Complexity Economics and Sustainable Development:

A Computational Framework for Policy Priorities

Omar A. Guerrero & Gonzalo Castañeda

Chapter 4

A Computational Model

We present an agent-computing model of a government that allocates resources across several policy issues, and the policymaking agents that transform those resources into policy outcomes by making use of the existing government programmes.

4.1 The model

Let us begin by describing an economy with $1, 2, \ldots, N$ policy issues; these could be social, economic, environmental, technological, etc. In this economy, there is a government or a central authority that is responsible for financing the policy programs designed to improve some of these issues. We say *some* because the policy space may be too diverse for the government to have specific programs designed to directly impact each issue. For example, we could safely assume that not all countries have policy instruments to adequately deal with cybersecurity threats, so this responsibility relies, for the most part, on private companies. Alternatively, a policy issue–such as GDP (which is the result of many socioeconomic processes)– may be too broad or aggregate, so it is hardly arguable that any government has a policy program that directly impacts GDP with a reliable degree of control. Hence, our central authority has direct influence over $n \leq N$ policy issues, which we call *instrumental issues.*

While the government agent may only be able to directly affect n policy issues, it could have goals or aspirations for all the N topics as they are all socially relevant. Thus, the central authority monitors the progress of each issue through N development indicators, each one representing a higher degree of development if its value is larger. An *instrumental indicator* quantifies the level of development in an instrumental policy issue. If the associated issue is not instrumental, we say that the indicator is *collateral*, as its progress depends on spillovers and other societal factors that are out of the direct control of the government. Of course, which indicator is instrumental and which one collateral depends on each country and government level (national vs subnational). This is something that each user needs to determine when preparing the data for the model, as it may reflect context specificities that are important when translating the model inferences into policy prescriptions.

The readers of this book may be methodologically diverse. For example, some will be very comfortable with aggregate statistical analysis such as regressions, others will be strong advocates of models with economic micro-foundations, while some others may be more familiar with generic stochastic processes. To meet everyone in a middle point, we structure the presentation of the model uses the indicator dynamics as a focal point. That is, first we explain the mathematical specification that we use to model the aggregate dynamics of the indicators. This consists of a key equation that ties theory and data in an intuitive manner. Next, we develop the micro-foundations that provide the theoretical backbone to the main parameters of this equation. We do it by explaining the behavioural model of the policymaking agents and, then, the heuristic behind the budgetary allocation of the central authority.

4.1.1 Indicator dynamics

In real world data, development indicators exhibit a wide variety of behaviours, for example, positive/negative trends, high/low volatility, non-linearities, periods of continuous/intermittent growth, etc. A flexible framework to model such varied dynamics is stochastic processes. Therefore, we specify a random walk with a drift that can be used to match three key empirical features: (1) positive and negative trends, (2) the actual initial and final values, and (3) empirical probabilities of growth. As we will show, each of these features is associated to a parameter that has an intuitive interpretation or theoretical micro-foundations.

The evolution of indicator I_i follows a random walk determined by (1) a probability of growth and (2) a step size. Every period t, indicator i grows with probability $\gamma_{i,t}$. On the other hand, the indicator decreases with probability $1 - \gamma_{i,t}$. If the indicator grows, it does so by a factor (or step size) α_i. If it decreases, it does by α'_i. Let $\xi_{i,t}$ be the binary outcome (0 or 1) of a random draw with probability $\gamma_{i,t}$. Then, we can write the indicator's random walk as

$$I_{i,t+1} = \begin{cases} I_{i,t} + \alpha_i & \text{if} \quad \xi_{i,t} = 1 \\ I_{i,t} - \alpha'_i & \text{otherwise} \end{cases} \tag{4.1}$$

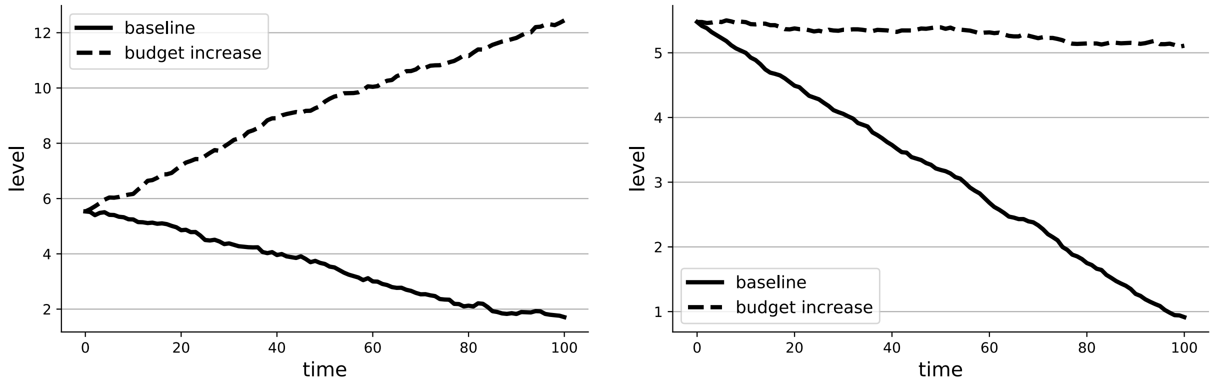
Notice that equation 4.1 has certain peculiarities that complicate mathematical tractability. First, the probability of success $\gamma_{i,t}$ (and hence the outcome $\xi_{i,t}$) is not constant. Thus, the probability of going up or down changes through time. As we will show ahead $\gamma_{i,t}$ provides a link to the micro-foundations of the model by accounting for the dynamic nature of public spending and network spillovers. For now, it is useful to know that, if indicator i receives more funding or if it receives more positive spillovers, $\gamma_{i,t}$ may increase (and so the chances of improving the indicator). A way to interpret the binary outcome $\xi_{i,t}$ is in terms of success and failure of a policy. The more resources or spillovers, the more likely that an existing government programme will succeed. However, such success relates to short/mid-term impacts, not to structural transformations. This is so because the we assume that the existing government programmes remain the same. Thus, a structurally flawed programme may still perform poorly even with substantial financial resources, an idea that we exploit in chapter *** to analyse development bottlenecks.

Next, let us talk about the long term and its structural factors, which we capture through α_i and α_i'. If a policy is successful, the improvement on the indicator is bound by α_i, which captures structural factors. Notice that α_i is constant through time, which means that it has a long-term nature. When an policy fails, the cost in terms of indicator performance is α_i', which also captures structural factors. Parameters α_i and α_i' are a reduced way of accounting for all other factors that contribute to the development of I_i, but that we do not model explicitly in this book, so we have to assume them constant. Nevertheless, due to the flexibility of agent computing, the reader could modify our code in order to account for additional factors, effectively decoupling their effect from α_i and α_i' at the moment of calibration.

The specification in equation 4.1 offers great flexibility to model the empirical data and to generate intuitive counterfactuals. To provide some examples before diving into the model details, assume that γ_i is constant through time. Then depending on the levels of α_i, α_i', and γ_i, we can generate different dynamics with interesting associated counterfactuals. For instance, consider an indicator i that is described by the parameter set $\alpha_i = 0.4$, $\alpha_i' = 0.7$, and $\gamma_i = 0.6$. Generally speaking, this random walk has a negative drift (due to $\alpha_i < \alpha_i'$), even if the policy is successful most of the times. Next, consider another indicator j with similar dynamics (a negative drift) which are governed by parameters $\alpha_j = 0.11$, $\alpha_j' = 0.27$, and $\gamma_j = 0.6$. Figure 4.1 shows the dynamics of both indicators in solid lines. Each line corresponds to the average trajectory of 100 runs. Clearly, both are qualitatively and quantitatively consistent, despite coming from random walks with different structural factors.

The dotted lines in Figure 4.1 show the dynamics of a counterfactual in which the probability of success in both indicators increases from $\gamma = 0.6$ to $\gamma = 0.7$. Suppose that such increment in the success rate of each policy comes from an increment in public expenditure. As shown by the panels, budget increments are highly effective in indicator i (panel 4.1a) as they are able to revert the negative drift into a positive one, despite having $\alpha_i < \alpha_i'$. In panel 4.1b (indicator j), in contrast, we observe ineffectiveness in the sense that, expenditures increments in the same amount as in i are unable to revert the negative trend. Thus, even if there are improvements in the (negative) performance of indicator j, it the intention of an increased expenditure is to revert the negative trend, one could say that such strategy fails. Next, we elaborate on the micro-foundations that generate the probability of success $\gamma_{i,t}$.

Figure 4.1: Error minimisation behaviour



(a) Example where budget is effective

(b) Example where budget is ineffective

Notes: .

4.1.2 Policymaking Agents

Let us consider the n instrumental policy issues. We assume that, for each issue, there is a government program. Furthermore, for each government programme, there is a policymaking agent in charge of implementing it. That is, we assume that the government programs are given and remain the same throughout a simulation. Therefore, our modelling emphasises how public expenditure is transformed into policy outcomes. The role of each policymaking agent consists of receiving resources from the central authority and transforming them into successful policy.[1] More formally, let $P_{1,t}, \ldots, P_{n,t}$ represent a vector that characterises the allocation of resources across the n instrumental policy issues in period t. We also call this vector the *allocation profile*, and its particular configuration is determined by the central authority. The time sub-index denotes the fact that the government may adjust these allocations dynamically, something that we explain in more detail in section 4.1.3.

Each policymaking agent i is in charge of their allocation $P_{i,t}$. While their job is to use $P_{i,t}$ towards the implementation of the corresponding government programme, an agent may have incentives to be inefficient for several reasons. For example, the agent may have career aspirations in the private sector, so favouring specific contractors through manipulated public tenders may be a strategy that is their interest but not in the

[1] Notice that, while public spending may be efficiently used in a government programme, the outcome of the programme may still be poor as its design and operation could be flawed; in which case we say that the program needs to be redesigned due to structural factors that can only addressed in the long-term (which are captured in α_i and α'_i).

public one. Alternatively, the embezzlement of public funds (for the agent or other purposes such as political campaigns) is a well known case of the inefficiencies that may arise when the incentives of the principal do not align with the agent's ones. Such misalignment of incentives gives place to the classic principal-agent problem ***cite; something pervasive in public administrations around the world (in both developing and developed countries). Formally, we say that, out of $P_{i,t}$, agent i uses only $C_{i,t} \leq P_{i,t}$ towards the government programme. Thus, we call $C_{i,t}$ the *contribution* of the agent and $P_{i,t} - C_{i,t}$ the level of inefficiency.[2]

The policymaking agents face a trade-off. On the one hand, being a proficient public servant contributes to certain political status if such proficiency is adequately reflected in the policy outcomes. On the other hand, a private gain from being inefficient (like the examples mentioned above) may steer the agent towards a lower contribution. In order to model this trade-off, let us specify the agent's benefit function

$$F_{i,t+1} = \underbrace{\Delta I^*_{i,t} \frac{C_{i,t}}{P_{i,t}}}_{\text{proficiency}} + \underbrace{(1 - \theta_{i,t}\tau_{i,t}) \frac{(P_{i,t} - C_{i,t})}{P_{i,t}}}_{\text{inefficiency}} . \tag{4.2}$$

Equation 4.2 describes the benefits (or utility) F_i received by policymaking agent i. The equation is structured according to the agent's trade-off: a summand representing the benefits from proficiency and a another one capturing the personal gain from being inefficient. These terms are weighted according to the contribution $C_{i,t}$ and the inefficiency $P_{i,t} - C_{i,t}$ as a proportion of the allocation $P_{i,t}$. Proficiency is praised with political status which, in turn, is signalled by the improvement of indicator $I_{i,t}$ with respect to the previous period. Such change, however, is relative to the progress made by the other policymaking agents; representing the importance of standing out to gain political status. The relative change of the indicator that brings political status to the agent is

$$\Delta I^*_{i,t} = \frac{I_{i,t} - I_{i,t-1}}{\sum_j I_{j,t} - I_{j,t-1}} . \tag{4.3}$$

Coming back to equation 4.2, in its second addend, we can see that the benefit from an inefficiency is damped by $(1 - \theta_{i,t}\tau_t)$, which is a factor related to public procurement. Parameter $\tau_{i,t}$ captures the penalty incurred when an inefficiency is spotted, which we interpret as the *quality of the rule of law*.[3] For simplicity, we assume $\tau_{i,t} = \tau$ in most of the chapters.

The size of the penalty, however, is only half of the story when we speak of public procurement. The other half is the reduction of opportunities (of being inefficient) through monitoring. This is modelled through the

[2] ***A note on the nature of inefficiencies and PPI not being able to disentangle them.

[3] This parameter can be specific to each indicator as there may be varying procurement cultures across different sectors or professions (such as a disbarment or a permanent ban from running for public office). In its simplest form, $\tau_{i,t}$ could be assumed homogeneous and constant (so $\tau_{i,t} = \tau$ for all indicators and periods).[4] Alternatively, $\tau_{i,t}$ can be dynamic, as part of one of the indicators, something done by ***citeCorruptionPaper and covered in chapter ***.

binary variable $\theta_{i,t}$, which takes the value one when monitoring has been successful in spotting an inefficiency, and zero otherwise. The probability of getting $\theta_{i,t} = 1$ depends on the *quality of monitoring* and on the relative size of the inefficiency. Quality of monitoring is an institutional variable $\varphi_{i,t}$ that interacts with the size of the inefficiency in order to produce the probability

$$\lambda_{i,t} = \varphi_{i,t} \frac{P_{i,t} - C_{i,t}}{P_t^*}, \tag{4.4}$$

where $P_t^* = \max(P_{1,t}, \ldots, P_{n,t})$, to make the inefficiencies relative. By using a relative terms we account for the fact that inefficiencies are often the result of social norms. For example, a society where corruption if prevalent and, to some degree, accepted, will not only have lower governance parameters, but will also require an agent to stand out of the social norm in order to be spotted. This social-norm component is an important departure from the traditional economic approach to the principal-agent problem, and tries to conciliate this agent-centric perspective with the sociological view of social norms ***cite.

Now that we have explained the benefits of the policymaking agents, we need to describe the behavioural model through which they determine the contribution levels that are most beneficial to them. In this model, as in the real world, agents are rationally bounded and face substantial uncertainty due to the complexity of the environment. Hence our modelling choice for the behavioural component is *reinforcement learning*. In particular, we use a specification called *directed learning*, in which the agents either increase of decrease their contributions, depending on the change in benefits with respect to the previous period and the direction of the agent's action. The intuition behind this model is straightforward: if the last action increased the contribution, and this was followed by an increment in benefits, then the agent will increase their contribution the next period. If, on the contrary, increased contributions are followed by reduced benefits, then the agent will *adapt* and change its actions in order to decrease the contribution the next period.

The actions that an agent may take to increase or decrease their contribution may be different and of various types, as they depend of the nature of the inefficiencies involved. To simplify this complexity, we model all possible actions as a variable $X_{i,t+1}$ that can take any value in the line of the real numbers. An increment $X_{i,t+1} > X_{i,t}$ means that the agent increases their contribution, while $X_{i,t+1} < X_{i,t}$ indicates a reduction. For modelling purposes, we are interested in the updating rule of X_i under directed learning. Such a rule can be formalised as

$$X_{i,t+1} = \underbrace{X_{i,t}}_{\text{inertial term}} + \underbrace{\text{sgn}((X_{i,t} - X_{i,t-1})(F_{i,t} - F_{i,t-1}))}_{\text{direction}} \underbrace{|F_{i,t} - F_{i,t-1}|}_{\text{adaptation size}}, \tag{4.5}$$

where sgn$(\dot{})$ corresponds to the sign function, which captures the main intuition of directed learning. Factor

$|F_{i,t} - F_{i,t-1}|$, on the other hand, accounts for the size of the change that the agent takes towards their updated actions.

Now that we have established how the policymaking agents adapt, the final step is to map their actions into their contributions. We do this through the logistic function

$$C_{i,t} = \frac{P_{i,t}}{1 + e^{-X_{i,t}}}, \tag{4.6}$$

which is centred around zero and has symmetry between positive and negative actions. This function guarantees that a contribution cannot be larger than the resources allocated by the central authority, so the problem of the policymaking agent is to choose a contribution level in which it is beneficial to maintain a certain balance between proficiency and inefficiency; and this may be influenced by institutional, behavioural, and social mechanisms. Chapter 5 elaborates on the internal validation of this behavioural component.

4.1.3 Central Authority

Now we shift our attention to the government agent or central authority, who has to determine the allocation profile $P_{1,t}, \dots, P_{n,t}$. To speak of allocations, first we need to introduce the total budget B. The modelling of the central authority and the degree to which it is relevant to a particular application relies, partly, on the availability of expenditure data. Let us first discuss the ideal case in which there is open spending data for each indicator and period, and progressively introduce modelling assumptions as we restrict the availability of data more and more. This should provide a comprehensive picture of the flexibility of our framework, as it can accommodate applications with various degrees of data quality and institutional nuances.

First, suppose that we have highly disaggregated data on how much is spent every period in each government program. For T periods of time, this means that we have a *disbursement schedule*

$$\mathbb{B} = \begin{bmatrix} P_{1,1} & P_{1,2} & \dots & P_{1,t} & \dots & P_{1,T} \\ P_{2,1} & P_{2,2} & \dots & P_{1,t} & \dots & P_{2,T} \\ \vdots & \vdots & \dots & \vdots & \dots & \vdots \\ P_{i,1} & P_{i,2} & \dots & P_{i,t} & \dots & P_{i,T} \\ \vdots & \vdots & \dots & \vdots & \dots & \vdots \\ P_{n,1} & P_{n,2} & \dots & P_{n,t} & \dots & P_{n,T} \end{bmatrix}. \tag{4.7}$$

Essentially, equation 4.7 provides a complete mapping between public expenditure and government programs through time; the ideal dataset for any analyst. Each row of this matrix is a time series describing budgetary readjustments in each policy issue. If such data exists, then our concerns regarding the various motivations of such readjustments would be minor, as the model would just take these data as one of its inputs, and simulate the policymaking agents together with the indicator dynamics. Unfortunately, as we write this book, such a complete mapping does not exist, at lest not in a way that can unambiguously link each expenditure program to a development indicator. For this reason, we need to resort to socioeconomic theory and fill these data gaps.

Next, suppose we have a partial mapping between the expenditure programmes and the indicators. One such mapping, for example, would be expenditure data classified into broad development topics such as national budget tranches or the Sustainable Development Goals. In this case, one has a the time series of public expenditure at the level of the tranches or the SDGs, and uncertainty about how exactly these funds are distributed across the different expenditure programs that live within each tranche or SDG. For instance $B_{k,t}$ would describe the total expenditure destined to tranche k in period t, and a set $\mathbf{\Omega}_k$ of expenditure programmes that would potentially benefit from such resources.

Then, the natural question is: how is $B_{k,t}$ distributed across its government programs $\mathbf{\Omega}_k$? To answer this, one needs ex-ante knowledge about the heuristics or rules that a particular government follows when readjusting its budgets. For example, in the real world, a common practice is to allocate more resources to those policy issues that are most laggard, as they may represent bottlenecks that need urgent attention (a practice commonly promoted during through the Millennium Development Goals project ***cite). Another criterion relates to public governance, and to the principle that the central authority tends to reward those agencies of public servants who show good performance (through their indicators) or, alternatively, penalise the inefficient ones by withdrawing funds. Due to its computational nature, our model allows for any specification that the user wishes to implement. Here, and due to the various applications that we present throughout the book, we implement a heuristic that brings together the two principles above mentioned, in addition to one principle coming from the political science literature: the stochastic punctuated equilibrium approach ***cite.

Let $q_{i,t}$ denote the propensity of the central authority to spend in policy issue i in period t. The evolution of the propensities is given by

$$q_{i,t} = \underbrace{q_{i,t-1}}_{\text{inertial term}} + \underbrace{U(0,1)}_{\text{stochastic term}} \underbrace{\left\{\max\left(1, \sum_{h}^{t-1} \theta_{i,h}\right)\right\}^{-1} \sum_{h|\theta_{i,h}=1}^{t-1} \frac{P_{i,h} - C_{i,h}}{P_{i,h}}}_{\text{governance component}}, \tag{4.8}$$

which combines the three heuristics mentioned above. The term $q_{i,t-1}$ captures the historical inertia of the propensity. The stochastic term is represented by $U(0,1)$, which is an independent realisation of a random variable with a uniform distribution. The remaining terms represent the penalty for a history of spotted inefficiencies in policy issue i so far (until period $h \leq t$). The laggard-indicator prioritisation component enters through the initial conditions of $q_{i,0}$ in the form of gap closure as

$$q_{i,0} = \max[(G_i - I_{i,0}), 0], \tag{4.9}$$

where G_i corresponds to the development goal of policy issue i. The gap closure criterion relies on having information about the goals of each indicator. If no information is available, then the model sets random initial conditions for the propensities.

Next, let us account for structural and idiosyncratic factors that shape the government's expenditure decisions. These factors are represented in a vector $b_1, \dots, b_n$ that modulates the expenditure propensities. These are free parameters that, if calibrated, can be used to replicate empirical expenditure patterns,[5] or to study the impact of changes in the allocation profile when no disaggregate information is available.[6] Usually, we assume $b_i = 1$ for every instrumental policy issue, but it is important to present the full model specification here for future reference. With these factors, we construct the modulated propensities

$$\dot{q}_{i,t} = \left(\frac{q_{i,t}}{\sum_j q_{j,t}} \right)^{b_i}. \tag{4.10}$$

Finally, in order to determine the specific allocations that each expenditure programme receives, we employ the function

$$P_{i \in \mathbf{\Omega}_k, t} = B_{k,t} \frac{\dot{q}_{i \in \mathbf{\Omega}_k, t}}{\sum_j \dot{q}_{j \in \mathbf{\Omega}_k, t}}. \tag{4.11}$$

Through equation 4.11, it is easy to see how this model can be adapted to various data-availability situations. For example, suppose that no data on budgetary tranches exist. In that case the allocation profile can be fully determined using the total budget such that $P_{i,t} = B_t \dot{q}_{i,t} / \sum_j \dot{q}_{j,t}$. Alternatively, it may be the case that there is no inter-temporal budgetary information, in which case the allocation profile is determined by $P_{i,t} = B \dot{q}_{i,t} / \sum_j \dot{q}_{j,t}$, where B is an expenditure amount that does not change across time. Furthermore, it could be the case that no budgetary information exists. Then, one can simply set B to be an arbitrary number such as 1. This would limit cross-country comparisons, but still allow a within country analysis. Of course, the less budgetary data available, the weaker the inferences that can be obtained, and the more limited the interpretations and comparisons that can be made. In fact, when we first started this research programme, we were forced to work under the assumption of $B = 1$ ***cite. Fortunately, the data revolution and the open-gov movement have contributed in recent years to the availability of this

[5] ***a footnote here

[6] ***as shown in Chapter ***.

information, so this book will show applications that use budgetary data with various degrees of granularity.

A final note before proceeding to the next section is that the model cal accommodate other types of budgetary nuances. For example, it could be the case that one expenditure programme receives funds from multiple tranches. For instance, in the SDGs official database ***cite, there are some indicators that are classified into multiple SDGs ***example. Computationally-speaking, accommodating this feature is trivial, while mathematically it requires equation 4.11 to be changed into

$$P_{i\in\mathbf{\Omega}_k,t} = B_{k,t}\frac{a_{i,k}\dot{q}_{i\in\mathbf{\Omega}_k,t}}{\sum_j a_{j,k}\dot{q}_{j\in\mathbf{\Omega}_k,t}}, \tag{4.12}$$

where $a_{i,k}$ indicates the weight that tranche k represents in the budget of programme i. Of course, this requires prior knowledge about the relative proximity that government programmes have to the different tranches of SDGs. At the cost of double counting expenditure, one can simply assume $a_{i,k} = 1$ in order to account for this feature. As we will show in the next section, we do not think that this double counting is a big problem since the free parameters of the model can counterbalance this effect after calibrating them.

4.1.4 Policy Success

Finally, let us explain the last piece of the model; the one that connects the micro-level mechanisms to the macro-level dynamics: the probability of success $\gamma_{i,t}$. This is an endogenous dynamic variable that depends on (1) the resources that go to the government programme, and (2) the spillovers received from other policy issues (which could be instrumental as well as collateral). The probability of success of government programme i is determined by

$$\gamma_{i,t} = \underbrace{\beta_i}_{\text{expenditure returns}} \times \left[\underbrace{C_{i,t}}_{\text{contribution}} + \underbrace{\frac{1}{1+B_t}\sum_j C_{j,t}}_{\text{systemic efficiency}} \right] \times \underbrace{\left(1+e^{-S_{i,t}}\right)^{-1}}_{\text{spillovers}}. \tag{4.13}$$

The probability of success of a government programme depends, first, on the contribution of the policy-making agent. Therefore, if the public governance of an economy is poor, the impact of an allocation $P_{i,t}$ with be limited because the associated agent will have incentives to set low contribution levels. In the case of a collateral indicator $C_{i,t} = 0$ all the time. The overall level of efficiency (the ratio of total contributions to the budget) is also relevant for the probability of success as it connects the performance of the indicators to the general 'financial health' of the system (which also affects collateral indicators since government expenditure is a important component of development in general ***cite). The number one in the denominator $1 + B_t$ makes sure that the efficiency term is defined, even under the complete absence of public funds.

In order to normalise the budgetary information in such way that $\gamma_{i,t} \in [0,1]$, we introduce the free parameter β_i. One way to interpret this parameter is, partly, as returns to expenditure. For example, suppose that two indicators, i and j, exhibit the same performance, but i receives 100 times more resources than j. When calibrating their respective βs, we would obtain that β_j is, approximately, two orders of magnitude larger than β_i. Assuming all else constant, one would interpret that each dollar spent in β_j pays back 100 time than in β_j. This is something normal in real-world policymaking, as there exist topics that, inherently, require substantially more resources than others (e.g., government programmes related to public infrastructure).

Finally, we are left with the spillover term $\left(1 + e^{-S_{i,t}}\right)^{-1}$, where $S_{i,t}$ corresponds to the total amount of spillovers (positive and negative) that policy issue i receives in period t. These incoming spillovers are computed every period using the adjacency matrix $\mathbb{A}$ (of size $N \times N$), where entry $\mathbb{A}_{j,i}$ denotes the spillovers that, if realised, policy issue j sends to i. Thus, $S_{i,t} = \sum_j \mathbf{1}_{j,t} \mathbb{A}_{j,i}$, and $\mathbf{1}_{j,t}$ is the indicator function: 1 if indicator j grew in the previous period and 0 otherwise.

As we have discussed in Chapter 3, the network in $\mathbb{A}$ does not represent causal relations between the indicators. Instead it captures a stylised fact of co-movements that can be taken into account by the model. Notice that the spillovers are short-term events that realise from the growth dynamics of other indicators. Hence, and consistent with our previous argument, the network captures long-term conditional dependencies that allow for short-term spillovers to take place. Empirically speaking, matrix $\mathbb{A}$ is an exogenous input of the model, and it can be constructed in various ways (quantitative and qualitative). We elaborate on our method of choice for estimating $\mathbb{A}$ in Chapter 5.

4.1.5 Summary

As a summary, Table 4.1 presents all the variables and parameters of the model. Note that the only three free parameters that are necessary are α, α', and β, and that they capture meso and macro features. The micro-level variables, on the other hands, are all endogenous. This is a very attractive feature of this model because the user does not need to worry about finding micro-level data to parameterize the model, which is a major limitation when trying to operationalise behaviourally driven models. In Chapter 5 we show that this endogeneity of the behavioural mechanism is not trivial, since it generates intuitive and stable results.

In Figure 4.2, we provide a diagrammatic depiction of the model, starting at the bottom–with the expenditure allocations–and ending at the top–with the indicator dynamics. First, once the budget has been allocated, the dynamics take place in a decentralised yet interactive fashion. This is fundamentally different from traditional economic models in which homogeneous agents respond to a centralised price vector,

Table 4.1: Variables of the model

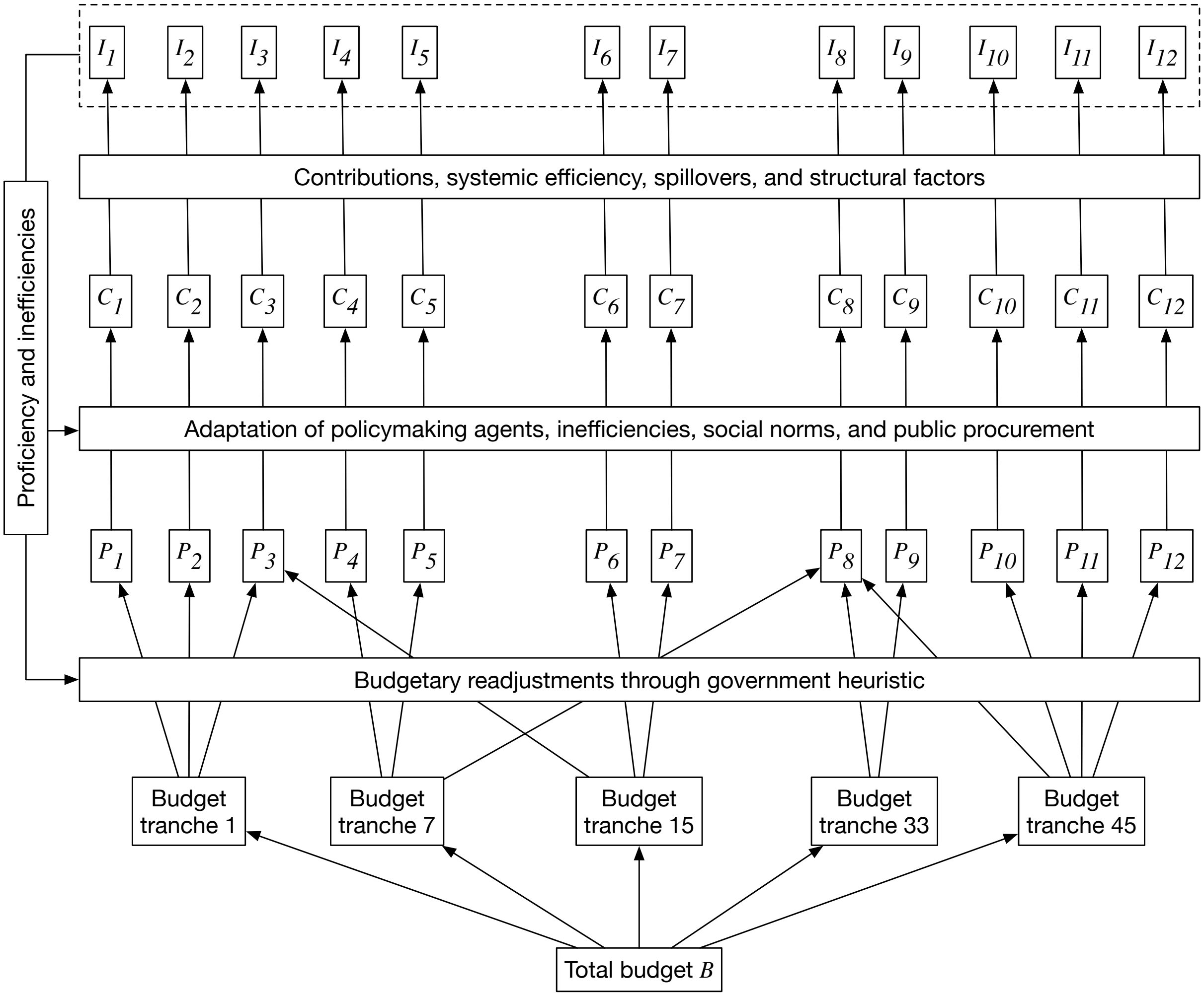
Symbol	Variable	Source	Type
α_i	limiting structural factors	calibrated	exogenous
α'_i	structural costs	calibrated	
β_i	expenditure returns	calibrated	
b_i	expenditure modulation parameter	imputed or calibrated*	
$I_{i,0}$	indicator initial level	indicators' time series	
$I_{i,-1}$	indicator final level	indicators' time series	
$\mathbb{B}$	disbursement schedule	open-spending data*	
$\mathbb{A}$	spillover network adjacency matrix	indicators' time series*	
$\varphi_{i,t}$	quality of monitoring	worldwide governance indicators**	
$\tau_{i,t}$	quality of the rule of law	worldwide governance indicators**	
$G_{i,t}$	development goal	development plans/documents*	
$X_{i,t}$	agent actions	equation 4.5	endogenous
$F_{i,t}$	agent benefits	equation 4.2	
$C_{i,t}$	agent contributions	equation 4.2	
$\lambda_{i,t}$	probability of spotting inefficiencies	equation 4.4	
$\theta_{i,t}$	binary outcome of monitoring	equations 4.2 & 4.4	
$\gamma_{i,t}$	probability of successful growth	equation 4.13	
$\xi(\gamma_{i,t})$	binary outcome of random growth process	equations 4.13 & 4.1	
$q_{i,t}$	propensity to spend	equation 4.8	
$q_{i,0}$	initial propensity to spend	equation 4.9	
$\dot{q}_{i,t}$	modulated propensity to spend	equation 4.10	
$S_{i,t}$	net incoming spillovers	equation 4.13	
$P_{i,t}$	government allocation	equation 4.11	
$I_{i,t}$	indicator level	equation 4.1	

Notes: * data on these parameters are optional, at the cost of obtaining weaker inferences. ** these parameters can also be endogenous if their respective indicators are part of the policy space on which the government allocates resources.

avoiding interactions or sociological considerations such social norms. Throughout the model bottom-up flow, we can see how different behavioural and network elements contribute to the indicators. Then, these aggregate dynamics feedback from top to bottom, generating multi-scale feedback loops. One of the virtues of specifying agent-level behaviour is that inferences can be made at both the macro and micro-level, which can be useful to validate the model in ways that is not possible with more traditional approaches. Such validation strengthen the inferences derived from intervention experiments in the sense that, in such experiments, the outcomes partially reflect more nuances regarding the adaptation of the agents to an intervention (not just the change of an aggregate variable).

Finally, in Algorithm 1, the reader can see the procedural logic of the model in the form of pseudocode.

Figure 4.2: Bottom-up and top-down causal links between budget and indicators


I_1 I_2 I_3 I_4 I_5 I_6 I_7 I_8 I_9 I_{10} I_{11} I_{12}
Contributions, systemic efficiency, spillovers, and structural factors
C_1 C_2 C_3 C_4 C_5 C_6 C_7 C_8 C_9 C_{10} C_{11} C_{12}
Proficiency and inefficiencies
Adaptation of policymaking agents, inefficiencies, social norms, and public procurement
P_1 P_2 P_3 P_4 P_5 P_6 P_7 P_8 P_9 P_{10} P_{11} P_{12}
Budgetary readjustments through government heuristic
Budget tranche 1
Budget tranche 7
Budget tranche 15
Budget tranche 33
Budget tranche 45
Total budget B

Notes: the notes.
Sources: the sources.

Algorithm 1: Model pseudocode

```
foreach period $t$ do
    foreach public servant $i$ do
        receive public funds $P_{i,t}$;
        evaluate the benefits from the previous contribution $C_{i,t-1}$;
        establish new contribution level $C_{i,t}$;
    foreach indicator $i$ do
        if the indicator is instrumental then
            implement public policy using the resources $C_{i,t}$;
        receive the incoming spillovers $S_{i,t}$;
        determine the probability of success $\gamma_{i,t}$ according to $C_{i,t}$ and $S_{i,t}$;
        if the public policy is successful (with probability $\gamma_{i,t}$) then
            improve the indicator according to the long-term structural factors $\alpha_i$;
        else
            worsen the indicator according to the long-term structural costs $\alpha'_i$;
    the government monitors the policymakers through imperfect mechanisms;
    the government penalises those who are found being inefficient;
    the policymakers receive the benefit from their chosen contributions;
    the government updates the allocation profile $P_{1,t}, \ldots, P_{n,t}$;
```

Chapter 5

Calibration and Validation

Now that we have discussed the model in detail, and its theoretical foundations, we move to more practical matters regrading its usability with real-world data. As we have mentioned previously, the model has been designed such that the micro-level behavioural variables emerge from the interactions and learning of the agents. Hence, the only type of parameters that the user of the model has to be concerned about are macro and meso-level ones, namely, vectors $\boldsymbol{\alpha}$, $\boldsymbol{\alpha}'$, and $\boldsymbol{\beta}$. In this section, we explain how to calculate those parameters through an efficient algorithm and, then, we demonstrate how–once calibrated–the model can be validated in various ways.

5.1 Calibration strategy

Let us discuss the task of finding the model's free parameters. On total, there are $3N$ parameters that need to be determined (3 per indicator). These are $\boldsymbol{\alpha} = \alpha_1, \ldots, \alpha_N$, $\boldsymbol{\alpha}' = \alpha'_1, \ldots, \alpha'_N$, and $\boldsymbol{\beta} = \beta_1, \ldots, \beta_N$. The objective function for calibrating $\boldsymbol{\alpha}$ and $\boldsymbol{\alpha}'$ is to minimise the difference between the average final value of the empirical indicators and the one of the simulated ones. For $\boldsymbol{\beta}$, on the other hand, we seek to minimise the difference between the empirical success rate of the indicators (the number of times positive growth is observed as a rate of the total number of periods) and the average success probability $\gamma_{i,t}$. While both objective functions seem straightforward, there is more than meets the eye.

One of the challenges of calibrating the model is that, due to the interdependencies caused by the agents and the spillover network (see Figure 4.2), the dynamics of an indicator are sensitive to evolution of the other other indicators. For instance, suppose we increase α_i and β_i while keeping all other parameters constant. If another indicator j is somehow linked to i its dynamics will be impacted too, for example, they would accelerate if i sends positive spillovers to j, meaning that we would need to readjust α_j and

β_j in the opposite direction; but this would also produce a dis-adjustment in other indicators that may be linked to j (or even in i). In earlier versions of the model, where we only had to calibrate $\boldsymbol{\alpha}$, we developed calibration algorithms that would assume *ceteris paribus* conditions (Guerrero and Castañeda, 2020a) so only one parameter could be evaluated at a time. Once we augmented the parameter space by including $\boldsymbol{\beta}$, this *ceteris paribus* strategy became computationally unfeasible. In addition, the interdependencies in the model produce a dynamic fitness landscape, that is difficult to handle through linear-programming and convex-optimisation frameworks.

In sight of these limitations, a heuristic optimisation approach is necessary. However, there are two additional problems: the computational cost and the loss of information. First, each evaluation of the model is computationally expensive because it requires several Monte Carlo simulations. For example, suppose that we are trying to calibrate $\boldsymbol{\alpha}$. A single simulation may yield a final value for indicator i that is close to the empirical one, but another one may generate a very different one. This is entirely possible due to the stochastic components of the model and to the potential presence of path-dependence created by learning and social norms; something quite common in complex systems ***citeExamples. Second, a problem with a reduced metric such as an overall average error is that information about each indicator's error is destroyed, and optimisation algorithms lose sensitivity to changes in indicator-specific parameters. For these reasons, several types of optimisation algorithms (e.g. evolutionary computing, swarm optimisation, simulated annealing, etc.) are unfit for purpose.

When it comes to optimisation algorithms that handle computationally intensive evaluations, we find different Bayesian strategies (e.g. the tree-structured Parzen estimator). From our experience, these methods do not perform well with this model, perhaps due to the large number of parameters and the rugosity of the fitness landscape. However in Guerrero and Castañeda (2021b), we finally developed an algorithm that works particularly well for this model. First, it uses a multi-objective function, which prevents the loss of indicator-specific error information, maintaining sensitivity to each individual parameter. Second, it readjusts the parameters simultaneously, so it is much more efficient than the *ceteris paribus* approach taken in Guerrero and Castañeda (2020a). Third, it uses a normalised gradient-descend rule to update each parameter, which contributes to its efficiency. Fourth, it allows introducing hyper-parameters to improve its efficiency. This algorithm scales well with the number of parameters (i.e. of indicators or policy dimensions). Furthermore, with enough Monte Carlo simulations per evaluation, the method achieves high precision levels, which allows a high degree of control on the error minimisation. Here we provide all the details.

5.2 Optimisation algorithm

The algorithm that we present here is an extension of the one from Guerrero and Castañeda (2021b) as, in this version of the model, we have to account for the vector $\boldsymbol{\alpha}'$ that enables downward dynamics. Let M denote a given number of independent Monte Carlo simulations; $I_{i,-1}$ is the empirical final value of indicator i; and $\hat{I}_{i,-1,m}$ its simulated final value in the m^{th} model run. The expected final value of a simulated indicator i is

$$\bar{I}_{i,-1} = \frac{1}{M}\sum_{m=1}^{M}\hat{I}_{i,-1,m}. \tag{5.1}$$

The α-error of indicator i is

$$e_{\alpha_i} = I_{i,-1} - \bar{I}_{i,-1}, \tag{5.2}$$

which is the same for α_i and α'_i.

Next, for an empirical indicator i, its change from period $t-1$ to t is

$$\Delta I_{i,t} = I_{i,t} - I_{i,t-1}. \tag{5.3}$$

Then,i's success rate is the number of times that it exhibits a positive change between two consecutive periods, divided by the number of changes, as described by

$$r_i = \frac{1}{T-1}\sum_{t=2}^{T}\mathbf{1}(\Delta I_{i,t}), \tag{5.4}$$

where $\mathbf{1} = 1$ if $\Delta I_{i,t} > 0$ and $\mathbf{1} = 0$ otherwise.

Next, the β-error of indicator i is

$$e_{\beta_i} = r_i - \bar{\gamma}_i, \tag{5.5}$$

where $\bar{\gamma}_i$ is the average success probability of the model, computed as

$$\bar{\gamma}_i = \frac{1}{TM}\sum_{t=1}^{T}\sum_{m=1}^{M}\gamma_{i,t,m}. \tag{5.6}$$

Now, let us define the normalised α-error as

$$\hat{e}_{\alpha_i} = \frac{e_{\alpha_i}}{|I_{i,-1} - I_{i,0}|}, \tag{5.7}$$

where $I_{i,0}$ is the empirical initial value of the indicator, so $I_{i,-1} - I_{i,0}$ represents the gap that was closed during the sampling period. The intuition behind normalising the α-error is that simulating indicators that closed bigger gaps (in the same amount of time) introduces more volatility, so the normalisation helps obtaining more stability in a gradient descent. This is not the case for the β-error, so it is not necessary to normalise it.

Note that both $\hat{e}_{\alpha_i}$ and e_{β_i} can be positive or negative. This is intentional as we exploit this feature to direct the gradient descent. The descent procedure seeks to readjust the relevant parameters in incrementally smaller magnitudes. For the normalised α-error, the readjustment rule of the associated parameter is

$$\alpha_i = \begin{cases} \alpha_i \times \min(1 + |\hat{e}_{\alpha_i}|, 1.5), & \hat{e}_{\alpha_i} \geq 0 \quad \text{and} \quad I_{i,-1} \geq I_{i,0} \\ \alpha_i \times \max(0.99 - |\hat{e}_{\alpha_i}|, 0.25), & \hat{e}_{\alpha_i} < 0 \quad \text{and} \quad I_{i,-1} > I_{i,0} \\ \alpha_i \times \max(0.99 - |\hat{e}_{\alpha_i}|, 0.25), & \hat{e}_{\alpha_i} > 0 \quad \text{and} \quad I_{i,-1} < I_{i,0} \\ \alpha_i \times \min(1 + |\hat{e}_{\alpha_i}|, 1.5), & \hat{e}_{\alpha_i} \leq 0 \quad \text{and} \quad I_{i,-1} \leq I_{i,0} \end{cases}. \tag{5.8}$$

There are four readjustment cases because the direction of the adaptation depends on the sign of the development gap. In a similar way, we define the readjustment of α'_i as

$$\alpha'_i = \begin{cases} \alpha'_i \times \max(0.99 - |\hat{e}_{\alpha_i}|, 0.25), & \hat{e}_{\alpha_i} > 0 \quad \text{and} \quad I_{i,-1} > I_{i,0} \\ \alpha'_i \times \min(1 + |\hat{e}_{\alpha_i}|, 1.5), & \hat{e}_{\alpha_i} \leq 0 \quad \text{and} \quad I_{i,-1} \geq I_{i,0} \\ \alpha'_i \times \min(1 + |\hat{e}_{\alpha_i}|, 1.5), & \hat{e}_{\alpha_i} \geq 0 \quad \text{and} \quad I_{i,-1} \leq I_{i,0} \\ \alpha'_i \times \max(0.99 - |\hat{e}_{\alpha_i}|, 0.25), & \hat{e}_{\alpha_i} < 0 \quad \text{and} \quad I_{i,-1} < I_{i,0} \end{cases}. \tag{5.9}$$

and for the β-error is

$$\beta_i = \begin{cases} \beta_i \times \min(1 + |e_{\beta_i}|, 1.5), & e_{\beta_i} \geq 0 \\ \beta_i \times \max(0.99 - |e_{\beta_i}|, 0.25), & e_{\beta_i} < 0 \end{cases}, \tag{5.10}$$

which does not depend on the sign of the development gap.

The principle behind these readjustment rules is twofold: to penalise deviations and to adapt the penalty size as the error shrinks. For instance, for the case of a positive trend, $\hat{e}_{\alpha_i} > 0$ means that the simulated indicator was slower than the empirical one since it ended at a lower value. Therefore, the adjustment is to increase α_i by a fraction not larger than 1/2. As this process continues, the fraction becomes lower than 0.5 because the error decreases, so the size of the readjustment is $|\hat{e}_{\alpha_i}|$.

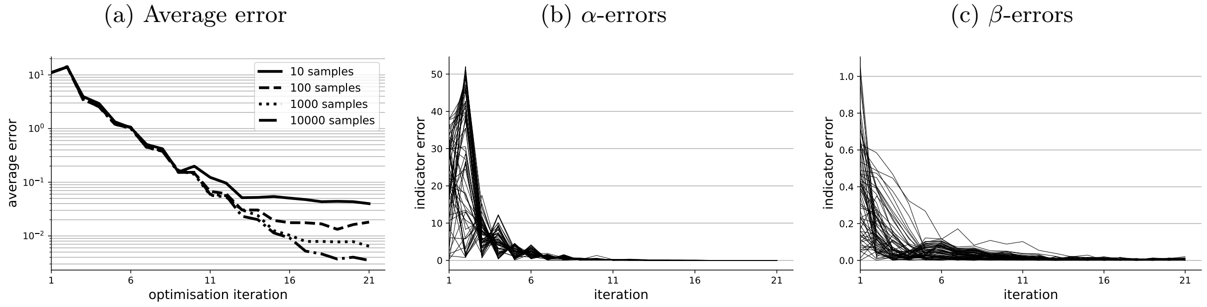
Putting together these elements, we construct an optimisation algorithm that iterates until a tolerance threshold is reached. We provide its pseudocode in Algorithm 2.

Algorithm 2: Calibration pseudocode

```
initialise vectors $\boldsymbol{\alpha}$, $\boldsymbol{\alpha}'$, and $\boldsymbol{\beta}$ with random values;
while a tolerance threshold is not met do
    run $M$ Monte Carlo simulations;
    compute the errors $\hat{e}_{\alpha_1}, \dots, \hat{e}_{\alpha_N}$ and $e_{\beta_1}, \dots, e_{\beta_N}$;
    foreach indicator $i$ do
        adapt parameters according to equations 5.8, 5.9, and 5.10;
```

The threshold criterion is a choice of the user. From our research, we have found that a criterion that achieves a high goodness of fit is to stop the calibration once the worse performing parameter achieves a minimum goodness of fit according to the metrics defined in section 5.3. The computational cost is partly determined by M, the number of Monte Carlo simulations to be run in an evaluation. How many simulations should one run to calibrate the model? It depends on how conservative one is with regard to the average error threshold. The stricter the threshold, the more precision is required; and more precision demands a larger M in order to obtain more stable moments. In other words, more Monte Carlo simulations ensure more stability in the resulting distributions and their respective moments. Figure 5.1a confirms this by showing how, with more simulations per evaluation, it is possible to achieve lower average errors. At the indicator level, we show the dynamics of minimising the α- and β-errors in Figures 5.1b and 5.1c respectively. Notice how, in both cases, the error of a specific indicator may jump back into a higher level after a few iterations. This is due to the interdependency issue previously discussed, in which adjusting the parameters of one indicator may affect another. Nevertheless, we can see that, as the algorithms iterates further, all the indicator-specific errors go down.

Figure 5.1: Error minimisation behaviour



(a) Average error

(b) α-errors

(c) β-errors

Notes: These illustrative calibrations were performed for the case of Mexico.
Source: Authors' own calculations with data from the 2021 Sustainable Development Report.

Increasing the number of Monte Carlo simulations help achieving lower errors, at the expense of higher computational costs. To reduce mitigate this cost, we introduce three hyperparameters and a routine that allows to set M automatically as the optimisation proceeds. The idea is that, during the first optimisation rounds, the errors are very large, so a small M is enough to generate coherent responses when readjusting the parameters. Hence, the routine consists of starting with few Monte Carlo simulations and, then, start increasing them after a certain number of optimisation rounds. From our experience, an initial $M = 10$ for 20 iterations are enough to drop the average error substantially. The number of iterations that use a low M is the first hyperparameter. Then, the routine increases M periodically. We have found that increments of 10 every 20 iterations is a good balance between error reduction and computational cost for the applications that we present in this book. The size of the increments and the frequency with which they are applied are the second and third hyperparameters respectively. Here, we have determined the values of the three hyperparameters by experience and trial and error. In our applications, this has been enough to calibrate the model for a country in a few seconds, however, one could also design more sophisticated routines to optimise the hyperparameters, something that we leave to the computationally savvy reader.

5.3 Goodness of fit

We have seen that our calibration procedure is effective and efficient in minimising the different types of errors in the model. However, how good is this optimisation? Any quantitative method requires a goodness of fit or accuracy metric in order to assess this question. Often, the construction of such metric obeys to particular characteristics of the problem at hand, for example, the R^2 is popular in linear regressions to get a sense of how much variance is explained by a model, while the ratio of correct predictions to input samples is widely used to assess the accuracy of different non-regression machine-learning algorithms. In the case of our model, it is necessary to construct a goodness-of-fit metric that is coherent with our definitions of error. Here, we introduce such a metric and present some results from calibrating the model for all the countries in the SDR dataset.

Let Ψ_{α_i} denote the goodness of fit of parameter α_i (or α'_i). Following the error notation, we define the goodness of fit of this parameter as

$$\Psi_{\alpha_i} = 1 - \frac{e_{\alpha_i}}{|I_{i,-1} - I_{i,0}|}, \tag{5.11}$$

which corresponds to the complement of the normalised error $\hat{e}_{\alpha_i}$ defined in equation 5.7.

The intuition behind Ψ_{α_i} is that, in a good fit, the error e_{α_i} should represent a small fraction of the historical gap that needs to be closed in a simulation ($|I_{i,-1} - I_{i,0}|$). Therefore, this metric penalises extreme errors by setting negative values that, in turn, will affect the aggregate goodness of fit.[1]

The metric for the goodness of fit of parameter β_i follows the same logic, and it is

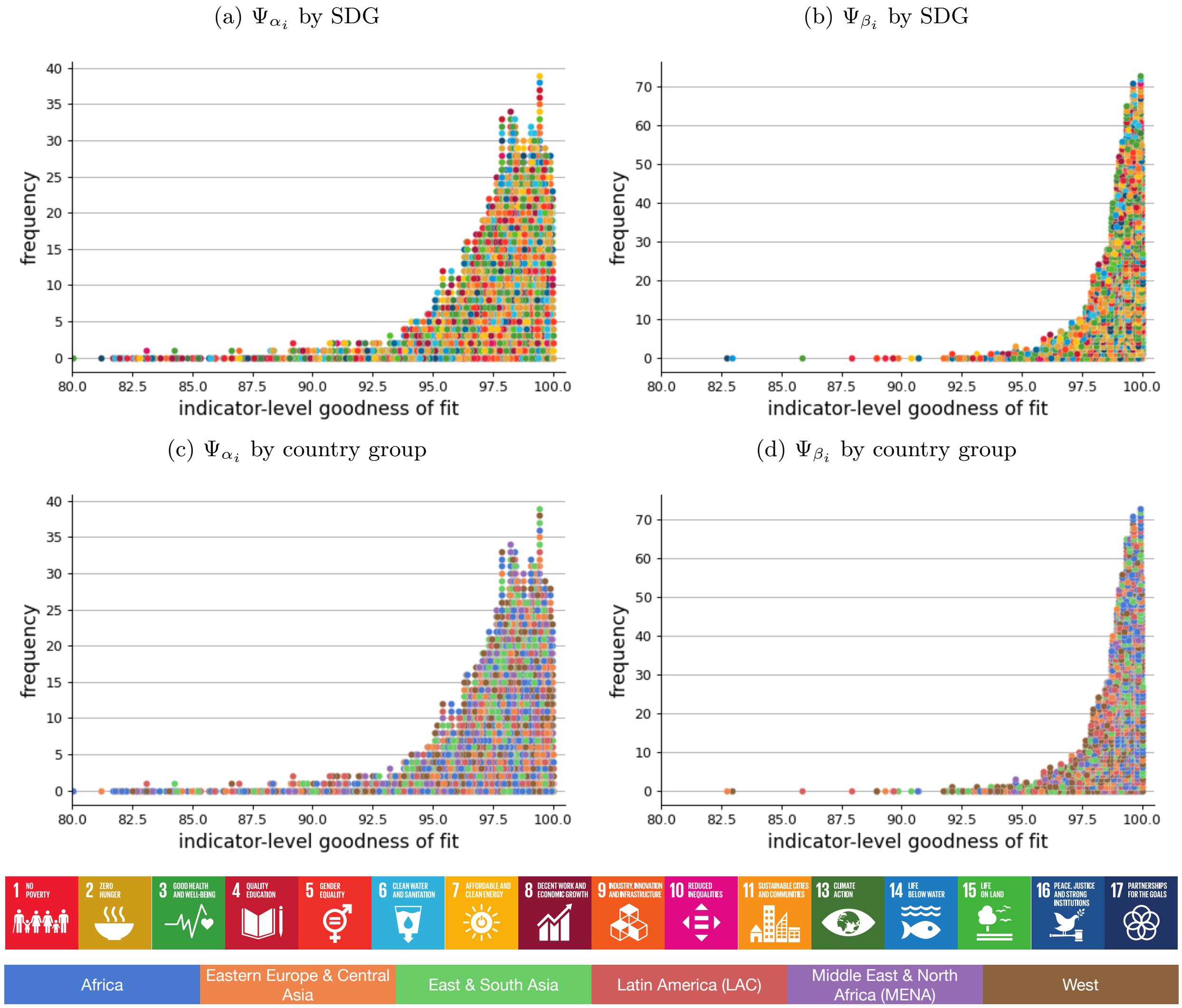
$$\Psi_{\beta_i} = 1 - \frac{e_{\beta_i}}{r_i}, \tag{5.12}$$

where r_i is the empirical success rate as defined in 5.4.

We calculate the Ψ_{α_i} and Ψ_{β_i} of each indicator for every country in the SDR dataset. Then, bin them into different levels and plot their frequencies in Figure 5.2. Figures 5.2a and 5.2b present the Ψ_{α_i} and Ψ_{β_i} coloured by SDG, while Figures 5.2c and 5.2d show the same information but coloured by country group. In this example, we have set the threshold criterion at 80%, and we can see that the algorithm is able to obtain goodness of metrics greater than that for all the indicators in the dataset.

[1] When testing alternative calibration methods, we find that they yield several indicators displaying a negative Ψ_{α_i}. This is not the case for our algorithm.

Figure 5.2: Distribution of goodness-of-fit metric by SDG and country group



(a) Ψ_{α_i} by SDG

(b) Ψ_{β_i} by SDG

(c) Ψ_{α_i} by country group

(d) Ψ_{β_i} by country group

Notes: The goodness of fit is in percentage.
Source: Authors' own calculations with data from the 2021 Sustainable Development Report.

5.4 On confidence and testing

Now that we have discussed the model and its calibration, it is important to talk about the quantification of uncertainty to perform tasks such as establishing statistical confidence. First, it is important to unpack and discuss a couple of basic concepts that, many times, are conflated in the teachings of statistics in the social sciences. Let us begin with the confidence intervals, to then discuss hypothesis testing.

5.4.1 Confidence intervals

Constructing confidence intervals for a metric or statistic of interest is, essentially, the task of quantifying the uncertainty of the estimation of such metric. For this, the modeller needs to assume the source of such uncertainty and device a way to 'propagate' it throughout the empirical exercise, all the way to the statistic of interest. Usually, the result of propagating this uncertainty is a distribution of the metric or statistic. In computational modelling, constructing confidence intervals is not always straightforward because the complex interdependencies of a system makes it difficult to track the propagation from the source of uncertainty to the statistic. Hence, in models such as the one developed in this book, the user has explicitly to consider (1) the source of uncertainty, (2) the variation of the source, and (3) a simulation strategy to propagate such variation to the metric of interest.

From our experience in the intersection of development economics and sustainability, we have noticed that the most common source of uncertainty brought up by colleagues and peers is the quality of the indicators (i.e., measurement uncertainty). This is natural since, in many developing countries, the procedures to collect such information, and the commitment of the authorities can be questionable at times. In this respect, our approach has been to propagate the uncertainty conveyed in the indicators through an ensemble-calibration approach. This computationally intensive task can be achieved by randomising the original indicator data according to their distribution intervals (sometimes provided by the source), and to generate one calibration for each randomised dataset (notice that this, effectively, generates a distribution of each free parameter). Then, with each calibration, one performs the counterfactual of interest (and the necessary Monte Carlo simulations) in order to arrive to the metrics of interest and to their distributions.

Readers experienced in using development-indicator data probably raised an eyebrow when we mentioned that the distribution intervals of the data since, in many cases, indicators come without such information. In this case, as with the normally-distributed assumption of a regression error, one needs to model the uncertainty in the data. For example, in Guerrero and Castañeda (2021b), we use the inter-temporal volatility of each indicator as a proxy for the variability in the quality of an indicator. The logic is that least-developed countries exhibit more volatility in their indicators, in part, because their infrastructure and methods for collecting such data can be more fragile than in developed nations. Another example can be found in our work with Daniele Guariso ***, where we model each indicator time series through a Gaussian process. Then, we use of these models to generate randomised synthetic indicators that we use create the calibration ensembles. Finally, in Guerrero and Castañeda (2020a), we analyse model uncertainty by implementing different model specifications for the government heuristic. While this exercise focuses on robustness and validation, these various alternative specifications can also be used to produce calibration ensembles.

5.4.2 Hypothesis testing

In the PPI research programme, we are not so interested in the statistical confidence of the parameter vectors $\boldsymbol{\alpha}$, $\boldsymbol{\alpha}'$, and $\boldsymbol{\beta}$. This stands in stark contrast with the regression frameworks to which quantitative social scientists are accustomed, where the estimated coefficients of a regression carry an explicit meaning and, hence, ought to be statistically tested against the null of being zero-valued. In PPI and other areas of the Computational Social Sciences, in contrast, testing for the significance of model parameters such as $\boldsymbol{\alpha}$, $\boldsymbol{\alpha}'$, and $\boldsymbol{\beta}$ is ill defined, as these parameters do no carry a meaning that is as precised as the coefficients of a linear regression. Thus, defining a null hypothesis seems rather ambiguous. Instead, the main interest lies in testing more meaningful statistics such as development gaps, time savings, efficiency gains, etc. In a computational framework such as PPI, building a null for these statistics becomes a bit more nuanced because, often, a zero-value statistic is not as meaningful as, for example, being able to determine where an observed empirical value would be expected in a world without the mechanism or intervention that is being tested. This way of thinking–creating a null model/simulation instead of deriving the distribution of the statistic–is common in fields such as network science, statistical mechanics, and computational biology. In contrast, social scientists who are trained in regression analysis often find this logic it difficult to grasp.

In general, the type of statistical significance testing exercises that will be of interest to the user of our model have to do with performing counterfactual simulations. In principle, that involves (1) generating a set of benchmark simulations and (2) simulating the counterfactual. For example, as we show in Chapter 5.5.2, if we want to verify whether positive spillovers indeed elicit incentives to be inefficient (as argued in the presentation of the model), we can produce a set of benchmark simulations with a spillover network, and another set without spillovers. Testing for statistical significance is usually done using the relevant distribution, and this could also include uncertainty if it is properly propagated through the null model. Note that there is no one-size-fits-all method on how to formulate statistical tests. Each problem requires carefully thinking about the meaning of the benchmark and the counterfactual, and how the information from the distributions should be used (e.g., a difference-in-means test, a paired test, a custom-built non-parametric test, etc.). Throughout the book, we will present some examples in which we have devised different strategies. What is important to remember is that, under this modelling framework, one should keep an open and creative mind to different ways to pose a concept of significance, rather than sticking to the *fitting-the-line* practice that is so prevalent in the social sciences.

5.5 Validation

The topic of validation is key in the generation of knowledge and for discriminating between competing models. While this is something commonly acknowledged, the meaning of validation varies from one field to another; sometimes it is mixed with related–but different–ideas such as generalisability and robustness. Furthermore, within specific fields or methodologies, validation can be tested in multiple ways. Hence, it is important to, first, identify the concepts of validation that are relevant in an agent-computing context; something that we have previously done in Guerrero and Castañeda (2020a), and which we discuss here.

In computational modelling (not just agent-computing), validation has many flavours that have evolved as more data and new methods have become available. Perhaps one of the pioneering works in categorising several of these flavours is Carley (1996), who identifies on eight levels of validation. By today's standards, Carley's validation levels can be classified in a hierarchical manner, with external and internal validation at the top, and various other nuanced concepts inside them. Here, we discuss and present some of these validation strategies applied in the context of our model.

5.5.1 External validation

External validation in agent-computing models typically means replicating one or more quantitative stylised facts (e.g., distributions, moments, or correlations) by generating them from the bottom up. Importantly, matching a stylised fact for validation purposes should not be an objective of the calibration exercise, otherwise the validation is trivial. Whenever possible, the target stylised fact should be constructed from a dataset that is independent (testing set) from the one used to calibrate the model (training set). In terms of Carley's validation levels, external validation would encompass parameter, point, distributional, and value validation levels.

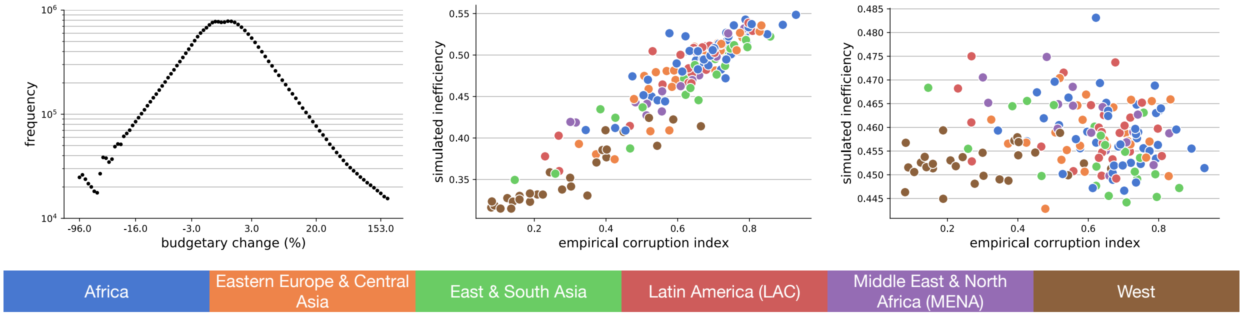
In our early work (Castañeda et al., 2018; Guerrero and Castañeda, 2020a, 2021a), we externally validate variants of this model by replicating two well-known stylised facts: (1) the heavy-tailed distribution of budgetary changes and (2) the negative relationship between development and corruption. Here, we would like to briefly revisit those validation strategies and show that, with this new model and data, they still hold.

First, we consider the distribution budgetary changes. A large literature in political science has documented non-normal tails in the distribution of changes in government budgets (total changes as well as disaggregated into policy issues) (Jones et al., 1998; John and Margetts, 2003; Jones and Baumgartner, 2005; Jones et al., 2009). This evidence is not entirely convincing on which one is the exact distribution generating the data. Nevertheless, an indisputable feature is that changes in government expenditure do not follow distributions with exponentially decaying tails (like a normal), but rather heavy-tailed ones. So, the question at hand is whether, without the influence of empirical budgetary data, our model is able to generate simulated budgetary changes that exhibit heavy tails.

To demonstrate that this is indeed the case, we perform 10000 Monte Carlo simulations with fully randomised data. That is, in each simulation, we generate (1) a random number of indicators (between 50 and 200 and randomly assigning which one is instrumental), (2) a random spillover network (with weights between -1 and 1), (3) random governance parameters, and (4) random free parameters (between 0 and 1). Figure 5.3a shows the resulting distribution. The plot is presented in log-log scale, which suggests that the tails are heavy when they show a linear decaying pattern. The graphic was constructed by taking all the changes of each individual government allocation $P_{i,t}$. Since this exercise does not use any empirical disbursement schedule, we can claim external validity because this stylised fact emerges from bottom-up thanks to the socioeconomic mechanisms within the model.

Figure 5.3: External validation

(a) Distribution of budgetary changes (b) Differentiated governance (c) Homogeneous governance



Notes: Panel (a) shows the simulated distribution of budgetary changes at the level of expenditure programs. Panel (b) compares Transparency International's corruption index against the model's endogenous level of corruption (they have a linear correlation larger than 93%). Panel (c) shows the association between Transparency International's corruption index and the model's corruption level under a counterfactuals where the governance parameters φ_i and τ_i equal 0.5 for every country.
Source: Authors' own calculations.

Next, let us turn to another external validation test. The SDR dataset reports International Transparency's corruption index for most of the countries in the sample. We intentionally removed this indicator from the dataset used in this book because it is redundant with the model's endogenous variable of inefficiency $P_{i,t} - C_{i,t}$, which we can interpret as corruption (at least in part). Thus, while the corruption index has been left out of the study, we can use it to validate the model by evaluating how well its endogenous inefficiency variable matches the index. To compute the level of inefficiency produced by the model across M simulations, we calculate $(\sum_{i,t,M} P_{i,t} - C_{i,t})/(M \times B)$, which is the fraction of the budget that is lost in inefficiencies. In parallel, we invert the directionality of the corruption index, so that higher values denote less corruption. In this way, if the model's emergent inefficiency has a positive correlation with the corruption index, then this validates the public governance mechanisms of the model. Figure 5.3b shows a strong association between the model's inefficiency and the corruption index. Their linear correlation is greater than 93%. In addition, Figure 5.3c shows a similar plot where, instead of using the empirical data on the governance parameters for 'quality of monitoring' and for the 'rule of law', we fix them in 0.5 for every country. The result clearly shows that the correlation dissipate because the agents' responses in terms of their contributions is not distinguishable across countries with the same quality of public procurement.

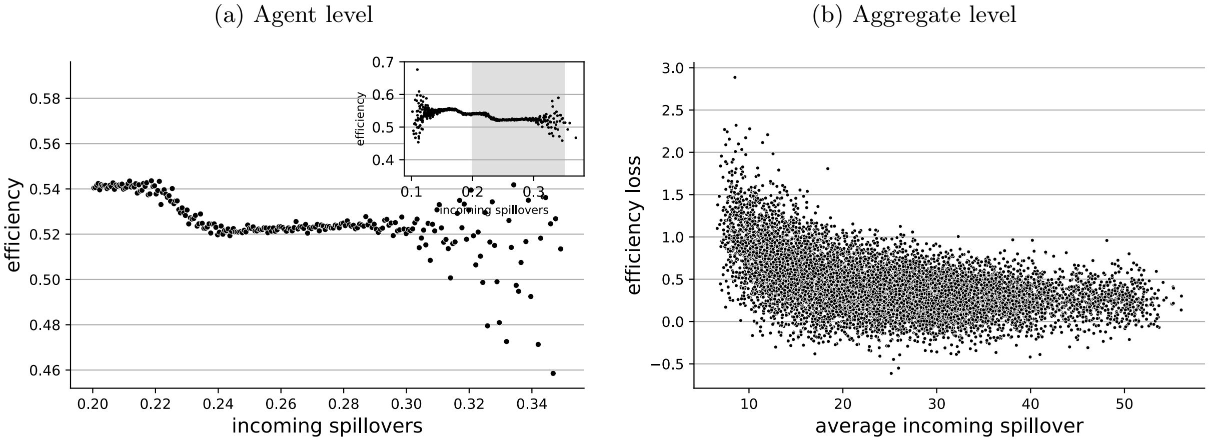
5.5.2 Internal validation

Internal validation consists of showing that certain theoretical expected outcomes (whether externally validated or not) are sensitive to the social and behavioural mechanisms specified in the model. On the one hand, internal validation can show that the micro-mechanisms are theoretically consistent. On the other, it suggests that these mechanisms bring new and useful information to the model because the expected outcomes are sensitive to the specification. In connection with Carley's work, internal validation corresponds to the theoretical validation level.

In Castañeda et al. (2018); Guerrero and Castañeda (2020a), we internally validate earlier versions of this model by analysing the relationship between positive spillovers and inefficiencies. Recall that, according to the theory behind the model, if a public servant's policy issue improves due to positive spillovers from other topics, this may elicit perverse incentives from the agent because they would be able to 'disguise' their inefficiencies using these positive externalities. Thus, theoretically, one should expect that, the more positive spillovers received by an agent, the less efficient (lower contributions) they will be. Notice that this connection–between spillovers and efficiency–is not easy to trace mathematically due to the complexity arising from the interactions and the behavioural components. Therefore, there is no guarantee of such logic holding in a simulation. This is where internal validation is important; it allows modellers to confirm that the mechanisms accounted for have, indeed, the intended influence in the system or in the agents' behaviour (at least partially). To demonstrate that this is the case in our model, we provide two alternative ways to internally validate it.

First, we validate the spillover→efficiency mechanisms at the micro level. That is, we show that agents who receive more positive spillovers tend (for the most part) to be less efficient. For agent i, the average amount of spillovers received across M simulations is $\sum_{t,m}^{T,M} S_{i,t,m}/(T \times M)/N$. We divide the spillovers over N because we compare simulations with different numbers of indicators, which affect how much spillovers occur in the system. We compute this quantity for each agent across 10000 sets of $M = 100$ Monte Carlo simulations. Each set of Monte Carlo simulations is produced under the same parameters, which are randomly generated for each set, in the same fashion as we do for the external validation (with the difference that we only generate positive spillovers). At the same time, we compute the agent-level average efficiency $\sum_{t,m}^{T,M}(C_{i,t,m}/P_{i,t,m})/(T \times M)$. Figure 5.4a shows that, indeed, our model is capable of eliciting inefficiencies through spillover effects. The plot suggests a negative non-linear relationship between the spillovers received and the level of efficiency for most levels of spillovers (they exhibit linear and non-linear correlations above -75%). The inset figure shows the full relationship, with large volatility at extreme spillover levels. Thus, we can claim that, for the most part, our model is internally valid with respect to the agent's incentives and their response to network effects.

Figure 5.4: Internal validation


(a) Agent level

(b) Aggregate level

Notes: Panel shows ***.
Source: Authors' own calculations.

Second, we internally-validate the spillover→efficiency relationship through aggregate evidence. For this, we use the same simulation as before, but calculate the aggregate level of incoming spillovers and efficiency, which are $\sum_{i,t,m}^{n,T,M} S_{i,t,m}/(n \times T \times M)$ and $\sum_{i,t,m}^{n,T,M} C_{i,t,m}/(B \times M)$ respectively. A distinctive for of this exercise is that, for each set of $M = 100$ Monte Carlo simulations, we assemble another set with the exact same parameters but without the network, so no spillovers occur in these counterfactual simulations. Once we compute the aggregate efficiency in both sets, we compute the different of the efficiency without a network and efficiency with one. If the difference is positive, then it means that the economy is more efficient without positive spillovers because agents are less incentives to engage in inefficient activities; we call this difference the loss in efficiency. Figure 5.4b shows the result of this exercise. The results are striking since, not only all the efficiency losses are positive, but also there is a clear negative association between the size of the loss and the average level of spillovers received by the agents. With these results, we provide evidence that validate our model internally at different levels of aggregation.

5.5.3 Soft validation

Soft validation is probably the most common one in the agent-computing literature, as it involves a qualitative assessment of an empirical pattern. It corresponds to what Carley's terms as pattern validation level. It is different from external validation because the validity assessment does not use a formal metric, but rather a qualitative judgement.

When studying policy coherence (Guerrero and Castañeda, 2020b), we provide a 'soft' validation exercise for a variant of our model. This consists of estimating an index of policy coherence for countries that are known to have been coherent with emulating specific economies in the past, for example, Korea following Japan, or Estonia adopting the Nordic development model. If the coherence index is consistent with this qualitative narrative of successful emulations, it provides further evidence that the inferred policy priorities contain valid information. Such exercise requires a balanced cross-national panel of development indicators, and a verifiable narrative as to why such qualitative pattern should be expected. In Guerrero and Castañeda (2020b), such narrative is provided by Akamatsu's flying geese, by scholarly work on the countries under study, and by the public discourse of government officials.

5.5.4 Stakeholder validation

In the literature of participatory modelling (Guyot and Honiden, 2006), researchers seek to involve the stakeholders of a problem in the modelling process, and this may be through role-playing games, experiments, consultations, workshops, and feedback activities, to mention a few possibilities. The idea is that stakeholders can help to specify the data and mechanisms that 'actually' take place, and to verify that the model 'makes sense'. For Carley, this corresponds to the face and process validation levels.

Our work through various projects with policymakers (Castañeda and Guerrero, 2019a,b,c; Guerrero and Castañeda, 2020a; Sulmont et al., 2021) has allowed us to build certain level of stakeholder validation. For instance, during a collaboration with the UNDP-Mexico, different stakeholders from the federal- and state-level governments and NGOs took part in two workshops where the methodology, data, and results were presented and discussed. The stakeholders took part in an exercise in which they had to classify the indicator database into instrumental or collateral and, then, the results were discussed to reach a consensus. They were also involved in developing the idea of fluid versus rigid allocations since fiscal rigidities are something that 'actually' occurs quite often in the public administration. Hence, the stakeholders provided early feedback in refining the data and the model.

5.6 Conclusions

***missing

Bibliography

Amos, R. and Lydgate, E. (2020). Trade, Transboundary Impacts and the Implementation of SDG 12. *Sustainability Science*, 15(6):1699–1710.

Asadikia, A., Rajabifard, A., and Kalantari, M. (2021). Systematic prioritisation of SDGs: Machine learning approach. *World Development*, 140:105269.

Baez-Camargo, C. and Passas, N. (2017). Hidden agendas, social norms and why we need to re-think anti-corruption. Working Paper, Basel Institute on Governance.

Benedek, D., Gemayel, E., Senhadji, A., and Tieman, A. (2021). A Post-Pandemic Assessment of the Sustainable Development Goals. *Staff Discussion Notes*, 2021(003).

Boeren, E. (2019). Understanding Sustainable Development Goal (SDG) 4 on "Quality Education" from Micro, Meso and Macro Perspectives. *International Review of Education*, 65(2):277–294.

Carley, K. (1996). *Validating Computational Models.* Working Paper. CASOS Program, Pittsburgh, PA.

Castañeda, G., Chávez-Juárez, F., and Guerrero, O. (2018). How Do Governments Determine Policy Priorities? Studying Development Strategies through Networked Spillovers. *Journal of Economic Behavior & Organization*, 154:335–361.

Castañeda, G. and Guerrero, O. (2019a). Inferencia de Prioridades de Política para el Desarrollo Sostenible. Reporte Metodológico, Programa de las Naciones Unidas para el Desarrollo.

Castañeda, G. and Guerrero, O. (2019b). 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. and Guerrero, O. (2019c). Inferencia de Prioridades de Política para el Desarrollo Sostenible: Una Aplicación para el Caso de México. Reporte Técnico, Programa de las Naciones Unidas para el Desarrollo.

Fader, M., Cranmer, C., Lawford, R., and Engel-Cox, J. (2018). Toward an Understanding of Synergies and Trade-Offs Between Water, Energy, and Food SDG Targets. *Frontiers in Environmental Science*, 0.

Fuso Nerini, F., Sovacool, B., Hughes, N., Cozzi, L., Cosgrave, E., Howells, M., Tavoni, M., Tomei, J., Zerriffi, H., and Milligan, B. (2019). Connecting Climate Action with Other Sustainable Development Goals. *Nature Sustainability*, 2(8):674–680.

González-Pier, E., Barraza-Lloréns, M., Beyeler, N., Jamison, D., Knaul, F., Lozano, R., Yamey, G., and Sepúlveda, J. (2016). Mexico's path towards the Sustainable Development Goal for health: An assessment of the feasibility of reducing premature mortality by 40% by 2030. *lancet.Global health*, 4(10):e714–e725.

Guerrero, O. and 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. and Castañeda, G. (2020b). Quantifying the Coherence of Development Policy Priorities. *Development Policy Review*, 00:1–26.

Guerrero, O. and Castañeda, G. (2021a). Does expenditure in public governance guarantee less corruption? Large non-linearities and complementarities of the rule of law. *Economics of Governance*, forthcoming.

Guerrero, O. and Castañeda, G. (2021b). How does government expenditure impact sustainable development? studying the multidimensional link between budgets and development gaps. *SSRN Working Paper*.

Guyot, P. and Honiden, S. (2006). Agent-Based Participatory Simulations: Merging Multi-Agent Systems and Role-Playing Games. *Journal of Artificial Societies and Social Simulation*, 9(4).

Ionescu, G., Firoiu, D., Tănasie, A., Sorin, T., Pîrvu, R., and Manta, A. (2020). Assessing the Achievement of the SDG Targets for Health and Well-Being at EU Level by 2030. *Sustainability*, 12(14):5829.

John, P. and Margetts, H. (2003). Policy punctuations in the UK: Fluctuations and equilibria in central government expenditure since 1951. *Public Administration*, 81(3):411–432.

Jones, B., Baumgartner, F., Breunig, C., Wlezien, C., Soroka, S., Foucault, M., François, A., Green-Pedersen, C., Koski, C., John, P., Mortensen, P., Varone, F., and Walgrave, S. (2009). A General Empirical Law of Public Budgets: A Comparative Analysis. *American Journal of Political Science*, 53(4):855–873.

Jones, B. D. and Baumgartner, F. R. (2005). A model of choice for public policy. *Journal of Public Administration Research and Theory*, 15(3):325–351.

Jones, B. D., Baumgartner, F. R., and True, J. (1998). Policy Punctuations: U.S. Budget Authority, 1947-1995. *The Journal of Politics*, 60(1):1–33.

Kroll, C., Warchold, A., and Pradhan, P. (2019). Sustainable Development Goals (SDGs): Are We Successful in Turning Trade-Offs into Synergies? *Palgrave Communications*, 5(1):1–11.

Luken, R., Mörec, U., and Meinert, T. (2020). Data Quality and Feasibility Issues with Industry-Related Sustainable Development Goal Targets for Sub-Saharan African Countries. *Sustainable development*, 28(1):91–100.

Lusseau, D. and Mancini, F. (2019). Income-Based Variation in Sustainable Development Goal Interaction Networks. *Nature Sustainability*, 2(3):242–247.

Machingura, F. and Lally, S. (2017). The Sustainable Development Goals and Their Trade-Offs. Technical report, Overseas Development Institute, London, United Kingdom.

McGowan, P., Stewart, G., Long, G., and Grainger, M. (2019). An Imperfect Vision of Indivisibility in the Sustainable Development Goals. *Nature Sustainability*, 2(1):43–45.

Mensi, A. and Udenigwe, C. (2021). Emerging and Practical Food Innovations for Achieving the Sustainable Development Goals (SDG) Target 2.2. *Trends in Food Science & Technology*, 111:783–789.

Moyer, J. and Hedden, S. (2020). Are We on the Right Path to Achieve the Sustainable Development Goals? *World Development*, 127:104749.

Pedercini, M., Arquitt, S., Collste, D., and Herren, H. (2019). Harvesting synergy from sustainable development goal interactions. *Proceedings of the National Academy of Sciences*, 116(46):23021–23028.

Philippidis, G., Shutes, L., M'Barek, R., Ronzon, T., Tabeau, A., and van Meijl, H. (2020). Snakes and Ladders: World Development Pathways' Synergies and Trade-Offs through the Lens of the Sustainable Development Goals. *Journal of Cleaner Production*, 267:122147.

Porciello, J., Ivanina, M., Islam, M., Einarson, S., and Hirsh, H. (2020). Accelerating Evidence-Informed Decision-Making for the Sustainable Development Goals Using Machine Learning. *Nature Machine Intelligence*, 2(10):559–565.

Pradhan, P., Subedi, D., Khatiwada, D., Joshi, K., Kafle, S., Chhetri, R., Dhakal, S., Gautam, A., Khatiwada, P., Mainaly, J., Onta, S., Pandey, V., Parajuly, K., Pokharel, S., Satyal, P., Singh, D., Talchabhadel, R., Tha, R., Thapa, B., Adhikari, K., Adhikari, S., Bastakoti, R., Bhandari, P., Bharati, S., Bhusal, Y., Bk, B., Bogati, R., Kafle, S., Khadka, M., Khatiwada, N., Lal, A., Neupane, D., Neupane, K., Ojha, R., Regmi, N., Rupakheti, M., Sapkota, A., Sapkota, R., Sharma, M., Shrestha, G., Shrestha, I., Shrestha, K., Tandukar, S., Upadhyaya, S., Kropp, J., and Bhuju, D. (2021). The COVID-19 Pandemic Not

Only Poses Challenges, but Also Opens Opportunities for Sustainable Transformation. *Earth's Future*, 9(7):e2021EF001996.

Sobczak, E., Bartniczak, B., and Raszkowski, A. (2021). Implementation of the No Poverty Sustainable Development Goal (SDG) in Visegrad Group (V4). *Sustainability*, 13(3):1030.

Sulmont, A., García de Alba Rivas, M., and 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.