Guariso, Daniele; Castañeda Ramos, Gonzalo; Guerrero, Omar A.

Article

Budgeting for SDGs: Quantitative methods to assess the potential impacts of public expenditure

Development Engineering

Provided in Cooperation with:
Elsevier

Suggested Citation: Guariso, Daniele; Castañeda Ramos, Gonzalo; Guerrero, Omar A. (2023) : Budgeting for SDGs: Quantitative methods to assess the potential impacts of public expenditure, Development Engineering, ISSN 2352-7285, Elsevier, Amsterdam, Vol. 8, pp. 1-12, https://doi.org/10.1016/j.deveng.2023.100113

Development Engineering 8 (2023) 100113

Contents lists available at ScienceDirect

Development Engineering

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

Budgeting for SDGs: Quantitative methods to assess the potential impacts of public expenditure

Daniele Guariso [a,*], Gonzalo Castañeda [b], Omar A. Guerrero [a,c]

[a] The Alan Turing Institute, London, UK
[b] Centro de Investigación y Docencia Económica (CIDE), Mexico City, Mexico
[c] Centro de Estudios Espinosa Yglesias (CEEY), Mexico City, Mexico

ARTICLE INFO

Keywords:
Public finance
Sustainable Development Goals
Regression analysis
Machine learning
Agent-based models
Impact evaluation
Open spending

ABSTRACT

Using a novel large-scale dataset that links thousands of expenditure programs to the Sustainable Development Goals for over a decade, we analyze the impact of public expenditure on more than 100 different development indicators. Contrary to the single-dimensional view of evaluating expenditure in terms of overall economic growth, we take a multi-dimensional approach. Then, we assess the effectiveness of three quantitative methods for capturing expenditure effects on development: (1) regression analysis, (2) machine learning techniques, and (3) agent computing. We find that, under the existing data and for this particular task, approaches (1) and (2) have difficulties disentangling sector-specific effects (i.e., target effects in the SDG semantics), which is consistent with results in previous empirical research. In contrast, by applying a micro-founded agent-computing model of policy prioritization, we can provide empirical evidence about potential impacts and bottlenecks across a high-dimensional policy space. Our findings suggest that, in the discussion of budgeting for SDGs, one should carefully evaluate the data available, the suitability of data-driven approaches, and consider alternative methods that are richer in terms of incorporating explicit causal mechanisms and scalable to a large set of indicators.

1. Introduction

The United Nations Sustainable Development Goals (SDGs) support several international agendas that advocate for the open access of public data. The Open Spending Data agenda is particularly relevant to the SDGs because it deals with one of the main instruments that governments have at their disposal to promote sustainable development: *public expenditure*.[1] Both the SDGs and the Open Spending Data movement converge at the 'budgeting for SDGs' paradigm, which seeks to coordinate the budgeting process of the fiscal cycle with the 2030 Agenda. Budgeting efficiently across the 17 SDGs is, in fact, an extremely challenging problem. From an operational point of view, there are many interactions among policy issues that obfuscate the influence exerted by government programs on development indicators. Thus, from an empirical perspective, establishing a clear link between expenditure data and development indicators is challenging as much of the measurable patterns could be classified as 'noise'.

The neoclassical economic literature analyzing the expenditure-indicator relationship focuses on a reduced number of functional categories or policy issues; for example, education, health, defense, social assistance, and infrastructure. The impact of allocation decisions is typically measured in terms of economic growth. Most of these studies use growth regressions (Devarajan et al., 1996; Haque, 2004; Agénor and Neanidis, 2011; Bojanic, 2013; Neduziak and Correia, 2017; Yilmaz, 2018), and fewer consider vector autoregressions (Balaev, 2019) or general equilibrium models (Baca Campodónico et al., 2014). At the cost of introducing strong assumptions, some of these studies try to incorporate interdependencies between policy issues and tackle the effect of public expenditure on a broader set of development outcomes (e.g., equilibrium, homogeneity, perfect information, and rationality).

On the heterodox side of economics, the allocation of public expenditure and its effect on several development indicators has been treated through a system-dynamics approach (Qureshi, 2009). Under

* Corresponding author.
E-mail addresses: dguariso@turing.ac.uk (D. Guariso), gonzalo.castaneda@cide.edu (G. Castañeda), oguerrero@turing.ac.uk (O.A. Guerrero).

[1] Along with discussions on open data and budgeting for SDGs, there is great interest in adopting machine learning and AI as new methods that can support policymaking. For example, the UN Department of Economic and Social Affairs developed a natural-language-processing tool that classifies documents into the 17 SDGs (https://linkedsdg.officialstatistics.org); UNSECO recently created an international AI center to support the SDGs (https://ircai.org/); the UN Global Pulse initiative (https://www.unglobalpulse.org) tries to harness big data for humanitarian aid; and numerous private projects have also emerged to bridge the gap between AI, big data, and the SDGs (e.g., 2030Vision and AI4Good).

https://doi.org/10.1016/j.deveng.2023.100113
Received 13 November 2022; Received in revised form 13 May 2023; Accepted 15 May 2023
Available online 23 May 2023

this methodology, it is also possible to consider a network of interdependencies between policy issues. Likewise, it allows to establish different causal mechanisms, yet these are only conceived from a macro point of view.

Finally, a more recent strand, under the umbrella of Policy Priority Inference (PPI), uses agent computing to model the bottom-up process through which public expenditure impacts development indicators, linking micro-behavior to macro-dynamics. This framework was first developed by Castañeda et al. (2018), and recently refined by Guerrero and Castañeda (2022). PPI has been used to analyze various aspects of multidimensional development such as ex-ante policy evaluation (Castañeda and Guerrero, 2019a), policy resilience (Castañeda and Guerrero, 2018), policy coherence (Guerrero and Castañeda, 2020b), public governance (Guerrero and Castañeda, 2021), the impact of government expenditure (Guerrero and Castañeda, 2020a, 2022), sub-national development (Guerrero et al., 2021), and aid effectivenes (Guerrero et al., 2023); all these applications from the point of view of public spending.[2]

Such a large set of alternative methodologies (and corresponding results) provides little guidance to development practitioners and decision-makers on which analytical tool might be the most suitable to understand the $expenditure \rightarrow indicator$ relationship according to their policy objectives and the data available to them. To the best of our knowledge, there does not exist an explicit comparison showing the virtues and limitations of alternative quantitative methods that can be used to assess the impact of public spending on development outcomes. Such a comparative approach is necessary to prevent methodological silos, support policymakers in their planning practices, and advance the budgeting-for-SDGs paradigm.

In this paper, we try to bridge this gap. We employ a unique new dataset with thousands of disaggregated expenditure programs from Mexico. Each of these programs has been manually linked (by experts from the Mexican Treasury) to one or more SDG targets. No other expenditure-SDG linked dataset with this level of granularity exists in the world. We use these data to study the relationship between changes in public spending and changes in indicators through three different quantitative methodologies: (1) regression analysis, (2) machine learning algorithms, and (3) agent computing.[3] We compare their performance and assess whether they are helpful (and in which ways) to establish an $expenditure \rightarrow indicator$ linkage in this particular context.

2. Data and methods

The methodologies to be compared in the following sections use the same datasets: a public expenditure dataset and development indicators data. These two sources offer information to establish matches between government programs and indicators whose dynamics are directly affected by public spending. From an evaluation perspective, it is important to analyze whether or not government programs, and their allocated budget, exert an influence on the evolution of specific indicators. Because of this, we make use of public fiscal data in which government spending is disaggregated with the highest possible level of granularity.

2.1. Public expenditure

Since 2008, the Mexican Ministry of the Treasury (SHCP, the Spanish acronym for Secretaría de Hacienda y Crédito Público) publishes data on annual federal budgets and expenditure at a highly disaggregated level.[4] These data provide information on thousands of expenditure programs that can be tracked across different fiscal years. Expenditure programs describe sets of processes, activities, and services that have the same purpose; for example, 'Support program for road infrastructure' and the 'Assistance program for people with disabilities'. They allow the government to organize resources and achieve its development objectives. Panel (a) in Fig. 1 shows the 10 tranches (i.e., macro categories of expenditure) that accumulate most of the federal budget over the sample period (2008–2020), together with their standard deviation.

A feature that distinguishes this public expenditure dataset from any other in the world is that the SHCP has manually linked several programs to one or more targets of the SDGs.[5] There are 169 targets among the 17 SDGs, so these expenditure-SDG linked data provide unique information to investigate the $expenditure \rightarrow indicator$ relationship.[6] To the extent of our knowledge, such data has never been used for this type of analysis, and no other open spending dataset provides such scale and granularity of the link between expenditure and development indicators. Panel (b) in Fig. 1 shows the proportional distribution of Mexico's federal mapped budget across the SDGs and across the different fiscal years analyzed.

After filtering the expenditure programs that are classified into the SDGs and that appear in more than one year, we obtain a sample of 558 programs spanning the 2008–20 period.[7] It appears that, during the sample period, a significant fraction of the budget was consistently allocated to programs related to SDG 3 ('Good Health and Well-being') and 4 ('Quality Education'). We also notice a steady reduction in the spending associated with SDG 16 ('Peace, Justice and Strong Institutions'), coupled with a larger proportion of resources directed to SDG 7 ('Affordable and Clean Energy'), especially in the last years of our sample.

2.2. Development indicators

We obtain the SDG development indicators from two main sources: the Statistics Division of the United Nations,[8] and the database on World Development Indicators of the World Bank.[9] Most of these indicators correspond to the ones officially supplied by Mexico's National Statistics and Geography Institute (INEGI by its Spanish acronym) which, in turn, the Mexican federal government uses to evaluate its progress. Both the United Nations and the World Bank indicators are already classified into SDG targets, completing the linkage between expenditure data and indicators.

[2] In addition, this model has been adopted by governments (Gobierno del Estado de México, 2020) and international organizations (Sulmont et al., 2021; Castañeda and Guerrero, 2019c,d,b; Palacios et al., 2022; Castañeda and Guerrero, 2022a,b) to assess issues related to public financial management.

[3] These are not the only quantitative methods available, but they are the ones with least limitations when it comes to scalability as the number of indicators grows. Hence, the paper focuses on them.

[4] Through its fiscal transparency portal: *Transparencia Presupuestaria* (https://www.transparenciapresupuestaria.gob.mx).

[5] In 2017, the SHCP published the first methodology on budget tagging for the SDGs (SHCP, 2017). This document was created by the technical staff of the SHCP and was the result of a long-term agenda on program-oriented budgeting that was set in motion a decade earlier. For this reason, today, the SHCP is the institution with the most experience in tagging budgets to the SDGs and, thus, this dataset is of the highest quality available.

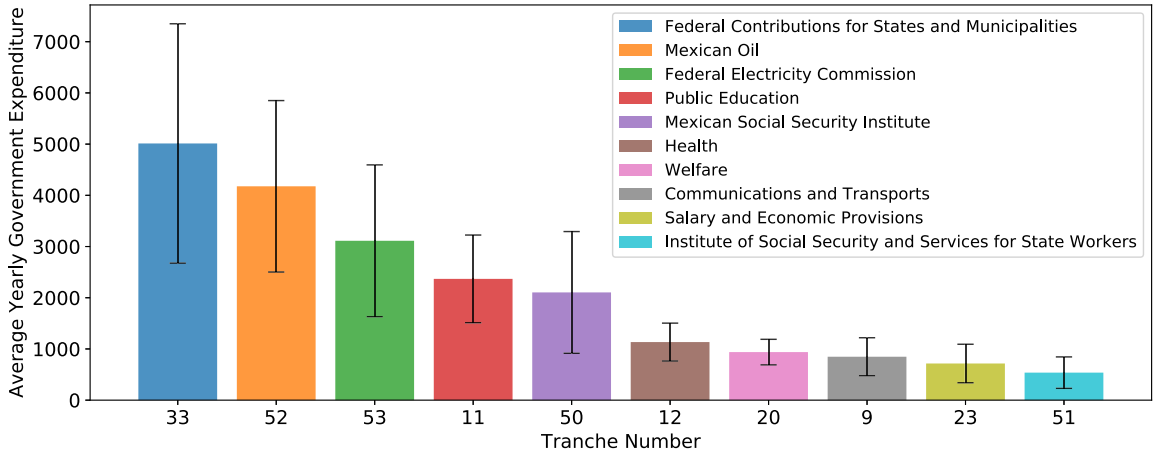
[6] Note that not all the expenditure programs within the federal budget are present in this classification. This is why we refer to the 'mapped' budget in panel (b) of Fig. 1.

[7] Note that expenditure programs are specific to each fiscal cycle, so not all of them are persistent across the different years.

[8] Source: *UNSD* (https://unstats.un.org).

[9] Source: *WDI* (https://datatopics.worldbank.org/world-development-indicators).

(a) Average expenditure across the top ten tranches



(b) Federal mapped budget across the SDGs

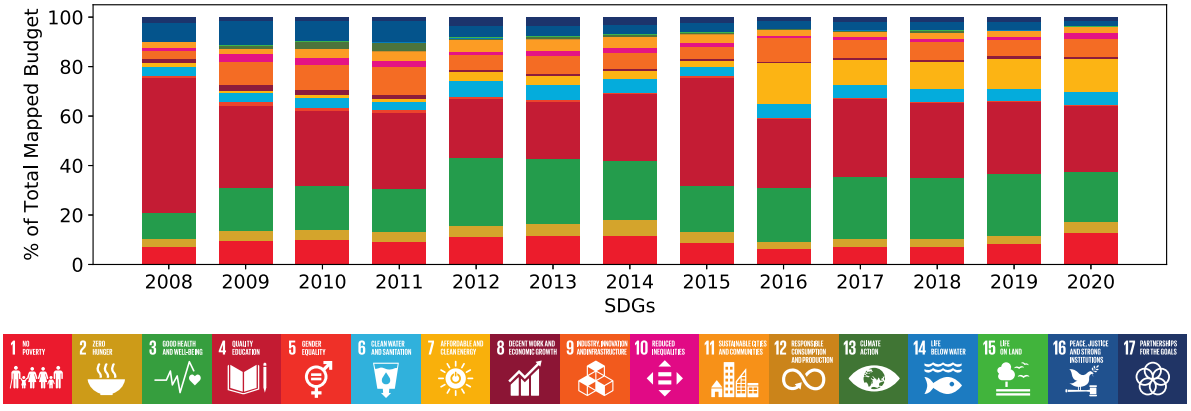


Fig. 1. Expenditure data. All expenditure data are deflated using the Mexican National Consumer Price Index. The values reported in panel (a) are in Millions of Mexican Pesos (MXN). Panel (a) shows the average yearly government expenditure for the top ten tranches over the sample period (2008–2020). The whiskers show the standard deviation of the expenditure level within each tranche, between the different years of the sample. Panel (b) shows the proportional distribution of the mapped budget across the different SDGs for each of the fiscal year considered.

To correct for panel imbalances, we impute missing values using Gaussian Processes (GPs).[10] GPs are a highly flexible class of supervised learning methods that allow modeling nonlinear data and have a wide variety of applications in the context of Bayesian inference. Generally speaking, a GP can be conceptualized as a probabilistic distribution over functions, and it is fully specified by a mean and covariance function (or kernel) (Williams and Rasmussen, 2006). Hence, the random variables are the functions' values at a given location point. In recent years, GPs have become a popular choice to input missing values in time series as they are able to capture non-linear dynamics. For instance, Becker et al. (2017) use GPs to analyze composite development indicators, while Guerrero and Castañeda (2022), Guerrero et al. (2021) use them to impute missing values across development indicators of different countries and subnational regions.

After cleaning and pre-processing the data, we obtain 143 unique indicators, spanning 82 targets and all 17 SDGs. As it is common practice in the development literature, the indicators have been re-scaled to be in the [0, 1] interval (see Appendix B for full details). For the purpose of interpretation, those indicators where lower values imply better outcomes (e.g., number of deaths from diabetes) were inverted through the operation $1 - normalizedValues$, which is also common in development studies as it does not affect the results; only how they are read (see table A.1 in Appendix A for a list with all the indicators).[11]

Table 1 shows, at the level of individual SDGs, the number of indicators, their mean, and standard deviation over the sample period (2008–2020). SDG 3 ('Good Health and Well-being') and 11 ('Sustainable Cities and Communities') concentrate a larger number of indicators (23 and 24, respectively). SDG 14 ('Life Below Water') and 2 ('Zero Hunger') display the highest average values (0.72 and 0.70), whereas SDG 15 ('Life on Land') and 13 ('Climate Action') have the largest volatility (a standard deviation of 0.32 and 0.30).

[10] For the imputation procedure, the maximum time coverage that we consider for the United Nations data is from 2000 to 2020. For the World Bank data, it is 1990–2020.

[11] For indicators where direction has an ambiguous interpretation in terms of the goodness of an outcome (e.g., public debt), we leave them in their original form.

Table 1
Descriptive statistics of development indicators.

SDG	N	Mean	Std Dev
1	18	0.5709	0.2903
2	7	0.6997	0.2179
3	23	0.6088	0.2726
4	17	0.6697	0.2871
5	5	0.6102	0.2890
6	6	0.6992	0.2206
7	5	0.4832	0.2806
8	15	0.6698	0.2438
9	8	0.5678	0.2511
10	4	0.5946	0.2786
11	24	0.6193	0.2864
12	2	0.5240	0.2800
13	8	0.5460	0.3031
14	3	0.7248	0.2000
15	2	0.4417	0.3254
16	7	0.6541	0.2529
17	13	0.6592	0.2672

All indicators have been normalized between 0 and 1, and higher values represent more development (see Appendix B for full details). The table reports (at the level of each SDG) the number of indicators, mean value, and standard deviation over the sample period (2008–2020).

2.3. The econometric approach: Average effects

The first methodology that we employ to study the $expenditure \rightarrow indicator$ relationship is *regression analysis*. Under such framework, the objective is to estimate the average effect of a change in public spending on the development indicators. To be more precise, let us specify a logistic regression to model the probability of observing a change in an indicator I_k due to changes in the public expenditure corresponding to the associated target.[12] That is, given an indicator representative of an SDG target k, we want to estimate the impact that a change in the associated budgetary allocation S_k has on its growth (improvement) probability. Formally, the statistical relationship can be established in the following terms:

$$logit(p_{k,t+1}) = \beta_0 + \beta_1 \Delta I_{k,t} + \beta_2 \Delta S_{k,t} + \boldsymbol{x}_t \boldsymbol{\gamma}, \tag{1}$$

where $p_{k,t+1} = \Pr\{\mathbb{I}_{k,t+1} = 1\}$.

In Eq. (1), $\mathbb{I}$ is a binary variable that yields 1 if indicator I_k improves and 0 otherwise. $\Delta I_{k,t}$ represents the relative change in the indicator during year t, and $\Delta S_{k,t}$ is the main term of interest: *the relative change in the budget associated to target* k. Vector $\boldsymbol{x}_t$ contains year dummies (i.e., binary variables that take the value of 1 for the corresponding year t) to control for shocks that may impact all the indicators in a given year.[13]

Although this type of statistical analysis does not allow establishing a causal relationship, formally, the fact that the expenditure change is set up with a lagged value–with respect to the dependent variable–indicates that future advances in the development indicators are associated with the current budgetary allocation. It is important to emphasize that these regressions do not impose budgetary restrictions across the different policy issues. Hence, they cannot disentangle the average effect of an increase in expense associated to a specific target from a reallocation between targets. While some studies attempt to introduce budgetary constraints, they do so in extremely narrow settings. In our context (the SDGs), the large number of development indicators renders this strategy unfeasible as scaling to a large N is not possible. However, by including time dummies, we partly control for yearly changes in budget size. In the following analysis, we first pool the data across all the targets in the 17 SDGs and then within each of them. We aim at measuring the average effect of a relative change in expenditure in terms of their levels or as a proportion of the total budget. With this procedure, we can infer whether or not there is a systematic association between current expenditures and future development.

2.4. The machine learning approach: Predictive accuracy

While linear regression is also considered part of machine learning (ML), here we refer to ML as all the other algorithms that allow detecting patterns in the data and make possible analytical predictions.[14] At a conceptual level, the most salient difference between the regression and the ML approach is that the former assumes a specific function connecting dependent and independent variables, while the latter is agnostic about the relationship between the variables. In other words, ML can be used to 'discover' or 'learn' such relationships, even if it is at the cost of interpretability.

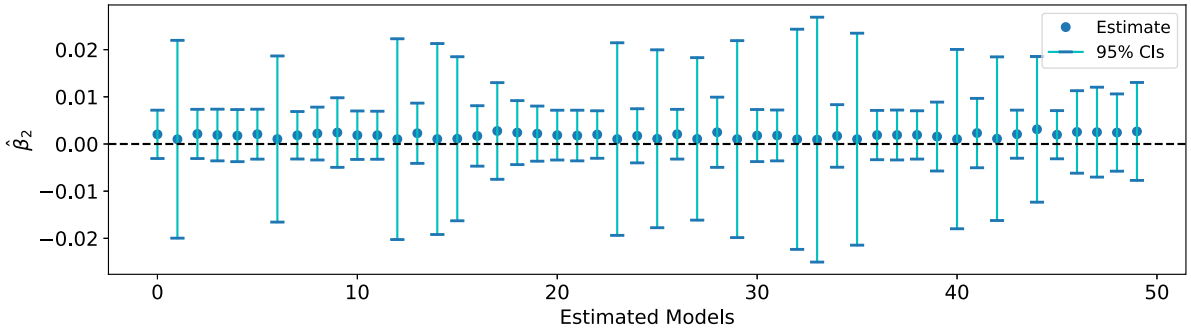
Interpretability is one of the key issues that differentiates how regression and ML methods are used, at least in the social sciences. With adequate controls and a convincing theoretical causal framing, regression results are often interpreted as average treatment effects. On the contrary, ML analysis typically focuses on prediction (although there is active work in trying to bring ML closer to causal inference). In other words, the traditional use of ML does not aim at estimating effects but at improving predictive accuracy Breiman (2001b). Thus, the type of insights that a policymaker would obtain is not so much about average impacts, but on the ability to foresee changes in the indicators following budgetary adjustments (and these predictions may be the result of multiple confounding factors, not only from changes in the relevant expenditure programs). Naturally, this may serve a different purpose than regression analysis. On the one hand, a regression can be used to compare expenditure effectiveness across different allocations, which can inform negotiations related to the construction of a new budget. On the other hand, an ML algorithm can be employed to generate short-term predictions from unexpected budgetary adjustments. In other words, regressions can be employed for ex-post analyses of budgetary decisions while ML approaches can be used for ex-ante analyses of budgetary changes.

[12] The following logistic regressions are estimated in R through the function `glm`, specifying `binomial` for the argument `family` (then `logit` is the default link function for this error distribution).

[13] One could argue that by dichotomizing the dependent variable we lose important information on the magnitude of an indicator's change. In addition, it could be argued that this model does not take into account the presence of time-invariant unobserved heterogeneity, which might capture structural factors specific to development areas. To address these concerns, we reproduce in, Figure C.1 of Appendix C, the results reported in Fig. 2 using the relative change in the indicator during year t+1 (i.e., $\Delta I_{k,t+1}$) as the dependent variable. Furthermore, as suggested by an anonymous reviewer, we estimate a panel model including fixed effects at the indicator level. We observe from both panel (a) and (b) that the main insights from our baseline model still hold: public spending appears to have no significant effect on development outcomes. Note that one could augment our specification including contemporaneous spatial lags of both the dependent variable and public spending, which would capture spillover effects from other development areas and budget programs. However, as discussed by Anselin et al. (2008), the parameters of such a dynamic spatial Durbin model will not be identified. Elhorst (2014) analyzes the assumptions that have to be made for solving the identification problem (e.g., the exclusion of exogenous or endogenous interaction effects). However, such parameter restrictions also limit the scope of empirical analyses focusing on short-term effects (Elhorst, 2011), which might be of particular interest for a government trying to optimize its budgetary allocation. The method proposed in Section 2.5 instead is able to incorporate this kind of interactions with less restrictive assumptions.

[14] Most machine learning textbooks consider linear regressions as part of this field. However, among social scientists, there is an implicit understanding that linear regressions are the 'more traditional' way of conducting quantitative research, while ML stands for a set of algorithms such as tree classifiers, neural networks, and clustering methods that are used to exploit big data, unconventional features, and non-structured information.

(a) Relative change in expenditure



(b) Relative change in expenditure as a proportion of the total budget

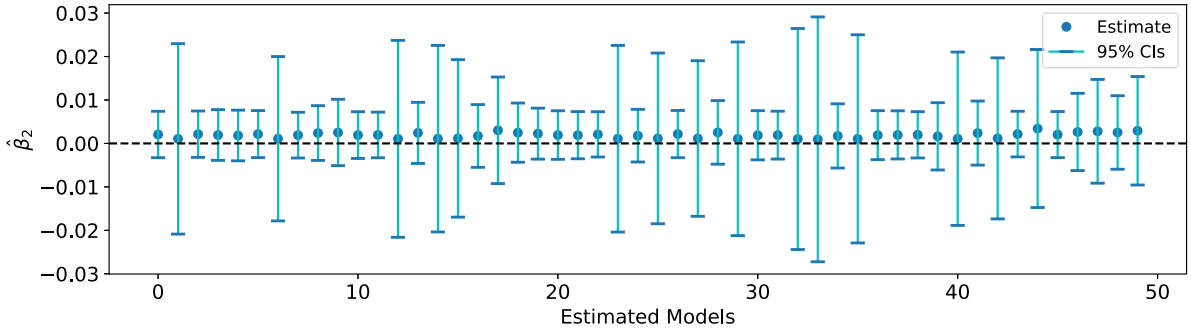


Fig. 2. Point estimates and confidence intervals (CIs) for ΔS. In the logistic regressions, the dependent variable is a binary outcome that takes the value of 1 if the indicator improves in a given year. For each model, one representative indicator is randomly chosen for every SGD target. We then estimate the parameter inferring the average effect of expenditure by pooling the data across all the SDG targets. The horizontal axis reports the number of the model, whereas the vertical one reports the estimated coefficient on the expenditure term, together with its 95% confidence interval. In panel (a) we use the relative change in the raw level of expenditure. In panel (b) we look at the relative change in the expenditure associated to a given SDG target, as a proportion of the total budget.

In ML jargon, expenditure programs, or rather their levels and changes, would be called 'features'. A common requirement of ML algorithms is that the number of observations has to be substantially larger than the number of features (typically, by orders of magnitude). Hence, if we want to account for all the indicators and expenditure programs, it is necessary to increase the size of our dataset. One way to achieve this is through data augmentation, which is a regularization technique common in different fields of data science (e.g., image recognition and signal processing). In this context, augmentation happens by relaxing the expert classification of expenditure, so that we have no prior about which programs are the most impactful in specific indicators. Hence, by linking each indicator with each expenditure program considered in the SHCP mapping, we generate more than 270,000 unique combinations of budgetary and indicator changes; enough observations to apply most ML algorithms. With a larger number of data points now available, we can enrich the specification by including higher-order lags and additional information on the expenditure programs. Now, we can predict $\mathbb{I}$ according to the following expression:

$$\hat{\mathbb{I}}_{i,t+1} = \hat{f}(\Delta I_{i,t}, \Delta I_{i,t-1}, \Delta S_{i,t}, \Delta S_{i,t-1}, \Delta S_{i,t-2}, \Delta S_{i,t-3}, \boldsymbol{r}_i) \tag{2}$$

where i represents a unique indicator-expenditure program combination, while the vector $\boldsymbol{r}$ contains dummy variables at the tranche level (i.e., binary variables that take the value of 1 for the tranche to which the expenditure program in i belongs).

In Eq. (2), the binary variable $\mathbb{I}$ yields 1 if the development indicator associated with the combination i improves and 0 otherwise.[15] In this application, we employ a specific ML classifier: *random forests* (Breiman, 2001a).[16] Random forests are decision-tree-based algorithms that have become increasingly popular in the social sciences, and they have already been used in development studies in the context of the SDGs (e.g., see Asadikia et al. (2021)).[17]

2.5. The agent-computing approach: Explicit mechanisms

Recently, agent computing (also known as agent-based modeling or multi-agent systems) has become a reliable empirical tool to study various social phenomena.[18] It is flexible enough to accommodate different types of theories without imposing unnecessary restrictions when applying an estimation methodology. Because causal chains are explicit in these models, causal effects are directly interpretable. Agent-computing models for the SDGs are not yet common in the development literature (see Guerrero et al. (2023) for a recent example using agent computing to estimate the impact of aid transfers). However, they have the capability of dealing with the difficulties of accounting for bottom-up causal channels in a policy space with complex interdependencies between the SDGs, a feature that seems prevalent across SDG studies.[19]

[15] Note that the term ΔS_k in Eq. (1) refers to the change in public spending allocated to a given SDG target k, according to the SHCP mapping (which may comprise multiple expenditure programs). In specification (2), instead, ΔS_i represents a change in spending at the level of a single expenditure program.

[16] We also train a multi-layer perceptron (with three hidden layers), obtaining similar results.

[17] We estimate the random forests through the class `RandomForestClassifier` of the Python library `scikit-learn`. Details on the optimization procedure are provided in footnote 26.

[18] See Castañeda (2020), for a textbook exposition and references cited there.

[19] Allen et al. (2016) conducted a survey on quantitative models used in the context of the SDGs in government documents and policy reports. They found that less than 1% are agent-computing ones.

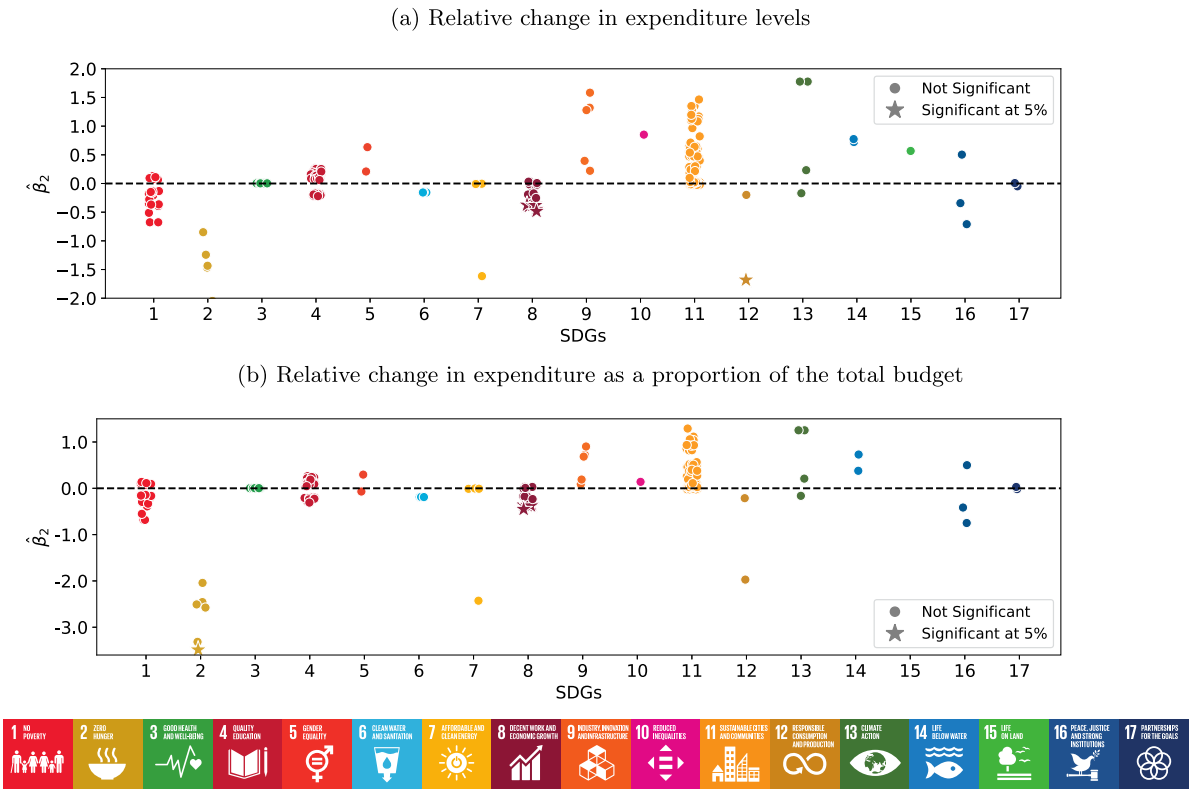


Fig. 3. Point estimates for ΔS at the level of individual SDGs. In the logistic regressions, the dependent variable is a binary outcome that takes the value of 1 if the indicator improves in a given year. For each model, one representative indicator is randomly chosen for every SGD target. We then estimate the parameter inferring the average effect of expenditure by pooling the data within each SDG. The horizontal axis reports the number of the SDG, whereas the vertical one reports the estimated coefficient on the expenditure term for each model. Stars denote results that are statistically significant at the 5% level. In panel (a) we use the relative change in the raw level of expenditure. In panel (b) we look at the relative change in the expenditure associated to a given SDG target, as a proportion of the total budget.

As an example of this systemic perspective, a framework called *Policy Priority Inference* (PPI) has been developed (Guerrero and Castañeda, 2020a, 2022). Here, we analyze how PPI can provide insights about the budget-indicator relationship by exploiting its ability to generate counterfactual data.

2.5.1. Model overview

First, let us provide a general description of the model underlying the PPI framework, highlighting the most empirically-relevant equations. The full model details are provided by Guerrero and Castañeda (2022), and a transcription can be found in Appendix D. The PPI model consists of a political-economy game between a central authority that allocates resources and policymaking agents that implement the government programs using part of those resources. That is, the contributions used to improve the performance of policy issues tend to be lower than the allocated resources because there are inefficiencies in the policymaking process. Each policymaking agent is in charge of government programs, and their performance is evaluated through specific indicators. This is a dynamic model in which, with each iteration, the central authority allocates resources across the expenditure programs, and the policymaking agents learn what is the most 'profitable' efficiency rate for those resources (as they have private incentives to be inefficient). To connect expenditure to policy outcomes, we say that the probability of success of policy i in period t is given by

$$\gamma_{i,t} = \beta_i \frac{C_{i,t} + \frac{1}{N}\sum_j C_{j,t}/P_{j,t}}{1 + e^{-S_{i,t}}}, \tag{3}$$

where β_i is a normalizing parameter and $S_{i,t}$ are the spillovers received by indicator i (these could be positive or negative). $S_{i,t}$ means that PPI accounts for a network $\mathbb{A}$ (its adjacency matrix) of interdependencies between the different indicators. This network is provided to the model as an input and can be estimated through various methods, as discussed by Ospina-Forero et al. (2020). Here, we follow Guerrero and Castañeda (2022) and estimate $\mathbb{A}$ using a Bayesian method for sparse graphs called `sparsebn` by Aragam et al. (2019) (see Appendix E for further details). The spillovers are computed every period according to $S_{i,t} = \sum_j \mathbf{1}_{j,t}\mathbb{A}_{j,i}$, where $\mathbf{1}$ is the indicator function: it yields 1 if indicator j grew in the previous period and 0 otherwise.

$P_{i,t}$ corresponds to the total amount of resources that the central authority allocates to indicator i in period t. $C_{i,t}$, represents the fraction of $P_{i,t}$ that the policymaker effectively uses in the relevant policy. $C_{i,t}$ is endogenous as it results from the agents' learning process (see details in Appendix D). $P_{i,t}$ could come from data if this information would be a one-to-one match between expenditure programs and indicators. Here, since programs are matched to targets, we use a version of PPI in which $P_{i,t}$ is endogenously determined within each target, while the allocations at the target level are given by the data.

Eq. (3) establishes a causal link between a budgetary allocation $P_{i,t}$ and the probability of success $\gamma_{i,t}$ of a policy. We say that these events–whether the indicator grows or not–are the result of short-term considerations such as the budgetary process (because budgeting does not change the existing programs, only how much resources go to them).

To complete the *expenditure* $\rightarrow$ *indicator* link, PPI models the indicator dynamics through the stochastic growth process

$$I_{i,t+1} = I_{i,t} + \alpha_i \xi(\gamma_{i,t}), \tag{4}$$

where parameter $\alpha_i > 0$ captures the long-term structural factors of the existing programs that may limit the effectiveness of public spending (e.g., badly designed programs, poor infrastructure, lack of

implementation capacity, ill-designed tenders, etc.). $\xi(\gamma_{i,t})$ represents the outcome of a binary random variable that can take values 0 and 1. The probability of getting 1 (policy success) is $\gamma_{i,t}$.

Of course, the ability of the model to inform policymaking is restricted by the validity and credibility of its underlying theories and the established causal mechanisms. Thus, in contrast to regressions and ML, agent-computing models often need to be subject to extensive internal and external validation tests. PPI has been validated and used in several different domains where policy prioritization is important. Thus, it provides a robust quantitative alternative with the additional benefit of including an explicit theoretical framework. In PPI, to estimate expenditure impacts, one needs to establish a benchmark case and perform counterfactual simulations in which the budget changes. In the next section, we explain how to calibrate the model to set the benchmark scenario.[20] Note that PPI does not require more data than regressions or ML, instead, special attention is given to the formulation, formalization, and validation of the underlying theoretical framework.

2.5.2. Counterfactual analysis: Budgetary frontiers

Since PPI endogenizes several aspects of the data-generating process of the indicators, it produces richer dynamics that are sensitive to the way the counterfactual is built. For this reason, the inference of expenditure impacts under PPI is an exercise that should be performed with prior knowledge about the desired change and the expected resulting allocation. It does not assume any particular–time invariant–functional form between the indicator and the associated expenditure.

To avoid problems related to exploring a vast space of potential alternative allocations and budget sizes, let us exploit a concept that was developed in Guerrero and Castañeda (2022): *the budgetary frontier*. A budgetary frontier is a theoretical situation in which a government has an unlimited and fully efficient budget. In the model, this means $\gamma_{i,t} = 1$ always. Thus, when operating at the budgetary frontier, the performance of the indicators depends exclusively on the structural factors captured by $\alpha_1, \ldots, \alpha_N$, which have been calibrated. Guerrero and Castañeda (2022) use the budgetary frontier to identify structural bottlenecks in the context of the SDGs. More specifically, if an indicator is unable to achieve its corresponding goal by 2030, it is said that there are structural bottlenecks that go beyond short-term expenditure issues and that the relevant government programs need to be revised.

In this paper, we apply a similar logic by implementing a counterfactual where the government operates at the budgetary frontier. More specifically, given the levels that the indicators achieved in 2020 and the calibrated parameters that explain these dynamics, we simulate the indicators for the same period, but with $\gamma_{i,t} = 1$ at all times. Imposing $\gamma_{i,t} = 1$ makes the model deterministic, so by iterating Eq. (4) forward, we obtain the period in which an indicator achieves the 2020 empirical level and, hence, the time saved when additional resources are available.[21] Therefore, our statistic of interest is 'time savings'. Despite being an extreme and hypothetical situation, the budgetary frontier can be very insightful to identify potential bottlenecks: policy issues that cannot be improved through sheer spending, and that require profound structural changes.

3. Results and discussion

3.1. Regression results: Statistical relationships

To estimate Eq. (1), first, we need to refine the input data since multiple indicators may be associated with the same SDG target.[22] Choosing a representative indicator for a policy issue is a common task in development economics and related fields. Typically, this is done by exploiting ex-ante knowledge on the topic or by building a composite indicator from multiple sources. While the specification of the econometric model is straightforward, its estimates may be subject to the particular availability or choice of data. Importantly, this is a limitation of methods that are not designed to handle multiple outputs (since there is only one dependent variable).[23] Rather than focusing on a particular selection of variables, we explore a large space of potential choices (all of them valid in the sense that indicators are already classified into targets through the mapping provided by the SHCP) and establish how insightful a model of this nature can be in terms of estimating expenditure impacts.

To explore different choices, we produce random samples of targets matched to one indicator only. For each of these samples, we estimate Eq. (1) and assess the direction and significance of β_2. After performing a large number of randomized samples, we find no evidence of impact from public expenditure to the indicators. Fig. 2 shows the result of 50 estimations, with none of them yielding a significant (although positive) coefficient. We perform this exercise for relative changes in absolute terms (panel (a)) and as a proportion of the total budget (which takes into account aggregate expansion/shrinking of public spending) (panel (b)).[24]

Next, let us pool data from sub-samples restricted to targets that belong to the same SDG, as it is typically done in development economics studies to reduce noise by analyzing more homogeneous sub-samples. When generating each sub-sample, we still limit the selection to maintain the matching of public expenditure and individual development indicators associated with a given target. Since the problem of multiple indicators per target remains, we exhaust all possible sub-samples within each SDG by running a regression for each one. Thus, we estimate Eq. (1) for narrower topics, hoping to tease out the average effect that public expenditure in a particular SDG has on its associated indicators.[25]

Fig. 3 reports the results of this exercise, for both expenditure in levels and as a proportion of the total budget. As expected, the magnitude of the effect varies across SDGs and, paradoxically, in a few cases, it exhibits a negative sign. However, like in the case of pooled data across SDGs, no systematic positive impacts are detected. For instance, the coefficients are predominantly negative in SDG 2 ('Zero Hunger'), while they are positive in SDG 9 ('Industry, Innovation and Infrastructure') but not significant. The coefficients seem to vary within the same SDG according to different sub-samples, like in SDG 16 ('Peace, Justice and Strong Institutions) where the estimated effects are both positive and negative but, still, not significant. Surprisingly, for the few SDGs in which we find a significant impact (SDGs 2, 8, and 12), the estimated effect is negative.

[20] See Appendix F for further details on data integration and model calibration.

[21] The simulations used to calibrate the model are stochastic because γ is endogenous. In that case, the calibration procedure relies on Monte Carlo simulations, as explained in Guerrero and Castañeda (2022).

[22] On average, we have around 2 indicators per target, with the maximum number of indicators being 18 in target 11.5 (related to a reduction in deaths by natural disasters).

[23] While regressions for systems of equations try to deal with this issue, their specification can be cumbersome as it requires a large number of assumptions to justify why certain expenditure programs relate to specific indicators through particular functional forms. Moreover, specifying such a system becomes a titanic endeavor when trying to explain the outcome of hundreds of indicators.

[24] One could try to isolate the effect of public spending on the indicators by performing regressions exclusively on the expenditure term (i.e., ΔS), and its lags (we thank an anonymous reviewer for pointing this out). However, omitting ΔI when it is expected to have an impact on the dependent variable would lead to model misspecification and biased coefficients. In any case, we reproduce the models of Fig. 2 estimating Eq. (1) without $\Delta I_{k,t}$, and adding several lags of the expenditure term (i.e., $\Delta S_{k,t-1}, \Delta S_{k,t-2}, \Delta S_{k,t-3}$). Across these alternative models, the coefficients on the expenditure terms are never statistically significant, supporting our main insights (results available upon request).

[25] Since the sample size for each estimation is now significantly reduced, we drop the year dummies from the specification.

Table 2
Classifier performance.

	Random Forest (1)	Random Forest (2)
Precision	0.9672	0.9880
	(0.0008)	(0.0002)
Recall	0.9644	0.9899
	(0.0010)	(0.0002)
F1-score	0.9657	0.9889
	(0.0009)	(0.0002)
Accuracy	0.9666	0.9891
	(0.0009)	(0.0001)
AUC	0.9933	0.9995
	(0.0004)	(0.0000)
N Train	164393	164393
N Validation	54799	54799
N Test	54799	54799

The table shows the average values of the metrics. Standard deviations are reported in parentheses. The size of the training set refers to the inner loop. Once the hyperparameters are optimized using the validation set, the model is retrained using both the training and validation sets and evaluated on the test set. For precision, recall, and the F1-score, we report the macro averages, so they do not account for label imbalance. In the second model, we do not include variables on changes in government expenditure.

3.2. Random forest results: Predictive accuracy and feature relevance

In the first column of Table 2, we report a range of performance statistics from applying random forests to estimate the model described in Eq. (2).[26] In the second column, we re-estimate the model excluding terms with information on changes in government expenditure (i.e., $\Delta S_{i,t}$, and its three lags). As expected from this method, for this size of data, predictive power is very high. Notice, however, that the performance of the classifier is greater in the second column, suggesting that the expenditure terms only add noise to the predictions. That is to say, most of the useful information is contained within the indicator's lagged values. This is a worrisome result if the purpose of these estimates is to understand the contribution of government expenditure to the prediction of the indicators. Likewise, the finding goes in the same direction as the result obtained from regression analysis: *expenditure appears to be ineffective*.[27]

Next, we analyze the features' relevance based on their 'mean decrease impurity' (MDI) (Louppe et al., 2013), which is a standard measure for determining the importance of a variable in tree-based methods (Breiman et al., 1984).[28] Panel (a) in Fig. 4 shows the MDI through the normalized Gini importance. Panel (b) presents an alternative global measure of feature importance: the 'mean decrease accuracy' (MDA), through its most common metric: permutation importance. Permutation importance is also a useful metric as it corrects well-known biases of the normalized Gini approach of the MDI (Sandri and Zuccolotto, 2008; Lundberg et al., 2018). The negative scores associated with the expenditure variables can be interpreted as the 'corrupted' models: those with their values randomly permuted being more accurate (by chance) than the original one.

3.3. PPI results: Time savings

The results presented in panel (a) of Fig. 5 show that there could be time savings ranging from 1 to 12 years. Panel (c) indicates that the relevance of time savings can be observed across all SDGs, although some degree of heterogeneity exists. The most prominent average savings are in SDG 7 'Affordable and Clean Energy' and the least ones in SDG 6 'Clean Water and Sanitation'. Finally, panel (b) shows that there is no systematic pattern between the initial level of the indicator and the years saved. This implies that the most lagged indicators at the beginning of the sample period are not necessarily the most sensitive ones to budgetary changes.

Before proceeding to the significance test results, it is important to discuss certain nuances of hypothesis testing in the context of PPI. In order to construct a relevant test, it is necessary to obtain the distribution of the time savings statistic. In contrast with hypothesis testing in a regression framework (where one tries to rule out a zero-value statistic), in PPI we derive the distribution of the statistic from simulations where indicator improvements are not the result of operating on the budgetary frontier, but of their inherent randomness. Hence, hypothesis testing consists of determining whether the empirical statistic of time savings would be expected under its 'null' distribution.

To construct the null distribution, we need to compute how much time would be saved under a random realization of the indicator, but with the original budget (not the frontier). We obtain random realizations of the indicators by using the Gaussian Processes (GPs) that were estimated to impute the missing observations in Section 2.2. Using the point-estimate distribution provided by the GP of each indicator, we generate random realizations of each series. In total, we generate 1000 realizations, which means 1000 alternative datasets and 1000 null statistics. Importantly, each null statistic is obtained by re-calibrating the model parameters using each null set of time series. Not being able to reject the null hypothesis means that budgetary increments have no significant impact since the same progress in the indicators would be expected due to their inherent randomness.

Fig. 6 presents these results. Panel (a) shows the indicators where the budgetary frontier has a significant impact in time savings. Panel (b) shows the rest of the indicators. In both panels, the dots indicate the empirical statistics, while the vertical lines denote their null distribution. Since we perform single-tail tests (at 95% confidence), significance means that the empirical estimator lies beyond the 95th percentile of the null distribution, often meaning that the dot is located above the vertical line in the plot.[29] Approximately one third of the indicators show a significant impact on the budgetary frontier. This

[26] We optimize the maximum number of random input variables considered at each split and the number of trees in a nested cross-validation framework (Varma and Simon, 2006). We use 4 folds in the inner loop and five in the outer one. To create the training, validation, and test sets, we randomly select unique indicator-budget program combinations at the year level. In the inner loop, we perform a grid search over 3 different values for the maximum number of features. These are $\sqrt{p}$ where p is the total number of features, a value typically suggested by the literature (Friedman et al., 2001), and two other values in its neighborhood. The three values considered are 7 ($\sqrt{44} \approx 7$), 4, and 10. For the number of trees, we consider several alternatives: 50, 100, 500, and 1000. Hence, the total number of combinations evaluated are 12.

[27] As an additional exercise, we isolate the effect of government expenditure by estimating the model described in Eq. (2) without the terms capturing the trend of the indicator (i.e., $\Delta I_{i,t}, \Delta I_{i,t-1}$). However, the (mean) predictive accuracy of the model drops significantly (0.5546) as all the other performance metrics (precision: 0.4670; recall: 0.4910; F-1 score: 0.4046; AUC: 0.5165), providing further support to our conclusions. We thank an anonymous reviewer for this suggestion.

[28] The 'purity' of a node is defined as the degree to which the explanatory variable in the node is able to split the data into groups that are characterized by a single label. It can be thought of as a measure of misclassification at the node level. The importance of a variable is then measured by its average contribution to the purity of the nodes when growing the trees over the training sample (e.g., see De'Ath (2007) for an application of variable importance in ecological sciences).

[29] In rare occasions, the statistic may be below the null distribution, meaning that the random fluctuations of the indicator are expected to produce a much higher impact than the one obtained through a budgetary increment in the frontier. This can happen because the null model involves a re-estimation of the parameters, so the estimated αs could change significantly if the random draws from the Gaussian process tend to yield time series with faster dynamics than the original ones.

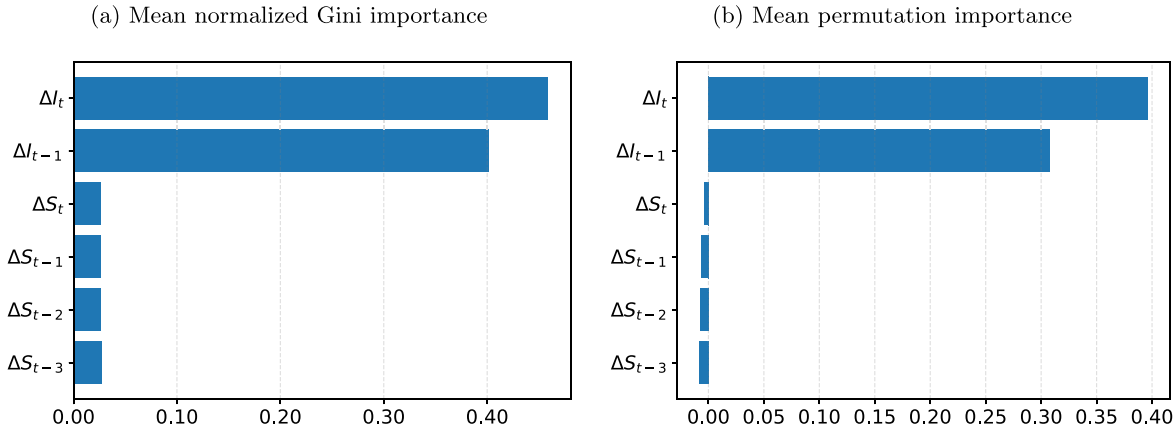


Fig. 4. Variable importance. Panel (a) reports the mean normalized Gini importance, whereas panel (b) shows the mean permutation importance. On the vertical axis, we list the different variables included in the model (the binary variables for the tranches are omitted). On the horizontal axis, we report the variables' score for the relevant metrics. On the right panel, negative scores result from models in which the random reshuffling of the variable's values improves (on average) the accuracy of the classifier compared to the original data.

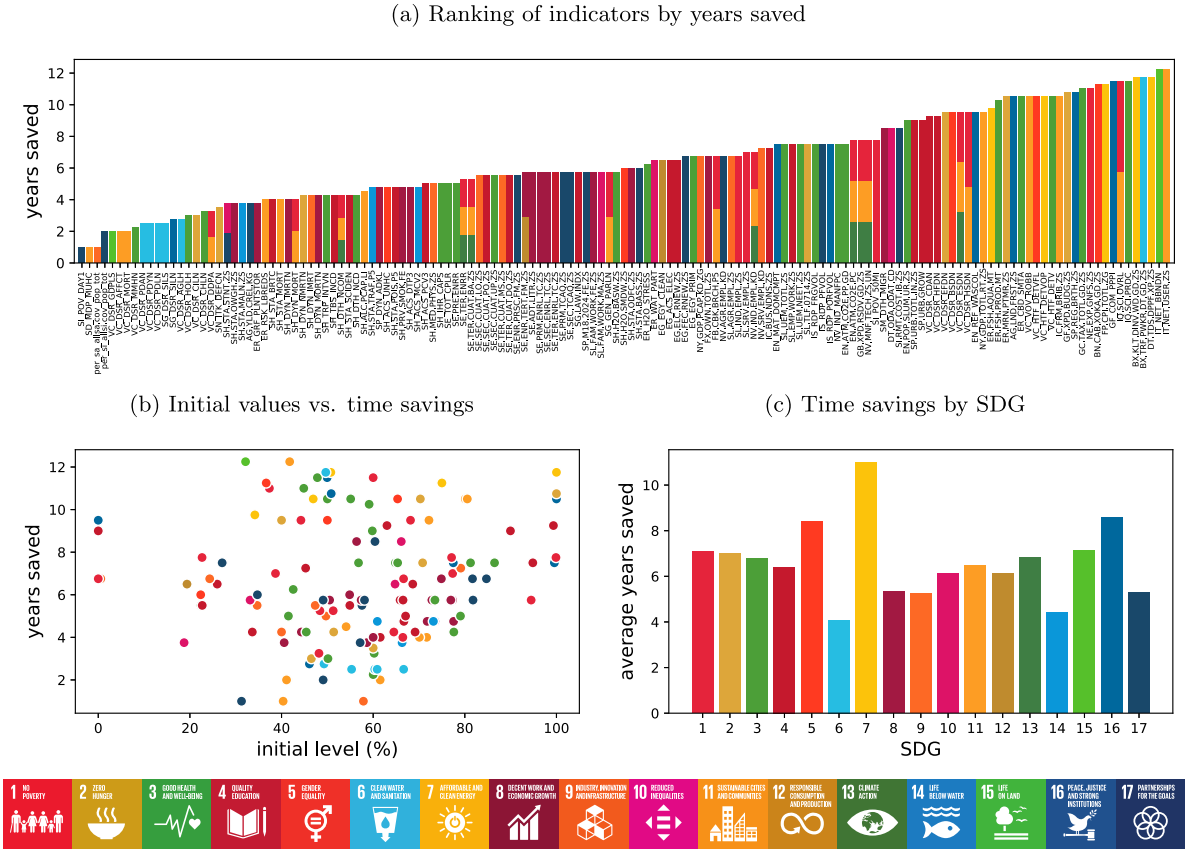


Fig. 5. Time savings from operating at the budgetary frontier. In panel (a), on the vertical axis, we report the number of years saved. In the horizontal axis, we list the indicators through their code (the description for each indicator can be found in Table A.1). The color of an indicator reflects its SDG. Multiple colors in a single bar in panel (a) suggest that the indicator is classified into multiple SDGs and that it receives funds from multiple targets and expenditure programs. In panel (b), we plot the years saved against the initial level of the indicators. The horizontal axis in panel (b) is in percentage. In panel (c), the vertical axis shows the average number of years saved for the indicators belonging to a given SDG, whereas the horizontal axis lists all the SDGs.

result represents an upper bound to realistic budgetary increments since the frontier is a hypothetical-extreme-scenario. In general, we observe significant impacts in indicators of all SDGs, with the exception of SDG 12 and 13.

3.4. Discussion

From appling regression analysis on our SDG-classified expenditure dataset, we obtain inconclusive results regarding the *expenditure* →

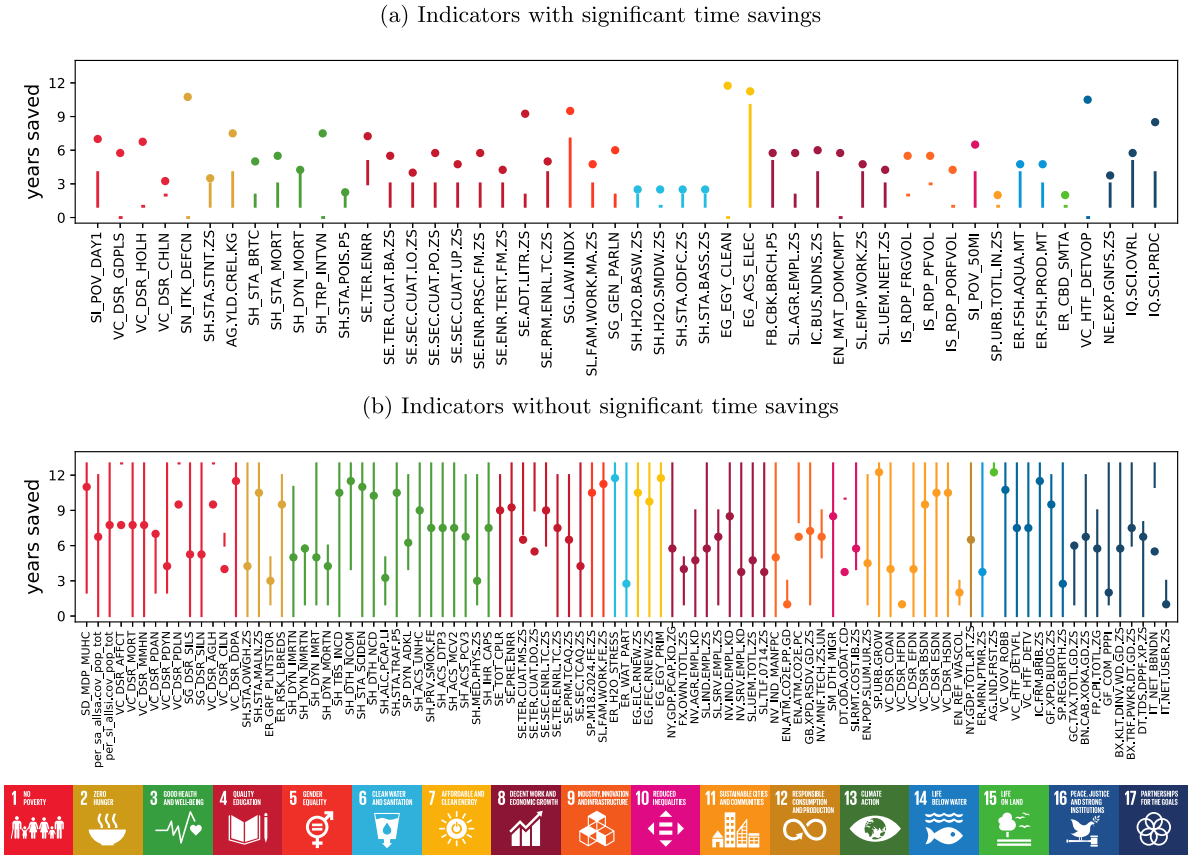


Fig. 6. Hypotheses tests under the budgetary frontier. In the vertical axis, we report the number of years saved. In the horizontal axis, we list the indicators through their code (the description for each indicator can be found in Table A.1). In both panels, the dots indicate the empirical statistics for each indicator and the vertical lines denote the null distribution until the 95th percentile. The color of an indicator reflects its SDG. Panel (a) presents indicators that exhibit significant time savings. Panel (b) shows those without significant time savings. In total, 48 indicators out of 143 exhibit significant time changes under the considered statistical test.

indicator relationship. In principle, one could argue for a non-existent impact of government spending. However, this conclusion is hard to defend given the vast economic literature asserting that public expenditure, at the macro level, is a necessary condition for the development of any country. Alternatively, one could blame the quality of the data since, in the end, the classification of expenditure programs into SDG targets is done by public servants through qualitative protocols.[30] This explanation is also difficult to defend given the fact that protocols for classifying the budgetary data have been designed in collaboration with many international experts and organizations. A more sensible explanation for the lack of a systematic significance of government spending could be that the assumptions on which the regression approach is built are too stringent to handle data with this level of temporal resolution and with a complex and unconventional structure.[31]

Users of the regression approach could argue that data pre-processing could help build a sample that is better fitted for this type of analysis. For example, one could further narrow the scope of the data and the topic of analysis, until there is a very clear mapping in which, beyond a reasonable doubt, an indicator is a reliable reflection of the outcomes of a specific expenditure program. Under this scenario, regression analysis would be likely to provide more insightful results. The problems with this strategy are that (1) it requires additional information that is not readily available, (2) it misses the point of understanding budgeting for SDGs since it lacks of a systemic approach (it ignores multidimensionality), (3) it discards theoretical insights that are informative under alternative methodologies, (4) it makes the analysis not scalable, and (5) any recommendation becomes highly specific and poorly generalizable to the context of national budgets. In addition, the need to provide an explicit functional form may be a methodological constraint that prevents us from properly disentangling useful information. Thus, if one aims at obtaining policy advice that takes advantage of these data, then looking at alternative methodologies may be the right direction. For this reason, we turn to a more flexible ML approach that does not impose restrictive functional forms, and whose focus is predictive accuracy.

Our findings after using ML methods highlight that the relevance of the expenditure terms within the predictive model is negligible. They suggest that changes in public spending are not relevant when building the random forests. The MDA analysis asserts that almost all the accuracy in the model is given by the autoregressive terms of the relative change in the indicator. This would imply that variation in public spending does not help predict the dynamics of development indicators, and that additional information on government expenditure cannot be exploited to improve short-term forecasts.

The results derived from the regression and ML analyses seem to suggest that these data-driven methods might not be the best suited ones to explain the relevance of expenditure in countries' development

[30] Or that the indicators have a poor quality; something difficult to argue in the case of Mexico since INEGI is an internationally recognized institution.

[31] One of the challenges posed by this type of data is given by the fact that information on development indicators and government spending is often collected on yearly basis. Hence, analysts usually end up dealing with very short time series that limit the capabilities of data-driven methods.

when highly disaggregated information is used. It is indeed a surprising outcome since it is hard to believe that the millions of dollars invested in a large number of different government programs do not make a difference in the indicators' performance. To tackle the paradoxical result observed in this context, we suggest that alternative quantitative methodologies, which are supported by sound theoretical frameworks, could better handle the complex structure of this data. The proposed agent-computing approach is flexible enough to include micro and macro causal mechanisms, and it can be scaled to introduce multiple dimensions of development and complex networks of interrelationships among policy issues.

Our results from applying PPI in the case of Mexico indicate that budget increments have a limited impact and that there may be structural factors (e.g., program design) that need to be analyzed. For instance, most of the indicators belonging to SDG 16 ('Peace, Justice and Strong Institutions') do not show significant time savings, suggesting that there might be more profound institutional or cultural constraints that hamper development in this policy area. Nevertheless, these results also suggest that there are indicators that could be improved by increasing their budget, and that these account for approximately one third of the policy issues in the dataset. Contrary to the data-driven frameworks, the finding from PPI implies that public expenditure does matter for development to take place in a large set of indicators, even if no changes are made in the operations and incentives of the established government programs.

Our study is the first to propose a comparative approach that allows understanding both the merits and limits of different quantitative methods that can be used to explore the $expenditure \rightarrow indicator$ link. Such comparison is much needed for policymakers to take informed budgetary decisions. To show how agent computing can shed new light on the $expenditure \rightarrow indicator$ relationship, let us take an example related to environmental policy and focus on SDG 14 ('Life Below Water). When looking at the PPI results (Fig. 6) we notice that two out of three indicators display significant time savings, whereas one appears to be characterized by structural bottlenecks. The first two indicators relate to total fisheries and aquaculture production, which we could intuitively assume that are impacted by public spending (e.g., through economic incentives). The one showing no significant effect measures the percentage of marine protected areas, which is clearly an issue that has to be addressed through more structured policy reforms. This heterogeneous impact across the indicators is lost in the regression analysis (Fig. 3), which only shows a positive, but not statistically significant average effect, providing no useful insight for policymakers at the level of single budget programs. PPI tries to capture this heterogeneity across individual outcomes by modeling the mechanisms that make explicit agents' decisions at the micro-level (i.e., those affecting the policymaking process) and produce interactions at the macro-level. However, one of the limitations of PPI is that it assumes that the set of policy interventions (i.e., the budget programs) is fixed. Hence, while it can identify those programs that appear to be not responsive to public spending (so that display structural constraints), it is not well suited to understand the effect of a policy reform on them.

4. Conclusions

The United Nations 2030 Agenda is a step in the right direction in so far as it considers the mutidimensionality and complexity of development. However, from the policymakers' perspective, it still lacks analytical tools to tackle many of the problems it entails. One of these is the issue of budgeting for the SDGs, in which policymakers have to decide how to allocate their resources among different government programs so that a set of predefined objectives can be attained within a certain time frame.

In this paper, we compare three alternative frameworks applied to a novel database with information from Mexico covering the 2008–20 period, to assess the strengths and weaknesses of existing quantitative methodologies when dealing with large-scale analysis of numerous development indicators and highly-disaggregated budgetary data. Formally speaking, the assumptions underpinning these quantitative tools are different and, hence, they are not entirely comparable. Pooled regressions establish statistical relationships, machine learning approaches offer predictive capabilities, and agent-computing models facilitate systemic causal inference. All of them can be helpful for policymakers interested in establishing an empirical link between expenditure and development indicators, either for estimating average effects, predicting changes in indicators, or inferring the consequences of different budgetary allocations. What these quantitative methods have in common is that they make use of the same type of inputs: large vectors of public expenditure and development indicators.

The main results of this paper are, on the one hand, that pooled regressions and random forests (an ML algorithm) have difficulties in validating the $expenditure \rightarrow indicator$ link in a high-quality SDG-linked expenditure dataset; on the other hand, that agent-computing simulations applied to the same dataset show that spending matters for development. The latter outcome alignes with the theoretical literature, experiences of particular governments' projects, and macroeconomic empirical results. Accordingly, we argue that any approach aiming at guiding empirical analyses on this type of data should incorporate micro-macro mechanisms that may be absent in data-driven methodologies that are not theoretically-oriented. We would like to stress that our claims do not pertain to these well-established analytical tools *per se*, rather to their suitability for this specific task and the limited insights that they can provide given the available data.

The PPI computational model has the additional advantage of identifying which government programs present structural long-term bottlenecks that preclude the advances of development indicators even if public funding were available. This type of outcome is extremely helpful for policymakers since it establishes a connection between budgeting and planning practices. The latter is associated with the definition of a coherent set of objectives to be reached in development indicators through the implementation of specific government programs. Once the objectives are established by the analysts, the model allows learning how different budgetary allocations generate time savings in comparison with the outcome of the historical budget profile.

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.

Data availability

All data are publicly avaialable in the sources cited along the article. The code developed for the analysis is available upon request.

Funding

The authors acknowledge funding from the Economic and Social Research Council (ESRC), UK. Grant code: ES/T005319/1. The funding source had no involvement in the development of the research.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.deveng.2023.100113.

References

Agénor, P.-R., Neanidis, K.C., 2011. The allocation of public expenditure and economic growth. Manch. Sch. 79 (4), 899–931.

Allen, C., Metternicht, G., Wiedmann, T., 2016. National pathways to the Sustainable Development Goals (SDGs): A comparative review of scenario modelling tools. Environ. Sci. Policy 66, 199–207.

Anselin, L., Gallo, J.L., Jayet, H., 2008. Spatial panel econometrics. In: The Econometrics of Panel Data. Springer, pp. 625–660.

Aragam, B., Gu, J., Zhou, Q., 2019. Learning large-scale Bayesian networks with the sparsebn package. J. Stat. Softw. 91 (11), 1–38.

Asadikia, A., Rajabifard, A., Kalantari, M., 2021. Systematic prioritisation of SDGs: Machine learning approach. World Dev. 140, 105269.

Baca Campodónico, J.F., Peschiera Cassinelli, J.R., Mesones, J.A., 2014. The Impact of Public Expenditures in Education, Health, and Infrastructure on Economic Growth and Income Distribution in Peru. Inter-American Development Bank (IDB).

Balaev, A., 2019. The structure of public spending and economic growth in Russia. Russ. J. Econ. 5 (2), 154–176.

Becker, W., Saisana, M., Paruolo, P., Vandecasteele, I., 2017. Weights and importance in composite indicators: Closing the gap. Ecol. Indic. 80, 12–22.

Bojanic, A.N., 2013. The composition of government expenditures and economic growth in Bolivia. Latin Am. J. Econ. 50 (1), 83–105.

Breiman, L., 2001a. Random forests. Mach. Learn. 45 (1), 5–32.

Breiman, L., 2001b. Statistical modeling: The two cultures (with comments and a rejoinder by the author). Statist. Sci. 16 (3), 199–231.

Breiman, L., Friedman, J., Stone, C.J., Olshen, R.A., 1984. Classification and Regression Trees. CRC Press.

Castañeda, G., 2020. The Paradigm of Social Complexity: An Alternative Way of Understanding Societies and their Economies. Centro de Estudios Espinosa Yglesias.

Castañeda, G., Chávez-Juárez, F., Guerrero, O., 2018. How do governments determine policy priorities? Studying development strategies through networked spillovers. J. Econ. Behav. Organ. 154, 335–361.

Castañeda, G., Guerrero, O., 2018. The Resilience of Public Policies in Economic Development. Complexity 2018.

Castañeda, G., Guerrero, O., 2019a. The Importance of Social and Government Learning in Ex Ante Policy Evaluation. J. Policy Model..

Castañeda, G., Guerrero, O., 2019b. Inferencia de Prioridades de Polítca Para el Desarrollo Sostenible. Reporte Metodológico, Programa de las Naciones Unidas para el Desarrollo.

Castañeda, G., Guerrero, O., 2019c. Inferencia de Prioridades de Política Para el Desarrollo Sostenible: El Caso Sub-Nacional de México. Reporte Técnico, Programa de las Naciones Unidas para el Desarrollo.

Castañeda, G., Guerrero, O., 2019d. Inferencia de Prioridades de Polítca Para El Desarrollo Sostenible: Una Aplicación Para El Caso de MÉxico. Reporte Técnico, Programa de las Naciones Unidas para el Desarrollo.

Castañeda, G., Guerrero, O., 2022a. El Presupuesto Público Nacional y los ODS en Colombia: Un Análisis de la Agenda 2030 desde la Metodología de Inferencia de Prioridades de Política (IPP). Documento de Desarrollo, Programa de las Naciones Unidas para el Desarrollo.

Castañeda, G., Guerrero, O., 2022b. Los Objetivos del Desarrollo Sostenible en Bogotá D.C. Un Análisis sobre las Asignaciones Presupuestales y su Impacto en los Indicadores del Desarrollo. Documento de Desarrollo, Programa de las Naciones Unidas para el Desarrollo.

De'Ath, G., 2007. Boosted trees for ecological modeling and prediction. Ecology 88 (1), 243–251.

Devarajan, S., Swaroop, V., Zou, H.-f., 1996. The composition of public expenditure and economic growth. J. Monetary Econ. 37 (2), 313–344.

Elhorst, J.P., 2011. Spatial Panel Models. The University of York, York, UK, p. 7.

Elhorst, J.P., 2014. Dynamic spatial panels: Models, methods and inferences. Spatial Econ. 9, 5–119.

Friedman, J., Hastie, T., Tibshirani, R., 2001. The Elements of Statistical Learning, Vol. 1. In: Springer Series in Statistics, New York, NY.

Gobierno del Estado de México, 2020. Informe de Ejecución del Plan de Desarrollo del Estado de México 2017–2023; a 3 Años de la Administración.

Guerrero, O., Castañeda, G., 2020a. Policy priority inference: A computational framework to analyze the allocation of resources for the Sustainable Development Goals. Data & Policy 2.

Guerrero, O., Castañeda, G., 2020b. Quantifying the Coherence of Development Policy Priorities. Development Policy Rev. 00, 1–26.

Guerrero, O., Castañeda, G., 2021. Does expenditure in public governance guarantee less corruption? non-linearities and complementarities of the rule of law. Econ. Govern. 22 (2), 139–164.

Guerrero, O., Castañeda, G., 2022. How Does Government Expenditure Impact Sustainable Development? Studying the Multidimensional Link Between Budgets and Development Gaps. Sustain. Sci..

Guerrero, O., Castañeda, G., Trujillo, G., Hackett, L., Chávez-Juárez, F., 2021. Subnational sustainable development: The role of vertical intergovernmental transfers in reaching multidimensional goals. Socio-Econ. Plan. Sci. 101155.

Guerrero, O., Guariso, D., Castañeda, G., 2023. Aid effectiveness in sustainable development: A multidimensional approach. World Development 168, 106256.

Haque, M.E., 2004. The composition of public expenditures and economic growth in developing countries. Global J. Finance Econ. 1 (1), 97–117.

Louppe, G., Wehenkel, L., Sutera, A., Geurts, P., 2013. Understanding variable importances in forests of randomized trees. Adv. Neural Inf. Process. Syst. 26, 431–439.

Lundberg, S.M., Erion, G.G., Lee, S.-I., 2018. Consistent individualized feature attribution for tree ensembles. arXiv preprint arXiv:1802.03888.

Neduziak, L.C.R., Correia, F.M., 2017. The allocation of government spending and economic growth: A panel data study of Brazilian states. Revista De Administraç 51 (4), 616–632.

Ospina-Forero, L., Ramos, G.Castañeda., Guerrero, O., 2020. Estimating networks of sustainable development goals. Inform. Manag. 103342.

Palacios, L., Quiroga, D., Romero, O., Ruiz, M., 2022. Sdg alignment and budget tagging: towards an sdg taxonomy. p. 60.

Qureshi, M., 2009. Human development, public expenditure and economic growth: a system dynamics approach. Int. J. Soc. Econ. 36 (1/2), 93–104.

Sandri, M., Zuccolotto, P., 2008. A bias correction algorithm for the Gini variable importance measure in classification trees. J. Comput. Graph. Statist. 17 (3), 611–628.

SHCP, 2017. Vinculación del Presupuesto a Los Objetivos de Desarrollo Sostenible. Anexo 2 de Los Lineamientos Para el Proceso de Programación Y Presupuestación Para El Ejercicio Fiscal 2018. Secretaría de Hacienda y Crédito Público, CDMX.

Sulmont, A., Rivas, M.García.de.Alba., Visser, S., 2021. Policy priority inference for sustainable development: A tool for identifying global interlinkages and supporting evidence-based decision making. In: Understanding the Spillovers and Transboundary Impacts of Public Policies. OECD Publishing, Paris.

Varma, S., Simon, R., 2006. Bias in error estimation when using cross-validation for model selection. BMC Bioinformatics 7 (1), 1–8.

Williams, C.K., Rasmussen, C.E., 2006. Gaussian Processes for Machine Learning, Vol. 2. MIT Press, Cambridge, MA.

Yilmaz, G., 2018. Composition of public investment and economic growth: Evidence from Turkish provinces, 1975-2001. Public Sector Econ. 42 (2), 187–214.