Sustainable Production and Consumption 30 (2022) 596–607

Contents lists available at ScienceDirect

Sustainable Production and Consumption

journal homepage: www.elsevier.com/locate/spc

Automatic modeling of socioeconomic drivers of energy consumption and pollution using Bayesian symbolic regression

Daniel Vázquez [a], Roger Guimerà [b,c], Marta Sales-Pardo [b,*], Gonzalo Guillén-Gosálbez [a,*]

[a] Department of Chemistry and Applied Biosciences, Institute for Chemical and Bioengineering, ETH Zurich, Vladimir-Prelog-Weg 1, 8093 Zurich, Switzerland
[b] Department of Chemical Engineering, Universitat Rovira i Virgili, Tarragona 43007, Catalonia, Spain
[c] ICREA, Barcelona 08010, Catalonia, Spain

ARTICLE INFO

Article history:
Received 22 October 2021
Revised 1 December 2021
Accepted 23 December 2021
Available online 27 December 2021

Editor: Prof. Ignacio Grossmann

Keywords:
Surrogate model
Symbolic regression
Stochastic impacts by regression on population
Affluence and technology (STIRPAT)
Greenhouse gas (GHG) emissions
Eora environmentally extended multi-region input-output database

ABSTRACT

Predicting countries' energy consumption and pollution levels precisely from socioeconomic drivers will be essential to support sustainable policy-making in an effective manner. Current predictive models, like the widely used STIRPAT equation, are based on rigid mathematical expressions that assume constant elasticities. Using a Bayesian approach to symbolic regression, here we explore a vast amount of suitable mathematical expressions to model the link between energy-related impacts and socioeconomic drivers. We find closed-form analytical expressions that outperform the well-established STIRPAT equation and whose mathematical structure challenges the assumption of constant elasticities adopted in the literature. Our work unfolds new avenues to apply machine learning algorithms to derive analytical expressions from data in environmental studies, which could help find better models and solutions in energy-related problems.

1. Introduction

Today's high living standards rely on the global trade, production, and use of resources, which results in a range of environmental impacts that could challenge the stability of the Earth system. Notably, energy consumption reached 583.9 EJ in 2020 (BP, 2020), leading to high levels of anthropogenic CO_2 emissions (31.5 Gt (IEA, 2021)) due to the heavy reliance on fossil energy resources. The COVID-19 pandemic has reduced these consumption levels, yet the expected increase in fossil fuels consumption for 2021 would reverse practically all the reductions achieved in 2020 (IEA, 2021). There is, therefore, a clear need to curb emissions by shifting to more sustainable energy sources and reducing the current high energy demand. This will require deepening our knowledge on the specific forces driving these energy-related environmental stressors so more effective policies can be developed.

The term "driver" refers to a range of factors that can explain the environmental footprint of a geographical or political unit. Several analytical tools were developed to identify impact drivers based on closed-form analytical expressions linking impacts to driving forces. Such formulas are calibrated with data and then used to predict how changes in the drivers affect changes in the impact values.

In a very influential work, Ehrlich and Holdren (1971) proposed the IPAT identity, which links the environmental impact with the population, affluence , and implemented technological levels. Impact here may refer to CO_2 emissions or any other pressure exerted on the environment. The identity assumes a multiplicative effect of the drivers that influence the impact, leading to the expression, $I = \mathrm{PAT}$, which gives name to the identity. The IPAT identity has been mainly used as an accounting equation to compute the implemented technology level from a known population, affluence, and environmental impact values.

A caveat of the IPAT identity is that it cannot account for non-monotonic or non-proportional effects of the driving forces, which motivated the development of an alternative methodology. Notably, York et al. (2003) put forward a stochastic version of the IPAT, termed STIRPAT, which relates impacts to drivers using the concept of ecological elasticity. This model allows for a more in-depth analysis of the impact drivers, providing a quantitative framework to relate changes in inputs (drivers) to changes in outputs (impacts)

* Corresponding authors.
E-mail addresses: marta.sales@urv.cat (M. Sales-Pardo), gonzalo.guillen.gosalbez@chem.ethz.ch (G. Guillén-Gosálbez).

https://doi.org/10.1016/j.spc.2021.12.025

Nomenclature

Abbreviations & sets

ANN	Artificial neural network
BMS	Bayesian machine scientist
GHG	Greenhouse gas
J	{j: Set of drivers to study}
LASSO	Least absolute shrinkage and selection operator
MCMC	Markov chain Monte Carlo
MINLP	Mixed-integer nonlinear programming
ML	Machine learning
N	{n: Set of data points}
SVM	Support vector machine

Variables & parameters

AP	Active population. Percentage of population with age comprehended between 15 and 64 years (%)
AT	Average temperature (K)
BIC	Bayesian information criterion
CDE	CO_2 emissions (kg CO_2)
CVE	Cross-validation error
DP	Population density. Number of habitants per square kilometer (inhabitants/km^2)
EC	Energy consumption (TJ)
GDP	Gross domestic product per capita (2020$/inhabitant)
ME	CH_4 emissions (kg CH_4)
MSE	Mean squared error
NOE	N_2O emissions (kg N_2O)
R^2	Coefficient of determination
TP	Total population (inhabitants)
UR	Urbanization rate. Percentage of population that lives in an urban area (%)
k	Number of parameters in the surrogate model
$\bar{y}$	Mean of the observed values
y_n	Observed value for data point n
$\hat{y}_n^{Model}$	Predicted value using the surrogate model for data point n

based on a linear relationship in logarithmic space. The STIRPAT model has been widely used in the literature to study multiple environmental burdens, ranging from CO_2 (Fan et al., 2006; Libao et al., 2017; Pao et al., 2012; Zhu et al., 2020), N_2O and CH_4 emissions (Le and Nguyen, 2020), and other greenhouse gas (GHG) emissions (Anser, 2019; Chekouri et al., 2020; Nosheen et al., 2020; Singh and Mukherjee, 2019), to water footprint (Zhao et al., 2014), water pollution (Zhang et al., 2017), phosphorus footprint (Jiang et al., 2019), ecological footprint (Yang et al., 2021), and energy consumption (Ibrahim et al., 2017; Lin et al., 2020; Ma et al., 2013). For example, these models showed that affluence and population are the main drivers of CO_2 emissions (York et al., 2003).

In recent years, the widespread use of machine learning (ML) in many scientific fields has unfolded new avenues for extracting knowledge from data by building predictive models. The STIRPAT equation is an example of an empirical relationship – linking anthropogenic drivers to environmental impacts–, which was likely found by manual procedures (York et al., 2003). By contrast, here we are interested in generating plausible mathematical expressions describing impact drivers in an automated manner without assuming any ad hoc fixed canonical formalism.

ML algorithms such as artificial neural networks (ANN) and Gaussian processes have recently become prevalent regression tools (Jean et al., 2016; Lee et al., 2018; Schweidtmann and Mitsos, 2019). Support vector machines (SVM) have already been used to predict CO_2 emissions (Saleh et al., 2016) from electricity consumption and the available technology in Indonesia, while random forest algorithms were applied to predict agricultural N_2O emissions (Saha et al., 2021) from intensively managed cropping systems using ammonium, nitrate and clay content in the soil as predictive variables. More classical ML methodologies, such as Least Absolute Shrinkage and Selection Operator (LASSO) regression and feed-forward ANN, have also been applied to predict environmental impacts. Hamrani et al. (2020) compared different ML approaches to predict GHG emissions from agricultural soils, using various inputs such as the air and ground temperature and wind speed. Hempel et al. (2020) predicted methane emissions of a dairy building with natural ventilation in Northern Germany using ANN, linear regression, and other ML methodologies and considering as predictive variables (or features) the temperature, wind speed, and direction. Romeiko et al. (2020) applied Gradient Boosting Regression Trees, ANN, and SVM to predict global warming and eutrophication impacts of corn production from the temperature, precipitation, soil organic content, soil texture, application rates of both nitrogen and phosphorus fertilizers, and farming practices. Nguyen et al. (2021) used random forests to create land use maps, which are useful to monitor and assess land-use change. Zhang et al. (2021) used an ANN to enhance the prediction of energy demand and its price during the COVID-19 pandemic.

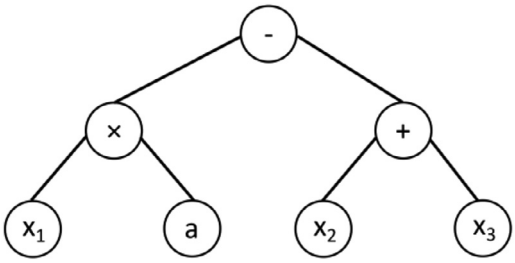

Fig. 1. Representation of the equation $f(x_1, x_2, x_3) = ax_1 - (x_2 + x_3)$ using a binary tree.

The ML regression models above assume a fixed mathematical structure encoded by a sequence of hyper-parameters –e.g., the number of neurons per layer and the number of layers in an ANN or the number of explanatory variables in a multi-linear regression– which control the learning process and are fixed beforehand. These approaches then seek the model parameters that provide the best fit to observed data. Similar to these ML approaches, the STIRPAT model (York et al., 2003) also relies on a fixed mathematical structure, i.e., a log-linear equation. However, assuming such a limited pre-defined structure may limit our ability to explain observed data precisely.

The problem of identifying governing principles and equations from data has recently emerged as an active area of research in machine learning. In mathematical terms, the symbolic regression problem (also called sparse regression and equation discovery) aims at finding both the structure of a model and its parameters from a set of observations. Symbolic regression is often addressed by representing mathematical expressions using a symbolic tree, as shown in Fig. 1. In these trees, leaf nodes correspond to either a variable or a constant, while all parent nodes correspond to mathematical operators. An algorithm can then be coupled with the symbolic tree representation to navigate through the space of feasible expressions and identify the one that best fits the data.

Recent years have witnessed an increasing interest in symbolic regression (Cozad and Sahinidis, 2018; Neumann et al., 2020). Unlike other ML methods, symbolic regression has the advantage of providing an explicit closed-form equation that can be manipulated algebraically, differentiated, and analyzed more in-depth. ANNs also lead to mathematical expressions, yet they combine ac-

tivation functions and weights in a single formalism that is hard to interpret.

Various symbolic regression approaches have been put forward typically with the sole aim of finding a unique best mathematical representation of the data. These approaches mainly differ in whether the optimization approach used to explore the space of symbolic trees is stochastic or deterministic. Stochastic methods, such as those using genetic algorithms to explore the search space (Žegklitz and Pošík, 2021), tend to be faster and easier to implement, yet they cannot guarantee convergence to the global optimum within an epsilon tolerance unless the algorithm is run for an infinite time. By contrast, deterministic methods often require more significant CPU times but can provide a bound on the minimum error that could be attained with the best model. More recently, deterministic optimization was applied to the symbolic regression problem by formulating a mixed-integer nonlinear program (MINLP) (Cozad and Sahinidis, 2018). This MINLP can be solved to local optimality by standard MINLP methods , such as the nonlinear branch and bound (Dakin, 1965) and outer approximation algorithms (Duran and Grossmann, 1986), or to global optimality using deterministic global optimization algorithms.

In a recent article, some of us have developed an alternative approach to symbolic regression that is rooted in probability theory and allows computing the plausibility of each mathematical expression given the observed data, which we call the Bayesian Machine Scientist (BMS) (Guimerà et al., 2020a). The BMS uses a stochastic Markov chain Monte Carlo (MCMC) algorithm to explore the space of possible mathematical expressions according to their plausibility in an ergodic manner. The plausibility of each model is obtained as the description length of the model and the data, which can be computed from the Bayesian information criterion (*BIC*) and the logarithm of the prior over the expressions considered. The prior, which controls for the structural complexity of the mathematical model and therefore acts as a structural regularizer, was obtained from a corpus of 4080 mathematical expressions found in Wikipedia. The BMS was shown to outperform state-of-the-art symbolic regression approaches as well as standard ML regression models, such as Gaussian processes, at fitting data and predicting out-of-sample data.

To the best of our knowledge, symbolic regression has never been applied to identify energy consumption and pollution drivers, which are still studied mainly using the STIRPAT method based on a fixed, known mathematical structure. To go beyond these approaches, here we use for the first time a Bayesian symbolic regression approach (i.e., the BMS) to identify environmental impact drivers from a set of socioeconomic descriptors. We find that the BMS leads to expressions that outperform the STIRPAT equation at explaining the variance of data. Moreover, our results challenge the assumption of constant elasticities, widely adopted in STIRPAT models, by which a constant increment in one driver always results in a constant change in impact. On the downside, the symbolic regression algorithm provides more complex analytical expressions (compared to the STIRPAT model), although the level of complexity can still be controlled through an appropriate tuning of the algorithm

The paper is structured as follows. First, we present the problem statement and the methodology. Then, we introduce the case study. Finally, we discuss the results and draw some conclusions.

2. Problem statement

Our goal is to find a mathematical expression to predict environmental impacts from a set of socioeconomic drivers to be identified from a pool of variables. Hence, the data available comprise a series of potential drivers and environmental impacts of a set of countries at different years. These input data, i.e., independent variables, have dimensions $x \times y$, where x is the number of pairs country-year studied and y the number of potential drivers. The output data, i.e., dependent variables, has dimensions $x \times z$, where z is the number of environmental impacts studied. We clarify here that impact can refer to (i) a direct damage to the environment; (ii) a set of emissions ultimately responsible for an impact, or (iii) the energy consumption level, ultimately leading to emissions and, hence, to impacts. The objective is to study the influence of socio-economical drivers on environmental impacts. For example, we might be interested in predicting the CO_2 emissions of a country from the population, affluence, and other socioeconomic descriptors.

3. Methodology

Fig. 2 shows a schematic representation of the overall methodology. We train two surrogate models, the BMS and the STIRPAT, to predict four environmental impacts using six socioeconomic drivers. For the obtained models, we then compute the elasticities, that is, the percentage change of the output in response to a percentage change in the inputs.

3.1. Data

We obtained data from the World Data Bank and the EORA database (Lenzen et al., 2013). We pre-processed the data from EORA to use consumption-based impacts in the calculations (rather than production-based), as discussed in Appendix A.

We study the effects of six environmental drivers, i.e., total population, GDP per capita, active population, population density, urbanization rate, and climate, on four environmental impact categories: CO_2 emissions(*CDE*), energy consumption(*EC*), N_2O emissions(*NOE*), and CH_4 emissions(*ME*) at the country level, with data spanning 25 years, from 1990 to 2015. Energy consumption includes power, fuels, and heating. CO_2 emissions are, therefore, strongly connected to energy consumption, as energy sources are mostly fossil-based. CH_4 is the second most important GHG emission in terms of global anthropogenic contribution to climate change (Marmier and Schosger, 2020). Approximately 36.9% of the world's anthropogenic CH_4 emissions are linked to natural gas and petroleum energy systems (IEA, 2020). For completeness, we also consider the third most important GHG emission, N_2O. Although N_2O is mainly emitted in agriculture, approximately 23.3% of the anthropogenic N_2O emissions are linked to the burning of biomass and to fossil fuels use and industrial processes (Tian et al., 2020).

Fig. 3 shows a breakdown of GHG emissions depending on the source for 2018, as presented in PRIMAP-hist (Gütschow et al., 2016).

Following the notation in EORA, CO_2 emissions correspond to the entry "Total CO_2 emissions from EDGAR", in kg. We obtain the energy consumption by adding the following terms: "Energy Usage, Coal", "Energy Usage, Natural Gas", "Energy Usage, Petroleum", "Energy Usage, Nuclear Electricity", "Energy Usage, Hydroelectric Electricity", "Energy Usage, Geothermal Electricity", "Energy Usage, Wind Electricity", "Energy Usage, Solar, Tide and Wave Electricity", and "Energy Usage, Biomass and Waste Electricity", in TJ. N_2O emissions correspond to the category "GHG emissions from PRIMAPHIST, N_2O, total excluding LULUCF" in kg. Finally, CH_4 emissions correspond to the entry "GHG emissions from PRIMAPHIST, CH_4, total excluding LULUCF".

Due to data gaps, we consider only the pairs country-year with complete information for all the drivers studied. Overall, our data set comprises 4183 data points for *CDE*, 4184 data points for *EC*, and 4185 points for both *NOE* and *ME*.

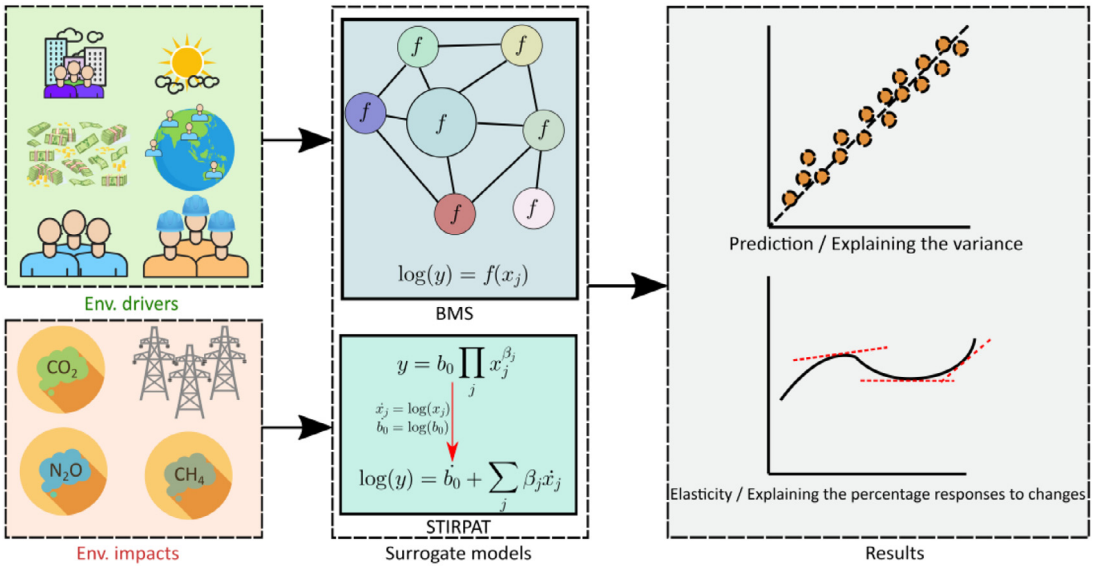

Fig. 2. Schematic representation of the methodology. We obtained two models for each impact using the Bayesian Machine Scientist (BMS) and the STIRPAT. For each of them, we compute a set of fitness metrics and the elasticities of the drivers.

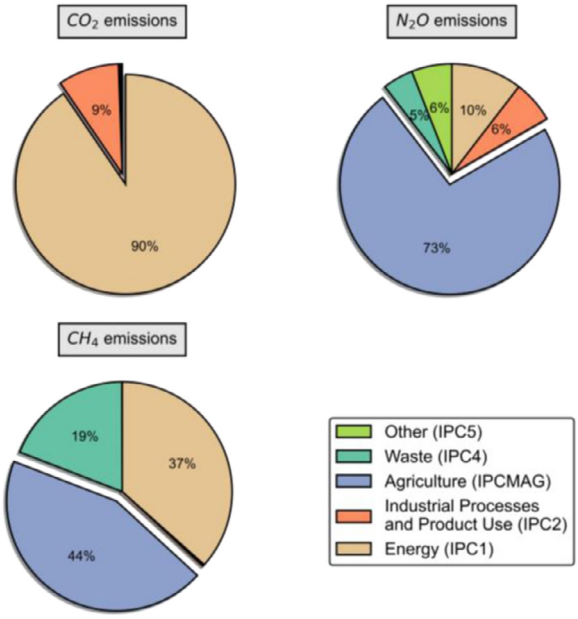

Fig. 3. Source of the greenhouse gas (GHG) emissions per sector in 2018 from PRIMAP-hist (Gütschow et al., 2016).

3.2. Comparison metrics

We estimate environmental impacts from six drivers using two distinct regression models, f_1' and f_2':

$$\begin{aligned} f_1'(x_1,\ldots,x_d)+e_1 &= y \\ f_2'(x_1,\ldots,x_d)+e_2 &= y \end{aligned} \tag{1}$$

where $x_1,\ldots,x_d$ denote the drivers thought to influence the studied variables, and y refers to the observed environmental impact. The error terms are denoted by e_1 and e_2. As surrogate models, we use the well-known STIRPAT (York et al., 2003), and a model obtained using the BMS (Guimerà et al., 2020a).

The STIRPAT and BMS models are trained to minimize the error and description length, respectively. The description length combines the Bayesian information criterion (*BIC*), which accounts for the error and the structural complexity of the surrogate model, with a prior. Details on how the description length is calculated can be found elsewhere (Guimerà et al., 2020a).

We define the predicted values for each model as Eq. (2)

$$\begin{aligned} f_{1n}'(x_1,\ldots,x_d) &= \hat{y}_n^{STIRPAT} \quad \forall n \in N \\ f_{2n}'(x_1,\ldots,x_d) &= \hat{y}_n^{BMS} \quad \forall n \in N \end{aligned} \tag{2}$$

where N is the set of data points indexed by n.

We evaluate the models in two ways. First, we evaluate the ability of the models to describe observed data. To that end, we compute three different metrics: the amount of variance explained (R^2), the mean squared error (*MSE*), and the *BIC*, as defined by Eq. (3), Eq. (4), and Eq. (5), respectively.

$$R^2 = 1 - \frac{\sum_n \left(y_n - \hat{y}_n^{Model}\right)^2}{\sum_n \left(y_n - \bar{y}\right)^2} \tag{3}$$

$$MSE = \frac{\sum_n \left(y_n - \hat{y}_n^{Model}\right)^2}{|N|} \tag{4}$$

$$BIC = k \cdot \log\left(|N|\right) + |N| \left(\log\left(2\pi\right) + \log\left(\frac{\sum_n \left(y_n - \hat{y}_n^{Model}\right)^2}{|N|} \right) + 1 \right) \tag{5}$$

Where y_n stands for the observed value for each data point, $\hat{y}_n^{Model}$ stands for the predicted value for each data point determined using either the STIRPAT or the BMS model, $\bar{y}$ stands for the mean of the observed environmental impacts, and k stands for the number of parameters in the surrogate model plus one.

Second, because looking at goodness of fit metrics favors complex models that overfit, we also evaluate the cross-validation error (*CVE*). We specifically follow a leave-one-country-out approach, where we first obtain the structure of the model (i.e., the mathematical dependency of the impact on the drivers) with all the points available. Afterward, we subdivide the data into subsets, or folds, using all the countries but one as the training set. The left-out country acts as the test set in each run. We repeat this procedure for each country, calculating the cross-validation error as the

average error in the test set of each fold, weighted considering the size of each test set.

With the models obtained from all the data, we study the elasticity of the drivers, a concept adopted from economics by the industrial ecology community to quantify the intensity of the links between impacts and drivers. Notably, elasticity refers to the proportional change in a dependent variable in response to a change in an independent variable, maintaining the other independent variables and parameters constant. As an illustrative example, for a general equation $y = f(x)$, the elasticity of the dependent variable y to a change in the independent variable x is given by Eq. 6.

$$E^x = \frac{\partial y}{\partial x}\frac{x}{y} \tag{6}$$

3.3. STIRPAT approach

The STIRPAT model is defined in Eq. 7.

$$I = aP^{\alpha}A^{\beta}T^{\gamma} \tag{7}$$

where P stands for the total population, A stands for affluence, often expressed as per capita gross product (GDP), and T stands for the technology factor, while I denotes the impact. This model can be further refined by reformulating T as the product of a set of drivers $j \in J$ with values D_j, resulting in the term T'^{γ}. The modified STIRPAT model we use in this work is shown in Eq. 8.

$$\begin{aligned} I &= aP^{\alpha}A^{\beta}T'^{\gamma} \\ T'^{\gamma} &= \prod_{j\in J} D_j^{\omega_j} \end{aligned} \tag{8}$$

where a is a constant, P^{α} refers to the contribution of the population, A^{β} refers to the contribution of the affluence, D_j refers to a set of additional drivers, and ω_j denotes the exponent (elasticity) of each driver. It is common practice to use an additive regression model where the variables are in logarithmic form, as shown in Eq. 9.

$$\log(I) = \log(a) + \alpha\log(P) + \beta\log(A) + \sum_{j\in J}\omega_j \log\left(D_j\right) \tag{9}$$

The main advantage of the STIRPAT methodology is that the constant elasticities are directly obtained from the values of the coefficients α, β, ω_i. A coefficient with a value of 0 indicates that the driver does not affect the impact. A coefficient with a value between 0 and 1 indicates an inelastic relationship, where the impact increases less than the increase in the driving force. A value of 1 indicates a unit elastic relationship, where changes in the impact are directly proportional to changes in the driver. A value above 1 indicates an elastic relationship, where an increase in the driver produces an even higher increase in the impact. These elasticities can be negative, in which case the direction of the proportionality is inverted, i.e., a negative elastic relationship means that the impact decreases at a greater proportion than the increase in the driver.

3.4. BMS approach

The BMS identifies models for observed data by navigating through the space of symbolic trees using a Markov chain Monte Carlo algorithm. The BMS assumes the following general mathematical dependency between the logarithm of the impact and socioeconomic drivers (Eq. (10)).

$$\log(I) = f\left(P, A, D_j\right) \tag{10}$$

We consider the same independent variables we use for the STIRPAT model: the population, the GDP per capita, and the additional drivers. However, we do not consider the logarithm of these independent variables but rather their original values. This is because the BMS has the freedom to explore the logarithmic transformation during the search over the space of mathematical expressions. The BMS generates new models by sampling over the distribution of the plausible mathematical models given the data. This plausibility can be computed in terms of the description length ($\mathcal{L}$), which we can approximate as shown in Eq. 11.

$$\mathcal{L} \approx \frac{BIC}{2} - \log(POE) \tag{11}$$

where POE is the prior over the mathematical expressions. Note that the BMS sampling is not an optimization algorithm, but because it is ergodic, it ensures that the best model can be found. From the sampling over models, we select the model with the shortest description length.

Table 1
Performance metrics for the different environmental impacts using STIRPAT and BMS.

Env. impact	Metric	STIRPAT	BMS
CDE	R^2	0.858	0.869
	MSE	0.677	0.626
	BIC	10,296	9996
	CVE	0.726	0.668
EC	R^2	0.861	0.870
	MSE	0.627	0.588
	BIC	9976	9715
	CVE	0.662	0.611
NOE	R^2	0.809	0.817
	MSE	0.649	0.620
	BIC	10,123	9915
	CVE	0.698	0.639
ME	R^2	0.809	0.825
	MSE	0.625	0.572
	BIC	9968	9605
	CVE	0.673	0.609

Table 2
Coefficients of the STIRPAT equation.

Env. Impact	Variable	Coefficient
CDE	Intercept	4.748**
	log(*GDP*)	0.566***
	log(*TP*)	0.914***
	log(*AP*)	2.741***
	log(*DP*)	−0.052***
	log(*UR*)	0.484***
	log(*AT*)	−2.318***
EC	Intercept	18.405***
	log(*GDP*)	0.460***
	log(*TP*)	0.911***
	log(*AP*)	2.309***
	log(*DP*)	−0.033***
	log(*UR*)	0.541***
	log(*AT*)	−6.211***
NOE	Intercept	20.299***
	log(*GDP*)	0.378***
	log(*TP*)	0.940***
	log(*AP*)	−1.656***
	log(*DP*)	−0.145***
	log(*UR*)	−0.052*
	log(*AT*)	−2.542***
ME	Intercept	−32.862***
	log(*GDP*)	0.302***
	log(*TP*)	0.931***
	log(*AP*)	0.540***
	log(*DP*)	−0.217***
	log(*UR*)	0.149***
	log(*AT*)	5.900***

* $p < 0.20$.
** $p < 0.15$.
*** $p < 0.01$.

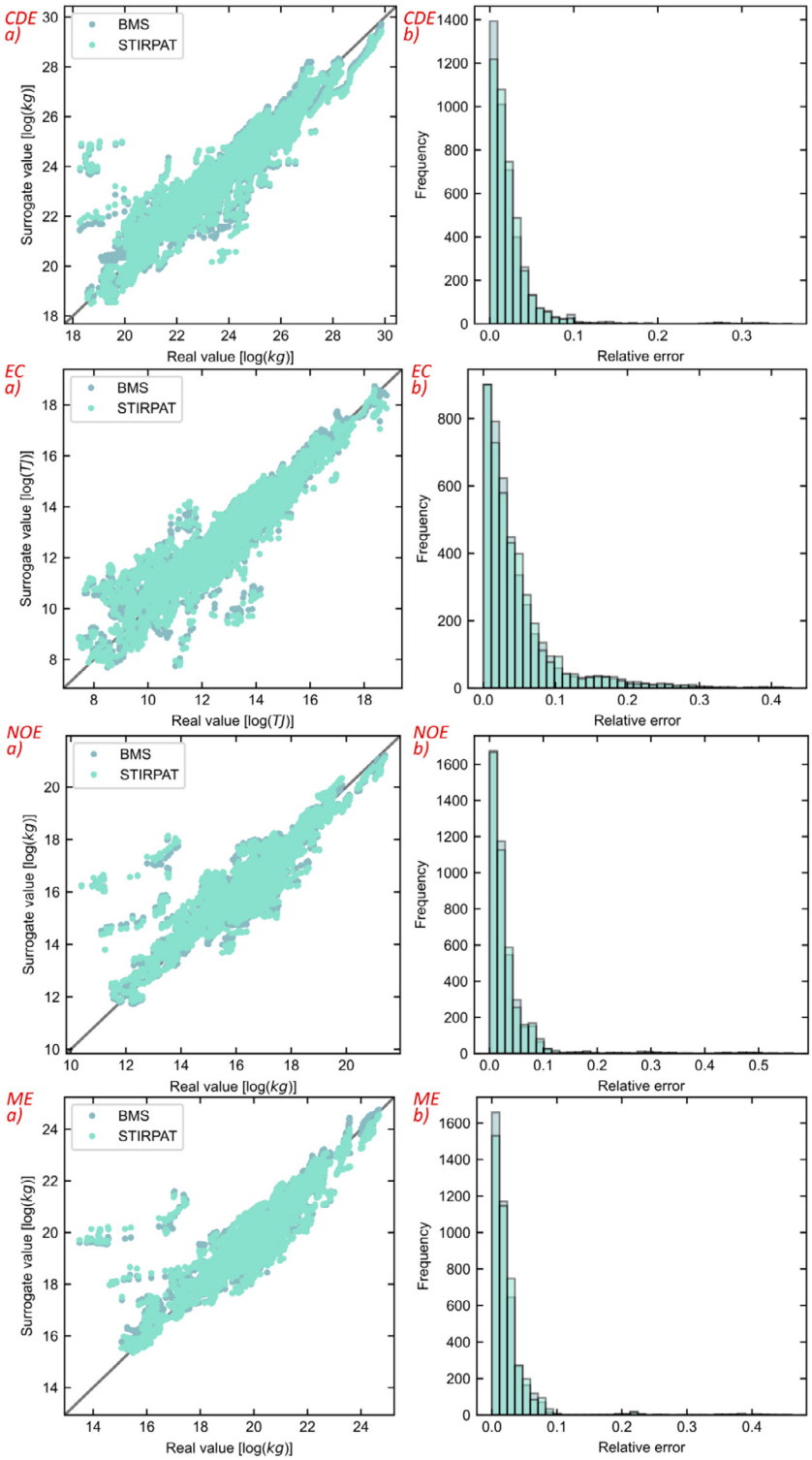

Fig. 4. Full data results for the four environmental impacts: a) Predicted value vs. real value, b) histogram of absolute relative errors for the STIRPAT and the Bayesian Machine Scientist (BMS) in the logarithmic space. The notation for the impact categories is as follows: *CDE* refers to CO_2 emissions, *EC* refers to energy consumption, *NOE* refers to N_2O emissions and *ME* refers to methane emissions.

Note that to compare with the STIRPAT approach, the BMS looks for models predicting the logarithm of the impact, but it is not a necessary requirement.

The STIRPAT model assumes constant elasticities. By contrast, the BMS does not. We can obtain the elasticities of the model found with the BMS using a symbolic differentiation algorithm, like the SymPy library of Python or the symbolic toolbox of MATLAB R2020a.

3.5. Chosen drivers

As potential drivers, we consider the affluence and population, often the main driving forces of anthropogenic impacts, expressed in GDP per capita (2020$/inhabitant) (The World Bank, 2020a) (*GDP*) and total population (inhabitants) (The World Bank, 2020b) (*TP*), respectively. Besides, based on the literature

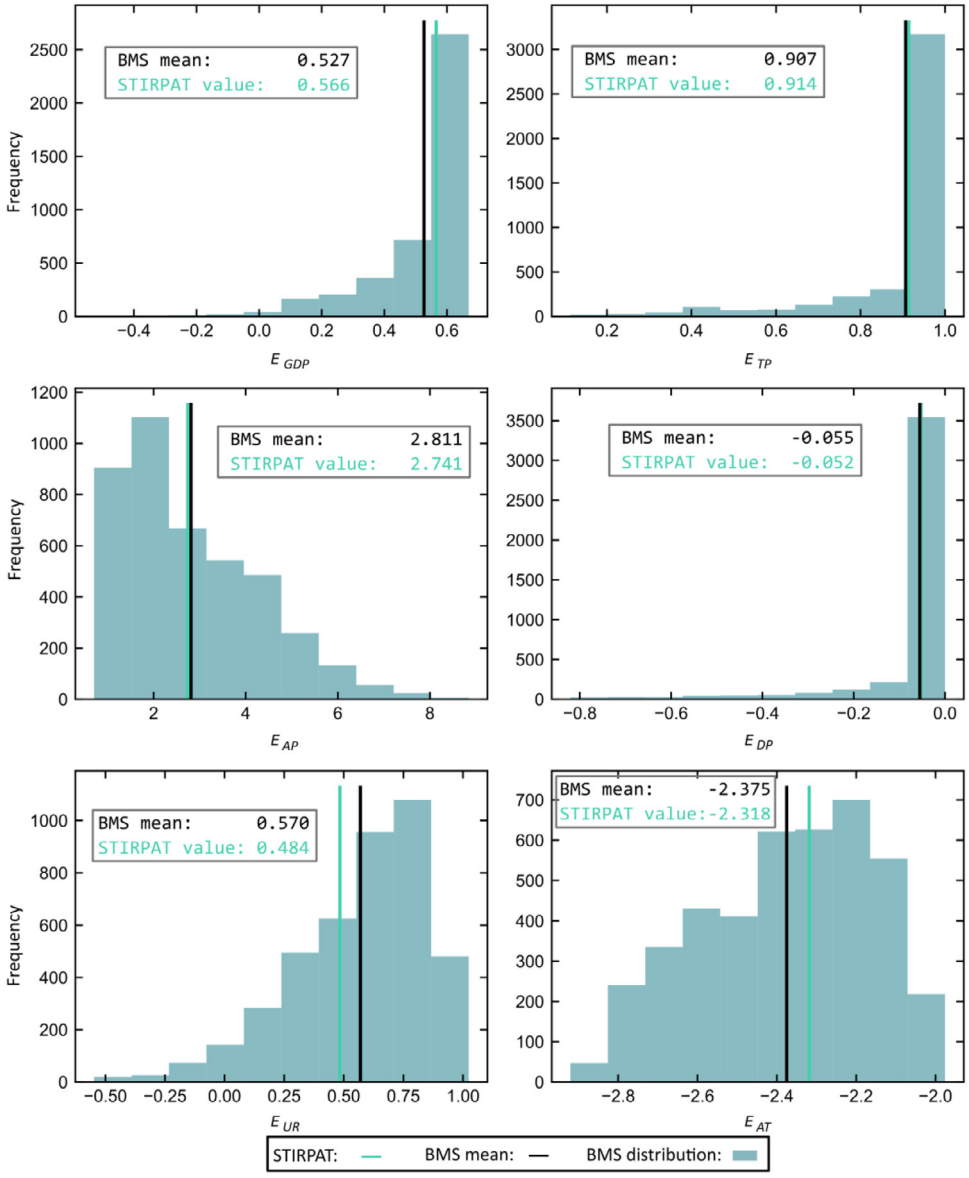

Fig. 5. Elasticities of models of CO_2 emissions. E_x refers to the elasticity with respect to driver x, where $x = GDP, TP$ (total population), AP (active population), DP (density of population), UR (urbanization rate), and AT (average temperature). The histogram shows the distribution of the elasticities obtained from the Bayesian Machine Scientist (BMS) model for each data point. Vertical lines correspond to the STIRPAT elasticity (turquoise) and the mean elasticity obtained from the BMS elasticity distribution (black).

(Teixidó-Figueras et al., 2016), we include the following four additional (potential) drivers:

- The active population (AP): percentage of people with ages comprehended between 15 and 64 years (source: World Bank Group (The World Bank, 2019a)).
- Population density (DP): Inhabitants per square kilometer (source: World Bank Group (The World Bank, 2019b)).
- Urbanization rate (UR): Percentage of population that lives in an urban area (source: World Bank Group (The World Bank, 2018)).
- Climate (AT): We use the average temperature registered in the country as a proxy for this driver (source: World Bank Climate Portal (The World Bank, 2021)).

Note that the STIRPAT equation includes all these drivers in the mathematical expression, while the BMS may or may not include them depending on the structure of the model showing the best description length. Hence, the BMS tackles the feature selection problem implicitly, i.e., which drivers are statistically relevant, during the search for the best model.

We clarify that the potential drivers here analyzed were defined based on the literature (Teixidó-Figueras et al., 2016; York et al., 2003) and considering as well data availability. We stress that the methodology is general enough to work with other drivers.

4. Results and discussion

We implemented the STIRPAT model in the Python 3.8 library statsmodels 0.12.2. We run the BMS model using the original BMS code provided in the repository (Guimerà et al., 2020b) together with Pandas 1.2.3 and SymPy 1.7.1. For the numerical calculations, we used NumPy 1.20.1. We ran 5000 MCMC steps of the BMS algorithm, requiring ca. five hours of CPU time, depending on the environmental impact. The STIRPAT model took less than 10 s for all cases. We performed all the calculations on an Intel Core i9–9900 CPU @ 3.10 GHz. We clarify that there is no single work covering

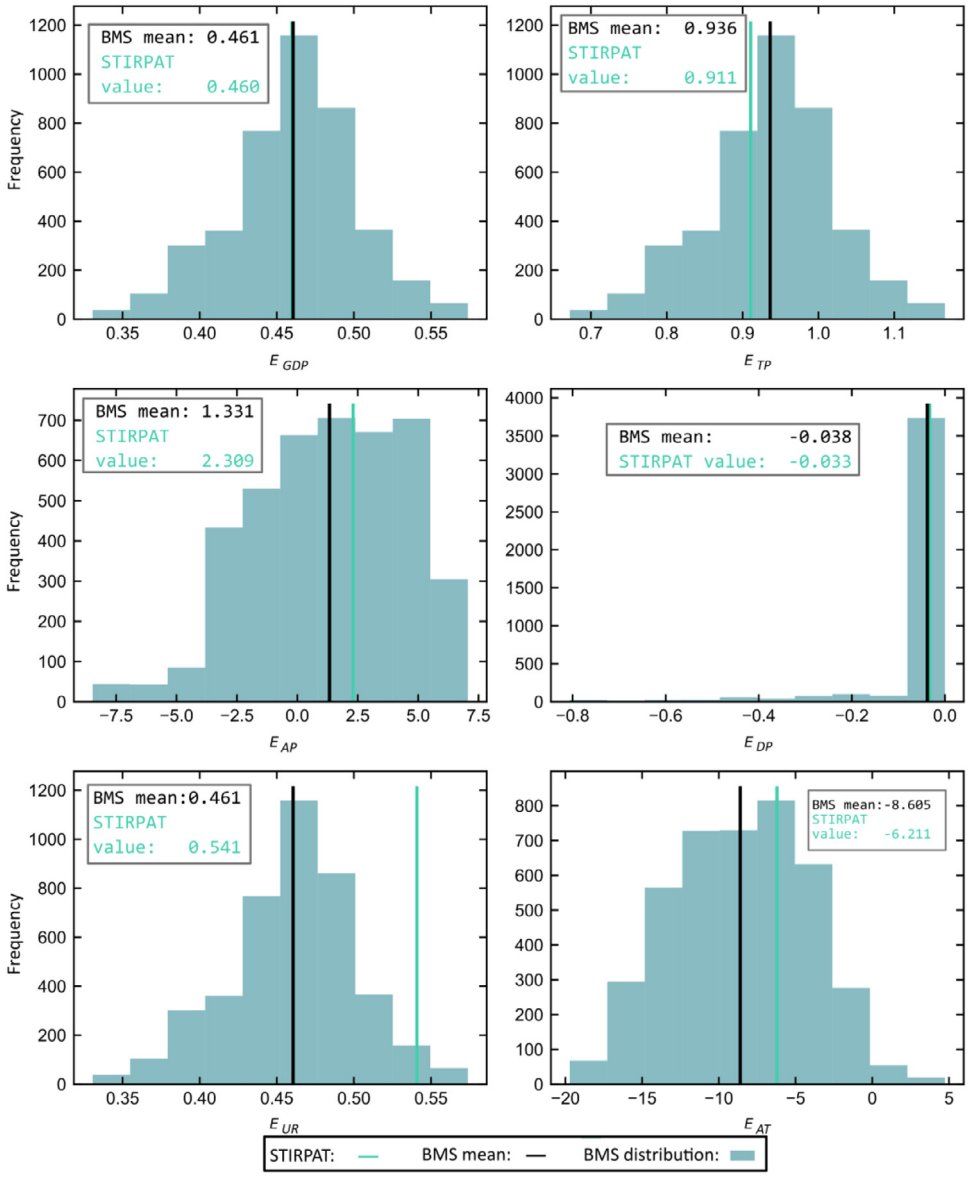

Fig. 6. Elasticities of the models of energy consumption. E_x refers to the elasticity with respect to driver x, where $x = GDP, TP$ (total population), AP (active population), DP (density of population), UR (urbanization rate), and AT (average temperature). The histogram shows the distribution of the elasticities obtained from the Bayesian Machine Scientist (BMS) model for each data point. Vertical lines correspond to the STIRPAT elasticity (turquoise) and the mean elasticity obtained from the BMS elasticity distribution (black).

the same range of impacts and drivers studied here, yet some investigated a subset of them. Hence, in what follows, we focus on assessing the performance of the BMS and compare the insight obtained with that generated in other works, whenever possible.

4.1. Performance metrics

Table 1 shows the performance metrics for the STIRPAT and BMS obtained using all the data. The STIRPAT provides good fits but is consistently outperformed by the BMS in all the performance metrics. This superior performance was expected since, theoretically, the STIRPAT equation is included in the search space of the BMS. However, finding a better model requires a higher computational cost (hours vs. seconds).

Fig. 4 shows the observed and predicted values and the histogram of absolute relative errors for both approaches. The two surrogate models perform similarly in the logarithmic space. However, both models produce some estimates that deviate significantly from the diagonal. Moreover, the centroid of the histogram of absolute relative errors is closer to zero in the BMS, highlighting its better predictive capabilities.

4.2. Structure and feature selection

Table 2 shows the STIRPAT parameters found for the different impacts.

On the other hand, the four models selected by the BMS are shown below, with the corresponding parameters being displayed in Table 3:

$$\log(CDE) = \frac{a_1 GDP + AP^{a_2} \log\left(\frac{(a_3 UR \cdot AT+1)^3 (a_4 GDP^{a_5} DP \cdot UR (TP+a_6)+a_7)}{DP \cdot UR \cdot AT^6}\right) - a_8}{AP^{a_2}}$$

$$\log(EC) = a_6 a_7 \left(a_4 + AP \left(a_7^2 GDP \cdot UR \left(\frac{a_5^{AP}}{TP} \right)^{a_2} \right)^{\frac{a_1 + a_3^{DP}}{AT}} + AP \right)$$

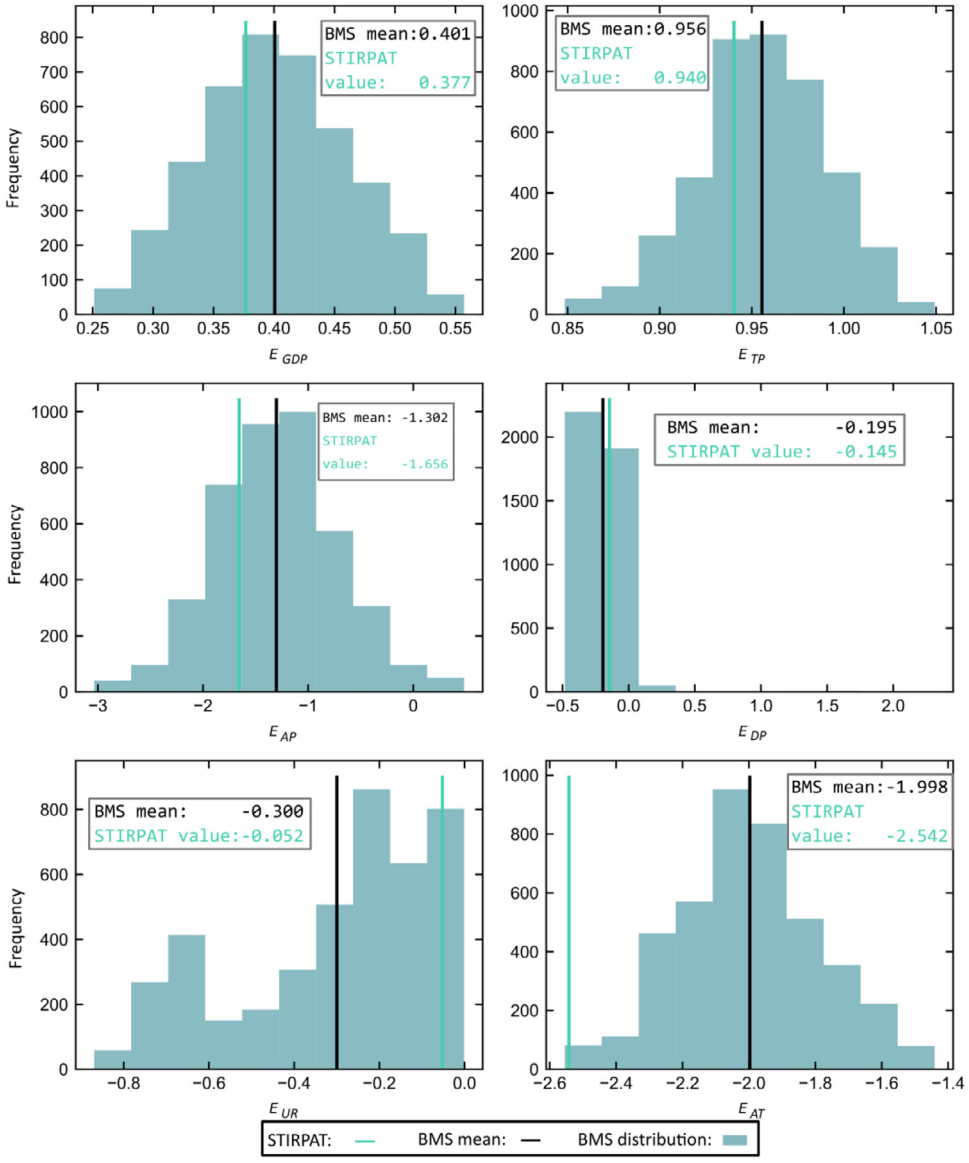

Fig. 7. Elasticities of the models of N_2O emissions. E_x refers to the elasticity with respect to driver x, where $x = GDP$, TP (total population), AP (active population), DP (density of population), UR (urbanization rate), and AT (average temperature). The histogram shows the distribution of the elasticities obtained from the Bayesian Machine Scientist (BMS) model for each data point. Vertical lines correspond to the STIRPAT elasticity (turquoise) and the mean elasticity obtained from the BMS elasticity distribution (black).

Table 3
Values of the parameters of the BMS models for the different environmental impacts.

Parameter	$log(CDE)$	$log(EC)$	$log(NOE)$	$log(ME)$
a_1	-1.200×10^{5}	5.804	90.005	-9.654×10^{-26}
a_2	5.503	−2.033	1.000	−0.278
a_3	1.720×10^{-5}	0.594	1.598	−21.508
a_4	1.808×10^{8}	−100.810	0.999	-2.858×10^{-2}
a_5	0.673	1.934	–	91.108
a_6	9.997×10^{4}	1.429×10^{-13}	–	-1.954×10^{3}
a_7	4.896×10^{18}	1.834×10^{12}	–	−11.293
a_8	2.186×10^{9}	–	–	–

Table 4
Feature selection. For each environmental impact, we show whether a driver appears in the model (X) or not (O) using STIRPAT|BMS approaches. For the STIRPAT we consider drivers with a p-value < 0.1. A green cell means that both models choose the driver, while an orange cell indicates a discrepancy between the models.

Env. Impact	GDP	TP	AP	DP	UR	AT
CDE	X | X	X | X	X | X	X | X	X | X	X | X
EC	X | X	X | X	X | X	X | X	X | X	X | X
NOE	X | X	X | X	X | X	X | X	O | X	X | X
ME	X | X	X | X	X | O	X | X	X | O	X | X

$$\log(NOE) = a_4 + a_4^{AT}\left(\frac{DP}{GDP} + \frac{UR^2}{a_1 GDP}\right)^{-\frac{a_3}{AP}} \log\left(a_2^{DP} TP\right)$$

$$\log(ME) = a_7 + \frac{a_3}{\frac{a_4 a_5}{GDP} - TP^{a_4}} + \frac{DP^{\frac{a_4^2 a_6}{DP}}}{a_1 GDP(-a_4 AT)^{-a_3} + a_2}$$

As seen, both the STIRPAT and the BMS provide a good fit, yet the BMS finds more complex expressions. Table 4 summarizes the drivers chosen by each surrogate model assuming a significance value of 0.1 for the p-values in the STIRPAT equation. Both approaches tend to consider all the drivers for all impacts with only two exceptions where at least one driver is omitted in at least one

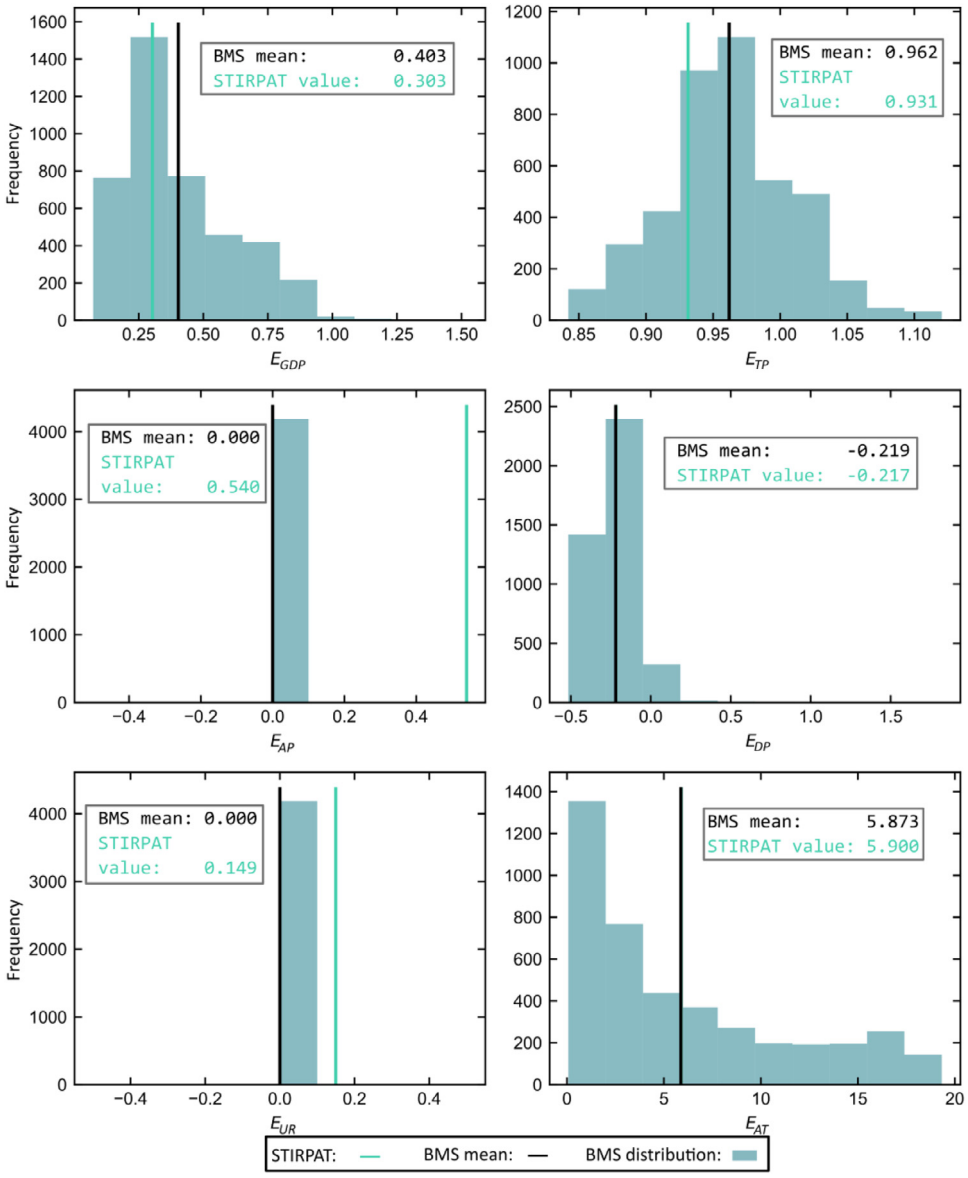

Fig. 8. Elasticities of the models of CH_4 emissions. E_x refers to the elasticity with respect to driver x, where $x = GDP$, TP (total population), AP (active population), DP (density of population), UR (urbanization rate), and AT (average temperature). The histogram shows the distribution of the elasticities obtained from the Bayesian Machine Scientist (BMS) model for each data point. Vertical lines correspond to the STIRPAT elasticity (turquoise) and the mean elasticity obtained from the BMS elasticity distribution (black).

of the approaches. For instance, the urban rate is not a significant driver for NOE in the STIRPAT approach and is not a driver for ME according to the BMS model. Similarly, the active population is not considered as a necessary driver for ME according to the BMS model, yet it is included in the STIRPAT.

4.3. Elasticities

Having observed that both methods tend to lead to similar combinations of drivers, we next analyze the intensity of the link drivers-impact using elasticities. In the STIRPAT model, the elasticity is constant due to its specific canonical form that simplifies the calculations, while it is variable and dependent on the drivers' values in all the BMS expressions found above.

Figs. 5-8 show the histogram of elasticities for the BMS models for all the impacts we consider. The elasticity has been evaluated in each data point, considering the values of the drivers in that particular observation. The figures also display the constant elasticity obtained from the STIRPAT equation. Notably, we found that for all environmental impacts, the *GDP* and *TP* mean (BMS) and constant elasticities (STIRPAT) lay between zero and one (inelastic positive relationship), meaning that an increase in these drivers tends to result in an increase of lower magnitude in all the environmental impacts. By contrast, looking again at the mean and constant elasticities, *DP* always shows a negative inelastic relationship. The mean and constant elasticities of the other drivers take positive or negative values depending on the environmental impact. In most cases, the mean elasticities for the BMS models lie close to the constant elasticities of the STIRPAT equation, except for the CH_4 emissions model (Table 5), where two drivers (*AP* and *UR*) are omitted by the BMS but kept by the STIRPAT. Although the mean elasticities from the BMS are close to those reported by the STIRPAT, they show a high variability depending on the data points where they are calculated, even shifting their sign from negative to positive. For example, the elasticity of the *AP* driver in energy consumption (Fig. 6) ranges from −8 to 7, challenging the assumption of constant elasticities in the STIRPAT approach. For the same impact, i.e., energy consumption, the AT driver presents even higher

Table 5
Elasticity summary. The pairs are ordered as STIRPAT| mean of BMS. P denotes that the elasticity is positive, N denotes that the elasticity is negative. E denotes elastic, while I denotes inelastic. O denotes that the elasticity is zero, and therefore, perfectly inelastic. In order to compare our result to previous works, we refer to three main references, "R1" (York et al., 2003), "R2" (Teixidó-Figueras et al., 2016) and "R3" (Le and Nguyen, 2020). Considering the relationship (positive or negative) between pairs driver – impact, bold states that our results match previous findings, while italics mean that the results do not match previous findings. A "–" indicates that no previous work that references the pair driver–impact was found.

Env. Impact	*GDP*	*TP*	*AP*	*DP*	*UR*	*AT*
CDE	PI|PI	PI|PI	PE|PE	NI|NI	PI|PI	NE|NE
	R1, R2	**R1**	**R1, R2**	**R2**	**R1, R2**	**R1, R2**
EC	PI|PI	PI|PI	PE|PE	NI|NI	PI|PI	NE|NE
	R1	**R1**	**R1**	–	**R1**	**R1**
NOE	PI|PI	PI|PI	NE|NE	NI|NI	NI|NI	NE|NE
	R3	–	–	–	*R3*	–
ME	PI|PI	PI|PI	PI|O	NI|NI	PI|O	PE|PE
	R3	–	–	–	**R3**	–

variance, with values ranging from −19 to 5. Most of the values are negative, in contrast to the AP driver where negative and positive values are more balanced. Overall, these results show that using symbolic, derivable expressions in this context challenges the understanding of how the drivers affect the environmental impacts in different countries and periods.

In general, the signs of the elasticities -indicating how a driver qualitatively affects the impact- are consistent with those found in previous works for the same drivers and impacts (Teixidó-Figueras et al., 2016; York et al., 2003), with only some exceptions (Le and Nguyen, 2020). Notably, of the 24 tuples driver-impact studied, we find the same qualitative relationship in 13 of them and two discrepancies, while the remaining nine cases were, as far as the authors are aware, never investigated before.

A more in-depth discussion of Table 5 and the comparison with existing works is available in Appendix B.

5. Conclusions

In this work, we applied Bayesian symbolic regression to build predictive models of energy consumption and pollution from a set of socioeconomic variables that could potentially act as impact drivers. We investigated six drivers using the standard STIRPAT methodology and a Bayesian learning algorithm (BMS) that automatically builds analytical expressions from data.

Using a data set encompassing 168 countries and spanning 25 years (+4180 data points in general per impact), we found that the BMS outperforms the STIRPAT approach in all the cases and under all the fitness metrics investigated. In terms of findings, although we are unaware of any study with a similar breadth (four impacts, six drivers, and +4180 data points per impact) and depth (analysis of any plausible mathematical expressions, rather than a fixed one, using symbolic regression), our findings seem to be consistent with those fragmented in the literature –often based on much fewer observations–. However, our approach challenges the use of constant elasticities, a widespread assumption in the literature. Notably, the predictions made by the widely used STIRPAT model can be outperformed by using other canonical expressions that do not rely on constant elasticities. The average elasticities found by the BMS tend to be close to the constant values provided by the STIRPAT. Yet, the BMS elasticities can take extreme values that differ substantially from the STIRPAT solution. Hence, the assumption of constant elasticities for the different drivers, often adopted in this type of studies, might prevent us from finding better equations leading to lower errors

Overall, this work paves the way for advanced ML methods based on symbolic regression to model how socioeconomic drivers impact energy consumption and pollution. By deriving analytical expressions from data, practitioners will be able to generate additional insight and perform in-depth analyses more efficiently. Future work should focus on customizing the method to this problem, defining better statistical criteria to guide the search based on a corpus of equations used in socioeconomic studies.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

R.G. and M.S.-P. acknowledge funding from MCIN/AEI/10.13039/501100011033 award FIS2016-78904-C3-P-1 and by the Government of Catalonia (2017SGR-896).

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.spc.2021.12.025.

References

Anser, M.K., 2019. Impact of energy consumption and human activities on carbon emissions in Pakistan: application of STIRPAT model. Environ. Sci. Pollut. Res. 26, 13453–13463. doi:10.1007/s11356-019-04859-y.
BP, 2020. Statistical Review of World Energy 66.
Chekouri, S.M., Chibi, A., Benbouziane, M., 2020. Examining the driving factors of CO_2 emissions using the STIRPAT model: the case of Algeria. Int. J. Sustain. Energy 39, 927–940. doi:10.1080/14786451.2020.1770758.
Cozad, A., Sahinidis, N.V., 2018. A global MINLP approach to symbolic regression. Math. Program. 170, 97–119. doi:10.1007/s10107-018-1289-x.
Dakin, R.J., 1965. A tree-search algorithm for mixed integer programming problems. Comput. J. 8, 250–255. doi:10.1093/comjnl/8.3.250.
Duran, M.A., Grossmann, I.E., 1986. An outer-approximation algorithm for a class of mixed-integer nonlinear programs. Math. Program. 36, 307–339. doi:10.1007/BF02592064.
Ehrlich, P.R., Holdren, J.P., 1971. Impact of population growth. Science (80-.) 171, 1212–1217. doi:10.1126/science.171.3977.1212.
Fan, Y., Liu, L.-.C., Wu, G., Wei, .Y.-M., 2006. Analyzing impact factors of CO_2 emissions using the STIRPAT model. Environ. Impact Assess. Rev. 26, 377–395. doi:10.1016/j.eiar.2005.11.007.
Guimerà, R., Reichardt, I., Aguilar-Mogas, A., Massucci, F.A., Miranda, M., Pallarès, J., Sales-Pardo, M., 2020a. A Bayesian machine scientist to aid in the solution of challenging scientific problems. Sci. Adv. 6. doi:10.1126/sciadv.aav6971.
Guimerà, R., Reichardt, I., Aguilar-Mogas, A., Massucci, F.A., Miranda, M., Pallarès, J., Sales-Pardo, M., 2020b. Bayesian machine scientist code repository [WWW Document]. https://bitbucket.org/rguimera/machine-scientist/src/no_degeneracy/
Gütschow, J., Jeffery, M.L., Gieseke, R., Gebel, R., Stevens, D., Krapp, M., Rocha, M., 2016. The PRIMAP-hist national historical emissions time series. Earth Syst. Sci. Data 8, 571–603. doi:10.5194/essd-8-571-2016.
Hamrani, A., Akbarzadeh, A., Madramootoo, C.A., 2020. Machine learning for predicting greenhouse gas emissions from agricultural soils. Sci. Total Environ. 741, 140338. doi:10.1016/j.scitotenv.2020.140338.
Hempel, S., Adolphs, J., Landwehr, N., Willink, D., Janke, D., Amon, T., 2020. Supervised machine learning to assess methane emissions of a dairy building with natural ventilation. Appl. Sci. 10, 1–21. doi:10.3390/app10196938.

Ibrahim, S.S., Celebi, A., Ozdeser, H., Sancar, N., 2017. Modelling the impact of energy consumption and environmental sanity in Turkey: a STIRPAT framework. Procedia Comput. Sci. 120, 229–236. doi:10.1016/j.procs.2017.11.233.
IEA, 2021. Global Energy Review 2021, Global Energy Review 2021. Paris.
IEA, 2020. Sources of methane emissions [WWW Document]. https://www.iea.org/data-and-statistics/charts/sources-of-methane-emissions-2 (accessed 6.9.21).
Jean, N., Burke, M., Xie, M., Davis, W.M., Lobell, D.B., Ermon, S., 2016. Combining satellite imagery and machine learning to predict poverty. Science (80-.) 353, 790–794. doi:10.1126/science.aaf7894.
Jiang, S., Hua, H., Sheng, H., Jarvie, H.P., Liu, Xin, Zhang, Y., Yuan, Z., Zhang, L., Liu, Xuewei, 2019. Phosphorus footprint in China over the 1961–2050 period: historical perspective and future prospect. Sci. Total Environ. doi:10.1016/j.scitotenv.2018.09.064.
Le, T., Nguyen, C.P., 2020. Determinants of Greenhouse gas emissions revisited: a global perspective. Singapore Econ. Rev 1–27. doi:10.1142/S0217590820500514.
Lee, J.H., Shin, J., Realff, M.J., 2018. Machine learning: overview of the recent progresses and implications for the process systems engineering field. Comput. Chem. Eng. 114, 111–121. doi:10.1016/j.compchemeng.2017.10.008.
Lenzen, M., Moran, D., Kanemoto, K., Geschke, A., 2013. Building EORA: a global multi-region input–output database at high country and sector resolution. Econ. Syst. Res. 25, 20–49. doi:10.1080/09535314.2013.769938.
Libao, Y., Tingting, Y., Jielian, Z., Guicai, L., Yanfen, L., Xiaoqian, M., 2017. Prediction of CO_2 emissions based on multiple linear regression analysis. Energy Procedia 105, 4222–4228. doi:10.1016/j.egypro.2017.03.906.
Lin, C., Gao, Y., Huang, J., Shi, D., Feng, W., Liu, Q., Du, X., 2020. A novel numerical model for investigating macro factors influencing building energy consumption intensity. Sustain. Prod. Consum. doi:10.1016/j.spc.2020.07.014.
Ma, Z., Liu, W., Wang, L., Ma, P.L., Wang, Y.X., Dong, D.M., Duan, H.Y., Wang, X.E., 2013. Study on energy consumption prediction and energy management in jilin province based on STIRPAT model. Appl. Mech. Mater. 281, 542–545. doi:10.4028/www.scientific.net/AMM.281.542.
Marmier, A., Schosger, J..-P., 2020. Methane as a Greenhouse Gas,From 'unconventional' methane production and recovery to biological mitigation options: a literature review relying on text-mining tools, publications office of the european union. https://doi.org/10.2760/09370
Neumann, P., Cao, L., Russo, D., Vassiliadis, V.S., Lapkin, A.A., 2020. A new formulation for symbolic regression to identify physico-chemical laws from experimental data. Chem. Eng. J. 387, 123412. doi:10.1016/j.cej.2019.123412.
Nguyen, H.A.T., Sophea, T., Gheewala, S.H., Rattanakom, R., Areerob, T., Prueksakorn, K., 2021. Integrating remote sensing and machine learning into environmental monitoring and assessment of land use change. Sustain. Prod. Consum. 27, 1239–1254. doi:10.1016/j.spc.2021.02.025.
Nosheen, M., Abbasi, M.A., Iqbal, J., 2020. Analyzing extended STIRPAT model of urbanization and CO_2 emissions in Asian countries. Environ. Sci. Pollut. Res. 27, 45911–45924. doi:10.1007/s11356-020-10276-3.
Pao, H.-.T., Fu, H.-.C., Tseng, .C..-L., 2012. Forecasting of CO_2 emissions, energy consumption and economic growth in China using an improved grey model. Energy 40, 400–409. doi:10.1016/j.energy.2012.01.037.
Romeiko, X.X., Guo, Z., Pang, Y., Lee, E.K., Zhang, X., 2020. Comparing machine learning approaches for predicting spatially explicit life cycle global warming and eutrophication impacts from corn production. Sustain 12, 1–19. doi:10.3390/su12041481.
Saha, D., Basso, B., Robertson, G.P., 2021. Machine learning improves predictions of agricultural nitrous oxide (N_2O) emissions from intensively managed cropping systems. Environ. Res. Lett. 16. doi:10.1088/1748-9326/abd2f3.
Saleh, C., Dzakiyullah, N.R., Nugroho, J.B., 2016. Carbon dioxide emission prediction using support vector machine. IOP Conf. Ser. Mater. Sci. Eng. 114. doi:10.1088/1757-899X/114/1/012148.
Schweidtmann, A.M., Mitsos, A., 2019. Deterministic Global Optimization with Artificial Neural Networks Embedded. J. Optim. Theory Appl. 180, 925–948. doi:10.1007/s10957-018-1396-0.
Singh, M.K., Mukherjee, D., 2019. Drivers of greenhouse gas emissions in the United States: revisiting STIRPAT model. Environ. Dev. Sustain. 21, 3015–3031. doi:10.1007/s10668-018-0178-z.
Teixidó-Figueras, J., Steinberger, J.K., Krausmann, F., Haberl, H., Wiedmann, T., Peters, G.P., Duro, J.A., Kastner, T., 2016. International inequality of environmental pressures: decomposition and comparative analysis. Ecol. Indic. 62, 163–173. doi:10.1016/j.ecolind.2015.11.041.
The World Bank, 2021. Monthly temperature per country [WWW Document]. https://climateknowledgeportal.worldbank.org/download-data (accessed 10.4.21).
The World Bank, 2020a. GDP per capita (current US$) [WWW Document]. https://data.worldbank.org/indicator/NY.GDP.PCAP.CD (accessed 10.4.21).
The World Bank, 2020b. Population, total [WWW Document]. https://data.worldbank.org/indicator/SP.POP.TOTL (accessed 10.4.21).
The World Bank, 2019a. Population ages 15-64 (% of total population) [WWW Document]. https://data.worldbank.org/indicator/SP.POP.1564.TO.ZS (accessed 10.4.21).
The World Bank, 2019b. Population density (people per sq. km of land area) [WWW Document]. https://data.worldbank.org/indicator/EN.POP.DNST (accessed 10.4.21).
The World Bank, 2018. Urban population (% of total population) [WWW Document]. https://data.worldbank.org/indicator/SP.URB.TOTL.IN.ZS (accessed 10.4.21).
Tian, H., Xu, R., Canadell, J.G., Thompson, R.L., Winiwarter, W., Suntharalingam, P., Davidson, E.A., Ciais, P., Jackson, R.B., Janssens-Maenhout, G., Prather, M.J., Regnier, P., Pan, N., Pan, S., Peters, G.P., Shi, H., Tubiello, F.N., Zaehle, S., Zhou, F., Arneth, A., Battaglia, G., Berthet, S., Bopp, L., Bouwman, A.F., Buitenhuis, E.T., Chang, J., Chipperfield, M.P., Dangal, S.R.S., Dlugokencky, E., Elkins, J.W., Eyre, B.D., Fu, B., Hall, B., Ito, A., Joos, F., Krummel, P.B., Landolfi, A., Laruelle, G.G., Lauerwald, R., Li, W., Lienert, S., Maavara, T., MacLeod, M., Millet, D.B., Olin, S., Patra, P.K., Prinn, R.G., Raymond, P.A., Ruiz, D.J., van der Werf, G.R., Vuichard, N., Wang, J., Weiss, R.F., Wells, K.C., Wilson, C., Yang, J., Yao, Y., 2020. A comprehensive quantification of global nitrous oxide sources and sinks. Nature 586, 248–256. doi:10.1038/s41586-020-2780-0.
Yang, B., Usman, M., Jahanger, A., 2021. Do industrialization, economic growth and globalization processes influence the ecological footprint and healthcare expenditures? Fresh insights based on the STIRPAT model for countries with the highest healthcare expenditures. Sustain. Prod. Consum. 28, 893–910. doi:10.1016/j.spc.2021.07.020.
York, R., Rosa, E.A., Dietz, T., 2003. STIRPAT, IPAT and ImPACT: analytic tools for unpacking the driving forces of environmental impacts. Ecol. Econ. 46, 351–365. doi:10.1016/S0921-8009(03)00188-5.
Žegklitz, J., Pošík, P., 2021. Benchmarking state-of-the-art symbolic regression algorithms. Genet. Program. Evolvable Mach. 22, 5–33. doi:10.1007/s10710-020-09387-0.
Zhang, C., Wang, Y., Song, X., Kubota, J., He, Y., Tojo, J., Zhu, X., 2017. An integrated specification for the nexus of water pollution and economic growth in China: panel cointegration, long-run causality and environmental Kuznets curve. Sci. Total Environ. doi:10.1016/j.scitotenv.2017.07.107.
Zhang, L., Li, H., Lee, W.J., Liao, H., 2021. COVID-19 and energy: influence mechanisms and research methodologies. Sustain. Prod. Consum. doi:10.1016/j.spc.2021.05.010.
Zhao, C., Chen, B., Hayat, T., Alsaedi, A., Ahmad, B., 2014. Driving force analysis of water footprint change based on extended STIRPAT model: evidence from the Chinese agricultural sector. Ecol. Indic. 47, 43–49. doi:10.1016/j.ecolind.2014.04.048.
Zhu, C., Wang, M., Du, W., 2020. Prediction on peak values of carbon dioxide emissions from the chinese transportation industry based on the SVR model and scenario analysis. J. Adv. Transp. doi:10.1155/2020/8848149, 20208848149.