5°C, Nat. Clim. Change **8**, 325 (2018)
4. F. Creutzig, P. Agoston, J.C. Goldschmidt, G. Luderer, G. Nemet, R.C. Pietzcker, The underestimated potential of solar energy to mitigate climate change, Nat. Energy **2**, 17140 (2017)
5. M.Z. Jacobson, M.A. Delucchi, Z.A. Bauer, S.C. Goodman, W.E. Chapman, M.A. Cameron, C. Bozonnat, L. Chobadi, H.A. Clonts, P. Enevoldsen, J.R. Erwin, S.N. Fobi, O.K. Goldstrom, E.M. Hennessy, J. Liu, J. Lo, C.B. Meyer, S.B. Morris, K.R. Moy, P.L. O'Neill, I. Petkov, S. Redfern, R. Schucker, M.A. Sontag, J. Wang, E. Weiner, A.S. Yachanin, 100% clean and renewable wind, water, and sunlight all-sector energy roadmaps for 139 countries of the world, Joule **1**, 108 (2017)
6. C. Clack, S. Qvist, J. Apt, M. Bazilian, A. Brandt, K. Caldeira, S. Davis, V. Diakov, M. Handschy, P. Hines, P. Jaramillo, D. Kammen, J. Long, M. Morgan, A. Reed, V. Sivaram, J. Sweeney, G. Tynan, D. Victor, J. Weyant, J. Whitacre, Evaluation of a proposal for reliable low-cost grid power with 100% wind, water, and solar, Proc. Natl. Acad. Sci. **114**, 6722 (2017)
7. S. Hong, C.J. Bradshaw, B.W. Brook, Global zero-carbon energy pathways using viable mixes of nuclear and renewables, Appl. Energy **143**, 451 (2015)
8. A. Berger, T. Blees, F.M. Breon, B.W. Brook, M. Deffrennes, B. Durand, P. Hansen, E. Huffer, R.B. Grover, C. Guet, W. Liu, F. Livet, H. Nifenecker, M. Petit, G. Pierre, H. Prévot, S. Richet, H. Safa, M. Salvatores, M. Schneeberger, B. Wornan, S. Zhou, Nuclear energy and bio energy carbon capture and storage, keys for obtaining 1.5°C mean surface temperature limit, Int. J. Glob. Energy Issues **40**, 240 (2017)
9. A. Berger, T. Blees, F.M. Bréon, B.W. Brook, P. Hansen, R.B. Grover, C. Guet, W. Liu, F. Livet, H. Nifenecker, M. Petit, G. Pierre, H. Prévot, S. Richet, H. Safa, M. Salvatores, M. Schneeberger, S. Zhou, How much can nuclear energy do about global warming? Int. J. Glob. Energy Issues **40**, 43 (2017)
10. X.-J. Xiao, K.-J. Jiang, China's nuclear power under the global 1.5°C target: Preliminary feasibility study and prospects, Adv. Clim. Change Res. **9**, 138 (2018)

11. B.K. Sovacool, P. Schmid, A. Stirling, G. Walter, G. MacKerron, Differences in carbon emissions reduction between countries pursuing renewable electricity versus nuclear power, Nat. Energy **5**, 928 (2020)
12. H. Fell, A. Gilbert, J. Jenkins, M. Mildenberger, Reply to 'Differences in carbon emissions reduction between countries pursuing renewable electricity versus nuclear power', by Sovacool et al. (2020), SSRN Electron. J. (2021) https://ssrn.com/abstract=3762762 or http://dx.doi.org/10.2139/ssrn.3762762
13. F. Wagner, CO2 emissions of nuclear power and renewable energies: A statistical analysis of european and global data, Eur. Phys. J. Plus **136**, 562 (2021)
14. *Conditions and Requirements for the Technical Feasibility of a Power System with a High Share of Renewables in France Towards 2050*, Technical report (IEA, Paris, 2021)
15. *BP Statistical Review of World Energy*, Technical report (BP, London 2019)
16. P. Bloomfield, *Fourier Analysis of Time Series: An Introduction* (Wiley, New York, 1976)
17. R.H. Shumway, *Applied Statistical Time Series Analysis* (Prentice Hall, Englewood Cliffs, NJ, 1988)
18. J. Hamilton, *Time Series Analysis* (Princeton University Press, Princeton, 1994)
19. I. Pardoe, *Multiple Linear Regression* (John Wiley & Sons, Ltd, Hoboken, NJ, 2012)
20. S. Da Veiga, F. Gamboa, B. Iooss, P. Clémentine, *Basics and trends in sensitivity analysis theory and practice in R*, (SIAM, 2021)
21. U. Grömping, Relative importance for linear regression in R: The package relaimpo, J. Stat. Softw. **17**, 1 (2006)
22. U. Grömping, Estimators of relative importance in linear regression based on variance decomposition, Am. Stat. **61**, 139 (2007)
23. J.W. Johnson, J.M. Lebreton, History and use of relative importance indices in organizational research, Organ. Res. Meth. **7**, 238 (2004)
24. *Data and Statistics*, Technical report (IEA, Paris 2020)
25. R. Christensen, Comment on Chevan and Sutherland, Am. Stat. **46**, 70 (1992)

Cite this article as: Daniel Perez. An attempt of reproduction of Sovacool et al.'s "Differences in carbon emissions reduction between countries pursuing renewable electricity versus nuclear power", EPJ Nuclear Sci. Technol. **8**, 24 (2022)