Policy Priority Inference: A Computational Framework to Analyze the Allocation of Resources for the Sustainable Development Goals

Supplementary Materials

Omar A. Guerrero & Gonzalo Castañeda

A. Data pre-processing

For clarity, let us introduce some notation. Let V denote a matrix where the rows are the units of observation (the development indicators) and the columns correspond to periods (years). Entry $V_{i,t}$ denotes the i^{th} element in period t. Then, if we perform an operation on one of the indices while holding a specific value for the other, we replace the operated index by the dot symbol $\cdot$. For example, if we want the maximum value across time for the i^{th} value of V, then we write $\max(V_{i,\cdot})$. Similarly, if we want the lowest indicator in a given period t, then we write $\min(V_{\cdot,t})$. Finally, if we omit the second index it means we have a vector.

A.1. Normalization

We normalize the indicators in the range [0,1]. The purpose of normalizing is to make the indicators comparable. Data expressed in percentages, may not need this normalization. However, if a large sample across countries can be collected, then normalization is recommended as it would provide more realistic bounds of what is achievable (e.g. a 100% coverage of forest to total land ratio is impossible). This is so because the levels of other countries act as benchmarks for how low and how high indicators are.
In order to normalize the indicators we employ the formula

$$\mathcal{I}_{i,t} = \frac{\mathscr{I}_{i,t} - \min(\mathscr{I}_{i,\cdot})}{\max(\mathscr{I}_{i,\cdot}) - \min(\mathscr{I}_{i,\cdot})}, \tag{17}$$

where $\mathscr{I}$ denotes the raw indicator and $\mathcal{I}$ the normalized one. The min and max operators are applied to the entire time series of indicator i across all available countries in the sample. In the case of Mexico, the data have been normalized across a larger sample covering 298 countries and territories for 27 years.

A.2. Imputation

PPI requires the initial and final values of each indicator. However, should the user want to estimate the spillover network through quantitative methods, then it is desirable to also have intermediate observations. While the data collected for this study has comprehensive time-series coverage in the sampling period, there are still some missing observations. To remedy this problem, we generate linear interpolations.

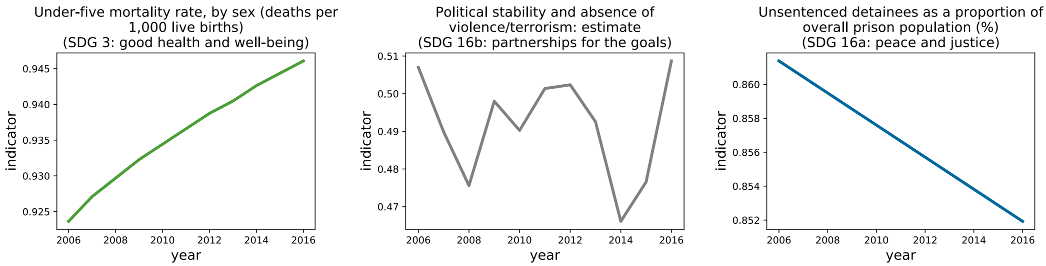
A.3. Reversing

For an easier interpretation, the higher values of an indicator should denote better outcomes. Since we normalized the indicators in the range [0,1], we can reverse them by using the complement $1 - \mathcal{I}_{i,t}$.

A.4. Goal adjustment

The model underlying PPI guarantees convergence to a vector of goals T. It does not matter if those goals are above or below the initial values of the indicators. Nevertheless, there is a problem with goals that are lower than their initial values. Assuming that the indicators have been pre-processed as suggested above and that higher values imply better outcomes, convergence to lower values would be inconsistent with PPI's logic. This is so because negative dynamics would imply that the government's systematic investment drives the indicators to worse outcomes. Moreover, the spillover network should already account for the negative externalities. Hence, the goal of an indicator should always be greater than its initial value.

Figure 13: Development indicators exhibit diverse dynamics



The indicators have been normalized in the interval [0,1] and have been reversed if necessary. Thus, higher values indicate better outcomes.

Figure 13 shows how different development indicators have various dynamics. PPI simplifies the challenge of modeling such diverse patterns through its stochastic growth process. However, this does not fix the problem of having indicators with final values lower than the initial ones. To remedy this issue for retrospective estimations, we propose an adjustment of the retrospective goals using the formula

$$T_i = I_{i,m} + |\min(I_{\cdot,m} - I_{\cdot,1})| + \epsilon, \tag{18}$$

where m is the final period in the sample and $\epsilon > 0$ is a small term close to zero. This calculation needs to be performed for all indicators, even if only one of them exhibits $I_{i,m} < I_{i,1}$. Given the normalization of the indicators, we choose an $\epsilon = 0.01$ for this application, since smaller values could produce extremely low αs, increasing the computational burden of the model.

Equation 18 shifts all the final values upwards, guaranteeing $T_i > I_{i,1}$ for every i. Effectively, it assigns the smallest historical gap (the difference between final and initial values) to the worst-performing indicator, and the largest one to the best-performing one. Thus, the historical gaps capture how much progress was achieved in each indicator during the sampling period.

Besides the argument of model consistency, a second motivation for this adjustment is linked to the growth factors α. Assume there are no network effects, and that the inefficiencies and allocations are the same across all nodes. Then, the only parameter that could explain the variation between historical gaps is α_i. Since all the empirical indicators arrived at their final values at the same time, it must be the case that the worst-performing indicator has the smallest growth factor and the best one has the largest α_i. What we have done here is mapping the performance of the indicators into the growth factors. This is why α_i can be interpreted as everything else that explains the indicator dynamics but that is not explicitly considered in the model. Furthermore, we interpret this additional information captured in α as long-term structural factors.

B. Model variables

Table 5 presents all the variables of the model. We have arranged them according to their sources. Clearly, most variables are endogenous. The free parameters that need to be calibrated are: the growth factors α_i (one for each indicator) and the number simulation steps before convergence. Regarding the exogenous variables, all of them can be obtained from publicly available datasets for most countries. Given a sample of development indicators for a given period, their initial values $I_{i,0}$ determine the initial conditions of the country, region or sector under study. The goals T, on the other hand, represent the aspirations that a government or society has, so they take specific values to be reached by each indicator. From a retrospective point of view: $T_i = I_{i,last}$, where $I_{i,last}$ denotes the final value of indicator i. In other words, we assume that the final values of the data sample represent the real aspirations that the government had in the past. This provides a benchmark to calibrate the model. Finally, the adjacency matrix $\mathbb{A}$ may be estimated via different methods. Whichever is the chosen approach to infer the spillover network, it is assumed to be exogenous.

Table 5: Variables of the model

Symbol	Variable	Source
α	growth factor	calibration
$\mathcal{T}$	retrospective convergence time	calibration
I	development indicators	data
$\mathbb{A}$	spillover network adjacency matrix	data
T	goals	data
φ	quality of monitoring	data
τ	quality of the rule of law	data
X	actions	endogenous
F	functionaries' benefits	endogenous
C	contributions	endogenous
D	relative private gains	endogenous
γ	probability of successful growth	endogenous
λ	probability of spotting inefficiencies	endogenous
θ	binary outcome of monitoring	endogenous
ξ	binary outcome of random growth process	endogenous
S	net incoming spillovers	endogenous
P	allocation profile	endogenous
G	goal-indicator gaps	endogenous
H	history of inefficiencies	endogenous

Sub-indices have been omitted.

C. Data

Here, we provide the complete list of development indicators.

Table 6: Development indicators

Description	Source	ODS	Instrumental	Reversed
Poverty gap at 5.50 dollars a day (2011 ppp) (%)	World Bank	1	yes	yes
Population in moderate poverty	CONEVAL	1	yes	yes
Population in extreme poverty	CONEVAL	1	yes	yes
Population that is vulnerable due to poor social capital	CONEVAL	1	yes	yes
Population that is vulnerable due to poor income	CONEVAL	1	yes	yes
Lack of health services	CONEVAL	1	yes	yes
Lack of social security	CONEVAL	1	yes	yes
Lack of quality and space in the dwelling	CONEVAL	1	yes	yes

Table 6: Development indicators

Description	Source	ODS	Instrumental	Reversed
Lack of basic house services	CONEVAL	1	yes	yes
Plant breeds for which sufficient genetic resources are stored (number)	UN	2	yes	no
Proportion of local breeds classified as being at unknown level of risk of extinction (%)	UN	2	yes	yes
Cereal yield (kg per hectare)	World Bank	2	no	no
Food production index (net, per capita)	FAO	2	yes	no
Prevalence of anemia among women of reproductive age (% of women ages 15-49)	World Bank	2	yes	yes
Under-five mortality rate, by sex (deaths per 1,000 live births)	UN	3	yes	yes
Number of new hiv infections per 1,000 uninfected population, by sex and age (per 1,000 uninfected population)	UN	3	no	yes
Tuberculosis incidence (per 100,000 population)	UN	3	yes	yes
Malaria incidence per 1,000 population at risk (per 1,000 population)	UN	3	yes	yes
Mortality rate attributed to cardiovascular disease, cancer, diabetes or chronic respiratory disease (probability)	UN	3	no	yes
Suicide mortality rate, by sex (deaths per 100,000 population)	UN	3	no	yes
Alcohol consumption per capita (aged 15 years and older) within a calendar year (litres of pure alcohol)	UN	3	no	yes
Proportion of the target population with access to 3 doses of diphtheria-tetanus-pertussis (dtp3) (%)	UN	3	yes	no
Proportion of the target population with access to measles-containing-vaccine second-dose (mcv2) (%)	UN	3	yes	no
Participation rate in organized learning (one year before the official primary entry age), by sex (%)	UN	4	yes	no
Internet access in schools, 1-7 (best)	World Economic Forum	4	yes	no
Quality of the education system, 1-7 (best)	World Economic Forum	4	yes	no
Quality of primary education, 1-7 (best)	World Economic Forum	4	yes	no
Quality of math and science education, 1-7 (best)	World Economic Forum	4	yes	no
Quality of management schools, 1-7 (best)	World Economic Forum	4	yes	no
Extent of staff training, 1-7 (best)	World Economic Forum	4	no	no
School enrollment, secondary (gross), gender parity index (gpi)	World Bank	4	yes	no
Proportion of seats held by women in national parliaments (% of total number of seats)	UN	5	yes	no
Proportion of women in managerial positions (%)	UN	5	no	no
Water body extent (permanent and maybe permanent) (% of total land area)	UN	6	yes	no
Proportion of population with access to electricity, by urban/rural (%)	UN	7	yes	no
Proportion of population with primary reliance on clean fuels and technology (%)	UN	7	yes	no
Access to clean fuels and technologies for cooking (% of population)	World Bank	7	yes	no
Annual growth rate of real GDP per capita (%)	UN	8	no	no
Number of commercial bank branches per 100,000 adults	UN	8	no	no
Unemployment rate, by sex and age (%)	UN	8	no	yes
Foreign direct investment, net inflows (% of GDP)	World Bank	8	yes	no
Index of economic complexity	Observatory of Economic Complexity	8	no	no
Efficiency of government spending	World Economic Forum	8	yes	no
Burden of government regulation, 1-7 (best)	World Economic Forum	8	yes	no
Burden of customs procedures, 1-7 (best)	World Economic Forum	8	yes	no
Regulation of securities exchanges, 1-7 (best)	World Economic Forum	8	yes	no
Business impact of rules on fdi, 1-7 (best)	World Economic Forum	8	yes	no
Strength of auditing and reporting standards, 1-7 (best)	World Economic Forum	8	yes	no
Protection of minority shareholders' interests, 1-7 (best)	World Economic Forum	8	no	no
Intensity of local competition, 1-7 (best)	World Economic Forum	8	yes	no
Effectiveness of anti-monopoly policy, 1-7 (best)	World Economic Forum	8	yes	no
Extent of market dominance, 1-7 (best)	World Economic Forum	8	yes	no
Efficacy of corporate boards, 1-7 (best)	World Economic Forum	8	no	no
Cooperation in labor-employer relations, 1-7 (best)	World Economic Forum	8	yes	no
Flexibility of wage determination, 1-7 (best)	World Economic Forum	8	yes	no
Pay and productivity, 1-7 (best)	World Economic Forum	8	no	no
Tax revenue (% of GDP)	World Bank	8	yes	no
New business density (new registrations per 1,000 people ages 15-64)	WDI	8	yes	no
Imports as a percentage of GDP	World Economic Forum	8	no	no
Strength of investor protection, 0-10 (best)	World Economic Forum	8	yes	no
Patent applications, residents	World Bank	8	no	no
Contribution of labor quality to GDP growth	The Conference Board	8	no	no
Exports of goods and services (% of GDP)	World Bank	8	no	no
Gdp, ppp (constant 2011 international dollars)	World Bank	8	no	no
Wage and salaried workers, total (% of total employment) (modeled ILO estimate)	World Bank	8	no	no
No. days to start a business	World Economic Forum	8	yes	yes
No. procedures to start a business	World Economic Forum	8	yes	yes
Rate of informal employment	INEGI	8	no	yes
Growth of total factor productivity	The Conference Board	8	no	no
Number of fixed internet broadband subscriptions, by speed (number)	UN	9	no	no
Internet users per 100 inhabitants	UN	9	no	no
Manufacturing value added per capita (constant 2010 united states dollars)	UN	9	no	no
Available airline seat km/week, millions	World Economic Forum	9	no	no
Quality of overall infrastructure, 1-7 (best)	World Economic Forum	9	yes	no
Quality of roads, 1-7 (best)	World Economic Forum	9	yes	no
Quality of air transport infrastructure, 1-7 (best)	World Economic Forum	9	yes	no
Quality of electricity supply, 1-7 (best)	World Economic Forum	9	yes	no
Availability of latest technologies, 1-7 (best)	World Economic Forum	9	yes	no
Firm-level technology absorption, 1-7 (best)	World Economic Forum	9	no	no
Fdi and technology transfer, 1-7 (best)	World Economic Forum	9	yes	no
Quality of scientific research institutions, 1-7 (best)	World Economic Forum	9	yes	no
Government procurement of advanced tech products, 1-7 (best)	World Economic Forum	9	yes	no
Soundness of banks, 1-7 (best)	World Economic Forum	9	yes	no
Venture capital availability, 1-7 (best)	World Economic Forum	9	no	no
Financing through local equity market, 1-7 (best)	World Economic Forum	9	yes	no
Availability of research and training services, 1-7 (best)	World Economic Forum	9	yes	no
Company spending on r&d, 1-7 (best)	World Economic Forum	9	no	no
Capacity for innovation, 1-7 (best)	World Economic Forum	9	yes	no
Availability of scientists and engineers, 1-7 (best)	World Economic Forum	9	yes	no
Quality of port infrastructure, 1-7 (best)	World Economic Forum	9	yes	no
Fixed telephone lines/100 pop.	World Economic Forum	9	yes	no
Investment in energy with private participation (current us dollars)	World Bank	9	yes	no
Investment in transport with private participation (current us dollars)	World Bank	9	yes	no
Mobile telephone subscriptions/100 pop.	World Economic Forum	9	no	no
Labour share of GDP, comprising wages and social protection transfers (%)	UN	10	no	no
Ease of access to loans, 1-7 (best)	World Economic Forum	10	yes	no
Income share held by lowest 10%	World Bank	10	no	no
Gini index (world bank estimate)	World Bank	10	no	yes
Pm2.5 air pollution, population exposed to levels exceeding who guideline value (% of total)	World Bank	11	yes	yes

Table 6: Development indicators

Description	Source	ODS	Instrumental	Reversed
Material footprint per capita, by type of raw material (tonnes)	UN	12	no	yes
Domestic material consumption per capita, by type of raw material (tonnes)	UN	12	no	no
Degree of customer orientation, 1-7 (best)	World Economic Forum	12	no	no
Ethical behavior of firms, 1-7 (best)	World Economic Forum	12	no	no
Adjusted net savings, excluding particulate emission damage (% of gni)	World Bank	12	yes	no
Coal rents (% of GDP)	World Bank	12	no	yes
Forest rents (% of GDP)	World Bank	12	no	yes
Mineral rents (% of GDP)	World Bank	12	no	yes
Natural gas rents (% of GDP)	World Bank	12	no	yes
Oil rents (% of GDP)	World Bank	12	no	yes
Total natural resources rents (% of GDP)	World Bank	12	no	yes
Intensity of emissions, meat and cattle	FAO	13	yes	yes
Temperature variation	FAO	13	no	yes
Average proportion of marine key biodiversity areas (kbas) covered by protected areas (%)	UN	14	yes	no
Average proportion of terrestrial key biodiversity areas (kbas) covered by protected areas (%)	UN	15	yes	no
Average proportion of mountain key biodiversity areas (kbas) covered by protected areas (%)	UN	15	yes	no
Red list index	UN	15	yes	no
Unsentenced detainees as a proportion of overall prison population (%)	UN	16a	yes	yes
Business costs of terrorism, 1-7 (best)	World Economic Forum	16a	no	no
Business costs of crime and violence, 1-7 (best)	World Economic Forum	16a	yes	no
Organized crime, 1-7 (best)	World Economic Forum	16a	yes	no
Reliability of police services, 1-7 (best)	World Economic Forum	16a	yes	no
Intentional homicides (per 100,000 people)	World Bank	16a	yes	yes
Public trust in politicians, 1-7 (best)	World Economic Forum	16b	no	no
Favoritism in decisions of government officials, 1-7 (best)	World Economic Forum	16b	yes	no
Transparency of government policymaking, 1-7 (best)	World Economic Forum	16b	yes	no
Property rights, 1-7 (best)	World Economic Forum	16b	yes	no
Intellectual property protection, 1-7 (best)	World Economic Forum	16b	yes	no
Judicial independence, 1-7 (best)	World Economic Forum	16b	yes	no
Government effectiveness: estimate	World Bank	16b	yes	no
Overall level of statistical capacity (scale 0 - 100)	World Bank	16b	yes	no
Legal rights index, 0-10 (best)	World Economic Forum	16b	yes	no
Political stability and absence of violence/terrorism: estimate	World Bank	16b	yes	no
Regulatory quality: estimate	World Bank	16b	yes	no
Corruption perception index	Transparency International	16b	no	no
Voice and accountability: estimate	World Bank	16b	yes	no
Debt service as a proportion of exports of goods and services (%)	UN	17	yes	yes
Prevalence of foreign ownership, 1-7 (best)	World Economic Forum	17	no	no
Prevalence of trade barriers, 1-7 (best)	World Economic Forum	17	yes	no
Gross national savings, % GDP	World Economic Forum	17	no	no
Inflation, annual % change	World Economic Forum	17	no	yes
Travel and tourism direct contribution to GDP percentage share of total GDP	World Travel & Tourism Council	17	yes	no

Table 7: Indicators with missing observations

Indicator	SDG	Missing years
Poverty gap at 5.50 dollars a day	1	2007, 2009, 2011, 2013, 2015
Population in moderate poverty	1	2007, 2009, 2011, 2013, 2015
Population in extreme poverty	1	2007, 2009, 2011, 2013, 2015
Population that is vulnerable due to poor social capital	1	2007, 2009, 2011, 2013, 2015
Population that is vulnerable due to poor income	1	2007, 2009, 2011, 2013, 2015
Lack of health services	1	2007, 2009, 2011, 2013, 2015
Lack of social security	1	2007, 2009, 2011, 2013, 2015
Lack of quality and space in the dwelling	1	2007, 2009, 2011, 2013, 2015
Lack of basic house services	1	2007, 2009, 2011, 2013, 2015
Plant breeds for which sufficient genetic resources are stored	2	2006, 2007, 2008, 2009, 2011, 2013, 2015
Malaria incidence per 1,000 population at risk	3	2006, 2007, 2008, 2009, 2011, 2012, 2013, 2014
Mortality rate attributed to cardiovascular disease, cancer, diabetes or chronic respiratory disease	3	2006, 2007, 2008, 2009, 2011, 2012, 2013, 2014
Suicide mortality rate, by sex	3	2006, 2007, 2008, 2009, 2011, 2012, 2013, 2014
Alcohol consumption per capita	3	2006, 2007, 2008, 2009, 2011, 2012, 2013, 2014
Proportion of women in managerial positions	5	2006, 2007, 2008, 2009, 2010, 2011, 2012
Proportion of population with primary reliance on clean fuels and technology	7	2006, 2007, 2008, 2009, 2011, 2012, 2013, 2014
Annual growth rate of real GDP per capita	8	2008, 2009, 2013
Unemployment rate, by sex and age	8	2006
Tax revenue	8	2006, 2007
Investment in energy with private participation	9	2009
Income share held by lowest 10%	10	2007, 2009, 2011, 2013, 2015
Gini index	10	2007, 2009, 2011, 2013, 2015
Pm2.5 air pollution, population exposed to levels exceeding who guideline value	11	2006, 2007, 2008, 2009
Unsentenced detainees as a proportion of overall prison population	16a	2006, 2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014, 2015

D. Model simulations and calibration

This section provides the details on how to calibrate the model's free parameters and how to perform Monte Carlo simulations to generate inferences.

D.1. Calibrations of growth factors

Finding the vector of growth factors can prove challenging because of the interdependencies between indicators. For instance, increasing α_i may affect the convergence time of other indicators through the network because i's 'steps' become larger and so do the spillovers. Thus, simultaneously calibrating all αs is not a trivial problem, and we have found that many non-linear optimization methods fail. For this reason, we have devised a heuristic strategy to solve this problem.

Our calibration method computes the marginal effect of each growth factor independently until an error term is minimized. In order to think about the error, let us assume that we want all indicators to converge simultaneously after $\mathcal{T}$ simulation steps. Then, the objective is to minimize the average deviation from $\mathcal{T}$ across all indicators and simulations.

First, let us determine an arbitrary vector of growth factors. Using this vector, we perform one simulation run until all indicators converge; then, we obtain a vector with the number of periods that it took each indicator to converge. By repeating this step m times, we obtain m convergence time vectors, which allows computing the average convergence time V_i for each indicator. Then, we compute the convergence error

$$r_c = \frac{1}{N} \sum_{i}^{N} |\mathcal{T} - V_i|. \tag{19}$$

Next, we want to identify those indicators whose individual convergence error $|\mathcal{T} - V_i|$ is greater than a tolerance threshold e_v. For one of these indicators, say i, we vary α_i marginally. Then, we perform m simulations and compute $|\mathcal{T} - V_i|$. We repeat these two steps for i, covering the range $(0, 1)$ for α_i. Then, we choose the level of α_i that minimizes the difference between convergence error and e_v. This greedy search is performed for each indicator with a convergence error $|\mathcal{T} - V_i| > e_v$, until we obtain a new vector of convergence factors. With this new vector, we re-estimate r_c and repeat all the previous steps. The procedure stops when $r_c < e_v$.

While our heuristic assumes a *ceteribus paribus* condition in every greedy search, it is effective in finding the growth factors. Since V_i consists of average convergence times, the error is sensitive to the size of m. That is, larger ms decrease the variance of convergence times. Consequently, the mapping from α_i to V_i during the greedy search becomes more accurate with more simulations. This of course, comes with a computational burden. Therefore, the model calibration greatly benefits from parallel computing. Algorithm 2 shows the pseudocode of the calibration procedure.

Algorithm 2: Calibration pseudocode

Input: $\mathcal{T}, \alpha_1, \ldots, \alpha_N$, initial $I, T, \mathbb{A}, \varphi, \tau, m, e_v$

```
while r_c > e_v do
    foreach i such that |T − V_i| > e_v do
        set α_i = argmin|T − V_i(α_i)|;
    update convergence error r_c;
```

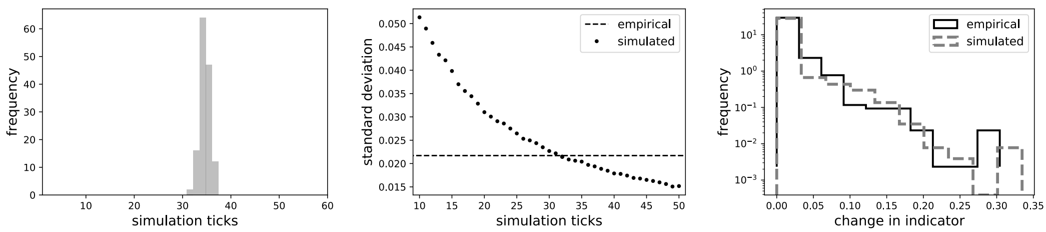
D.2. Calibration of simulation periods

Our calibration procedure assumes a number of periods $\mathcal{T}$ under which the model should converge. In order to calibrate $\mathcal{T}$, we aim at matching the total volatility of the indicators.

First, it is necessary to adjust the empirical data so that its volatility is comparable to the synthetic one. This is so because, while the empirical data may exhibit upward and downward dynamics, PPI's model only generates growth dynamics. Therefore, the adjustment consists of computing one-period changes in all the empirical indicators and, then, turning any negative change into zero. Now, we can calculate the standard deviation of the adjusted data. To calibrate $\mathcal{T}$, we need to find a number of periods under which the calibrated growth factors yield indicator dynamics with similar volatility to the empirical one. This means that the entire calibration procedure needs to be performed every time a different $\mathcal{T}$ is chosen.

Figure 14 shows the results of the calibration procedures. The left panel presents a histogram of average convergence times across indicators once the model growth factors have been calibrated. Clearly, the calibrated growth factors for $\mathcal{T} = 32$ generate a small divergence from the target number of periods. Here, the average error r_c is less than one. The middle panel shows the volatility of the simulated indicators obtained for different levels of $\mathcal{T}$ (each one with its calibrated growth factors). For this study, $\mathcal{T} = 32$ yields the best match between the empirical and the simulated volatility. In the right panel, we can see the histogram of the changes in the indicators (one for the empirical data and one for the simulated). Once all parameters have been calibrated, the model generates a similar distribution to the empirical one.

Figure 14: Model simulations and calibration



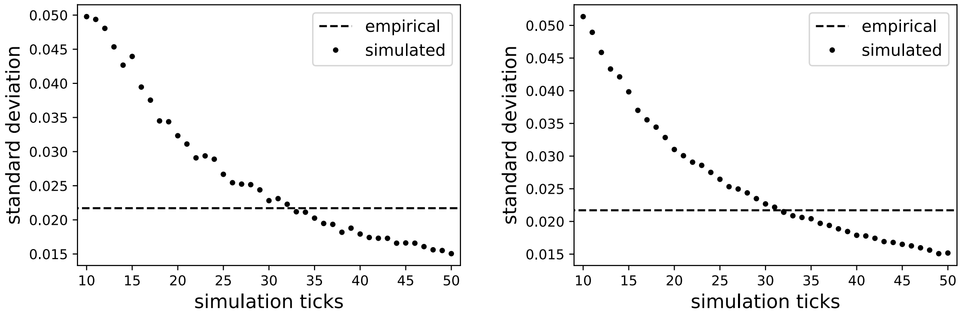
Model simulations for a calibration of $\mathcal{T} = 32$. Left: histogram of average convergence times across indicators. Middle: indicator volatility as a function of $\mathcal{T}$. Right: empirical and fitted distributions of the changes in the indicators.

D.3. Computational efficiency

Combined, the two calibration procedures can be computationally expensive. As we have mentioned earlier, one way to reduce the computational burden is to run these processes in parallel. However, through our understanding of the model, we can provide further advice on how to manage the computational burden of estimating the growth factors.

We have found that the volatility of the simulated indicators is not too sensitive to the precision with which the growth factors are calibrated, but rather to the assumed number of convergence periods $\mathcal{T}$. In other words, as an initial step to find an optimal $\mathcal{T}$, one may relax the convergence error threshold to produce a mapping like the one in the middle panel of Figure 14. By relaxing the error threshold, we do not need to run numerous simulations during the greedy search of algorithm 2, significantly reducing the computational burden. Once the mapping between $\mathcal{T}$ and the indicators' volatility has been produced, we can find the optimal $\mathcal{T}^*$. Then, we can re-calibrate the growth factors with more simulations and a more conservative error. As a verification step, one may repeat this last re-calibration with $\mathcal{T}^* - 1$ and $\mathcal{T}^* + 1$ to check that the best volatility is still given by $\mathcal{T}^*$. We have found this to be a very effective strategy to increase the computational efficiency of our calibration method.

Figure 15: Reduction in computational burden



Left: calibration procedure running 10 simulations in the greedy search. Right: calibration procedure running 100 simulations in the greedy search.

Figure 15 shows the result of the calibration procedure under different computational burdens. While there is some sensitivity to the number of simulations runs considered in the greedy search of the first step of the calibration, the overall mapping of $\mathcal{T}$ into volatility is robust, yielding $\mathcal{T}^* = 32$ even when relatively few simulations are produced.

D.4. Monte Carlo simulation

Once the model has been calibrated, PPI can be used to produce several inferences via Monte Carlo simulation. It is important to run independent realizations of the model because the uncertain environment under which the agents learn may lead to decision paths that are specific to a particular simulation. The idea behind the Monte Carlo approach is to generate many realizations of the world and to compute the expected values of the variables of interest. Besides the model parameters, it is also necessary to initialize the endogenous variables randomly (P, H, and X, for example.). Thus, in order to account for model uncertainty, one could construct confidence intervals from these distributions. The more simulations performed, the narrower those intervals become. In other words, inferences become more robust with a large number of simulations.

To demonstrate this point, let us concentrate on the main endogenous variable: the allocation profile P. In order to assess the robustness, we measure the similarity between two expected allocation profiles: one obtained from M simulations and another calculated from $M + 1$. That is, for two samples of sizes M and $M + 1$, we compute multiple pairs of expected allocation profiles and calculate their Spearman correlation. If sample size improves the estimation, then the variation of the Spearman correlation should decrease as we increase M. Figure 16 confirms this. After performing 1000 simulations, the estimated policy priorities are robust.

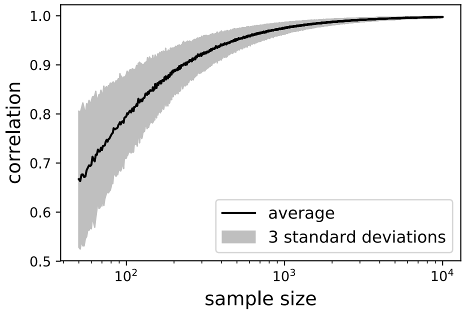
D.5. Robustness under alternative government specifications

Recall that the government's adaptive heuristic uses two sources of information to allocate resources to indicator i: (1) the historical gap G_i and (2) the historical inefficiencies H_i. Then, the propensity to allocate resources to policy issue i is given by

$$q_{i,t} = {G_{i,t}}^{1+H_{i,t}}.$$

In the main text, we have justified why the historical gaps should be more important than the historical inefficiencies. Nevertheless, given that this function is key to determine the shape of P, it is natural to

Figure 16: Robustness of allocation profiles across different sample sizes



question whether alternative functional forms would dramatically change the resulting allocation profile. Here, we show that this is not the case.
Of course, the alternative specifications should preserve our assumption of historical gaps being more important than historical inefficiencies. Otherwise, one would be dealing with a different underlying theory of how the government behaves, and significantly different allocations are to be expected. Then, given the previous requirement, let us define three alternative specifications for equation 10 from the main text: a quotient form ($q_{i,t} = \frac{G_{i,t}}{1+H_{i,t}}$), a multiplicative form ($q_{i,t} = (G_{i,t})(1 + H_{i,t})$) and an exponential form ($q_{i,t} = {G_{i,t}}^{2-H_{i,t}}$).
We run PPI for each of these specifications and compare all the resulting allocation profiles. Table 8 reports the Spearman correlation between each pair of allocation profiles. The upper numbers correspond to the correlation between allocation profiles. The results suggest robustness. The lower numbers in parenthesis report a second test in which, for a given simulation, we divide each allocation of the prospective priorities by each one of the retrospective ones. In other words, we obtain a vector of ratios between prospective and retrospective inferences and then evaluate if they are robust across different specifications. Clearly, the results show that this is precisely the case.

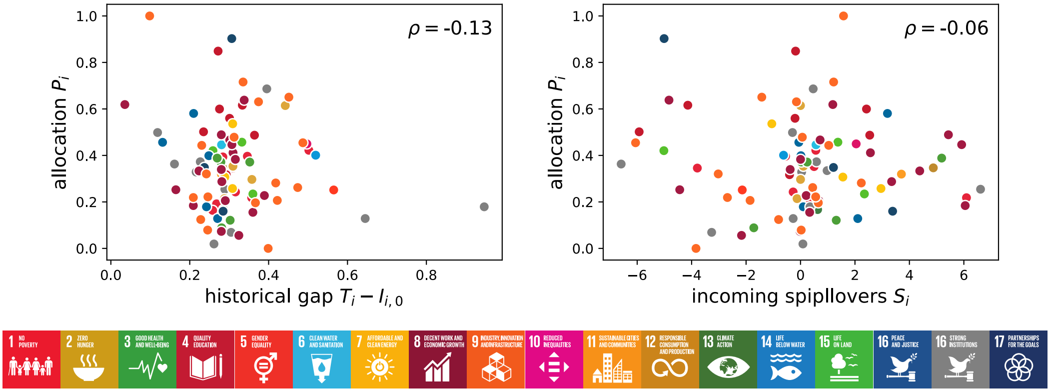
Table 8: Robustness to alternative specifications of the government's adaptive heuristic

Specification	$q_{i,t} = {G_{i,t}}^{1+H_{i,t}}$	$q_{i,t} = \frac{G_{i,t}}{1+H_{i,t}}$	$q_{i,t} = (G_{i,t})(1 + H_{i,t})$	$q_{i,t} = {G_{i,t}}^{2-H_{i,t}}$
$q_{i,t} = {G_{i,t}}^{1+H_{i,t}}$	1.00 (1.00)	0.974 (0.982)	0.976 (0.982)	0.971 (0.981)
$q_{i,t} = \frac{G_{i,t}}{1+H_{i,t}}$	- -	1.00 (1.00)	0.974 (0.988)	0.971 (0.987)
$q_{i,t} = (G_{i,t})(1 + H_{i,t})$	- -	- -	1.00 (1.00)	0.981 (0.987)
$q_{i,t} = {G_{i,t}}^{2-H_{i,t}}$	- -	- -	- -	1.00 (1.00)

E. Non-trivial inferences

A natural question regarding the inference of policy priorities is whether they could be obtained from simple back-of-the-envelope calculations. Figure 17 shows that this is not the case. Each panel compares the estimated policy priorities against simple data manipulations: the historical gaps and the sum of the weights of incoming edges. PPI's underlying model conveys non-trivial information because the inferred vector P does not correlate with either of these simple calculations. Therefore, inferences are not trivial.

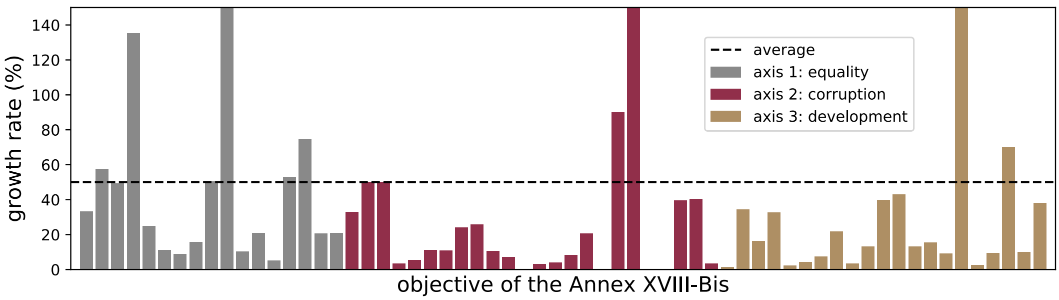
Figure 17: Non-trivial policy priorities



F. Constructing development goals from official documents

The Annex XVIII-Bis of Mexico's National Development Plan (NDP) identifies 234 specific policy issues that the government wants to improve in six years (from 2019 to 2024); let us call these issues objectives. Ideally, the specification of a development goal for an objective and the evaluation of the progress towards its achievement should be based on concrete measurements such as indicators. Unfortunately, the quantification of such a broad policy space is still challenging (even for the most advanced nations) and remains work in progress. Thus, from all the 234 objectives, the NDP identifies a subset of 67 that can be associated with specific indicators. By establishing a baseline year and a goal for each of these indicators, the NDP provides a picture of Mexico's development goals or aspirations. Figure 18 shows the distribution of the different development goals in terms of growth rates. That is, each bar represents the growth that the government intends to achieve in six years for each indicator of the NDP. The bars have been colored in three tones that match the three broad 'development axis' established in the NDP. In some indicators, the growth rates are extremely high, so the figure has capped their actual values. For example, the tallest bar in the *equality* axis corresponds to an indicator of spatial closeness to cultural infrastructure. In the baseline year, this indicator measures an average distance of 50km to cultural infrastructure. The government aims to reduce such distance to 5km. On average, the proposed growth rate for the NDP's indicators is approximately 50%. Not all the indicators presented in the NDP are suitable for usage in PPI. First, several indicators have been recently constructed, so their time-series do not have enough observations to match those indicators built for the retrospective analysis. Second, some of these data are highly specific to Mexico and cannot be normalized in rates or percentages. Third, some PND indicators are rather the result of political agreements to demonstrate that work is being done towards a goal, but not really to measure performance in such an indicator. An example of this can be found in indicator 1.6.1: *Number of projects in benefit of national development achieved through political agreements*.

Figure 18: Development goals from the National Development Plan



G. The role of public governance

Here, we provide a nuanced analysis of the role of public governance in the model. PPI accounts for public governance through two mechanisms: the monitoring of inefficiencies (φ) and the quality of the rule of law (τ). Here, we show the relevance of these mechanisms through their influence in the agent's efficiency choices. Recall that the contribution of public servant i is denoted by C_i. This is the effective amount of resources that this agent puts towards the public policy, out of the P_i resources allocated by the central authority. Then, efficiency is measured by the rate C_i/P_i.

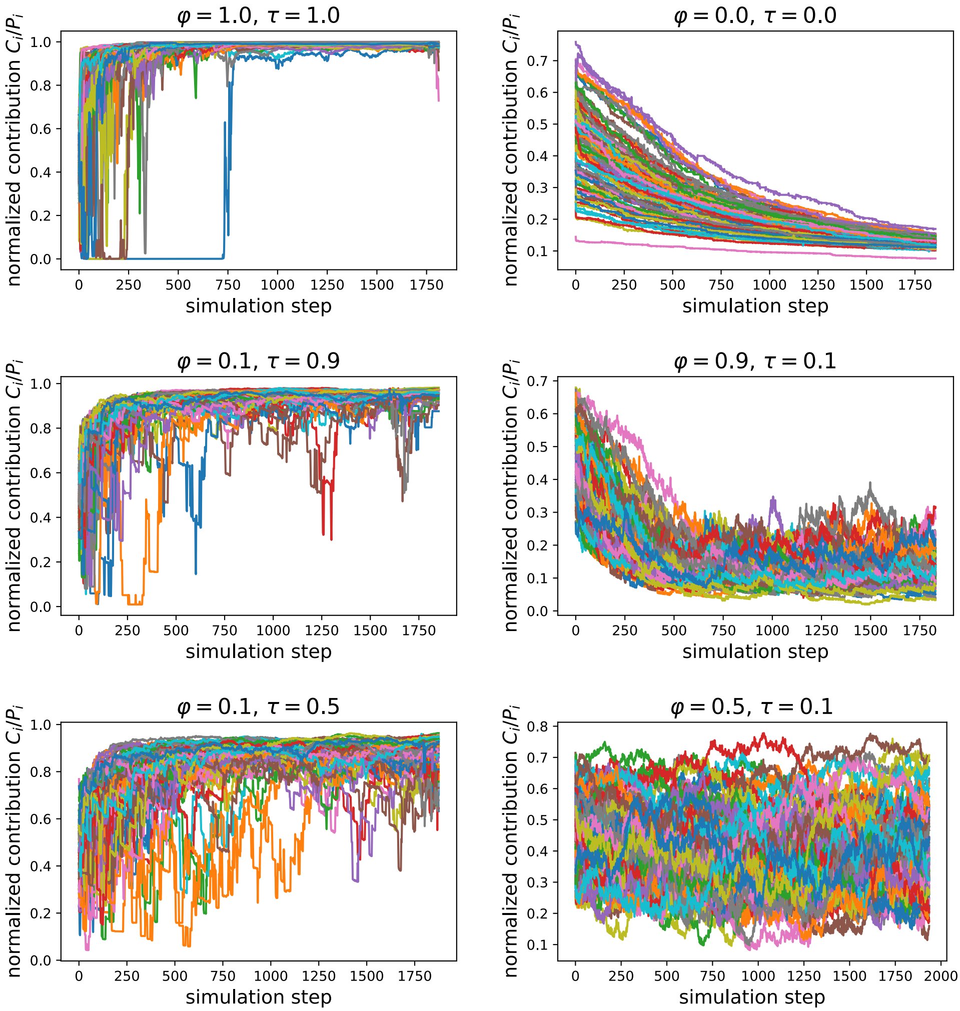
The first exercise consists of analyzing the evolution of agent-level efficiencies under different configurations of φ and τ. This analysis serves a double purpose: (1) to show the effect of different institutional settings and (2) to demonstrate that the learning model of the public officials works. For this, we present six parameterization cases of φ and τ in Figure 19.

Let us begin by looking at the two extreme cases in the upper panels: one with perfect monitoring and full effectiveness of the rule of law (left panel), and another where these mechanisms are absent (right panel). Clearly, when public governance is strong, agents learn that it does not pay to be inefficient, as their efficiency tends to 1. In contrast, in the absence of public governance, agents learn to become inefficient, decreasing their contributions towards zero. Note that, in both cases, agents exhibit different learning curves and occasional explorations. This is the kind of adaptive behavior that real agents exhibit in the presence of uncertainty and different forms of complexity in the environment. Thus, our learning model is well suited to study technical inefficiencies in the process of development.

Next, let us concentrate on the middle panels of Figure 19. Here, we present a case with poor monitoring and a strong rule of law (left panel); and another with good monitoring and a poor rule of law (right panel). When monitoring is weak, agents can spend several periods without being spotted. This allows them to explore being more inefficient and learn that it pays off. However, when they are caught, strict enforcement of the law inflicts high penalties, so agents react strongly by increasing their efficiency to almost 1. This is shown in the left panel, where agents tend to stay with high contributions, but infrequent supervision allows them to decrease them and reach low levels until they are caught (because of the stand out of a social norm); reverting to high contributions. On the right panel, we have a different type of dynamics, one of baring the costs of being inefficient. Here, strong monitoring efforts become an annoyance to the agents because the penalties are so low that they are willing to bear with them, i.e. they are the price to pay for being inefficient. In this scenario, contributions tend to be low and they exhibit upward spikes when penalized.

Finally, the bottom panels show the scenarios combining poor and mediocre qualities for monitoring and rule of law. They provide a mixture of dynamics where agents can explore the full spectrum of relative contributions. The important point to take from this exercise is that, depending on the institutional setting described by φ and τ, PPI can model a society with a law-abiding culture, one where inefficiencies are tolerated and penalties are the price to pay, or one undergoing a transition between the previous two.

Figure 19: Inefficiencies and public governance

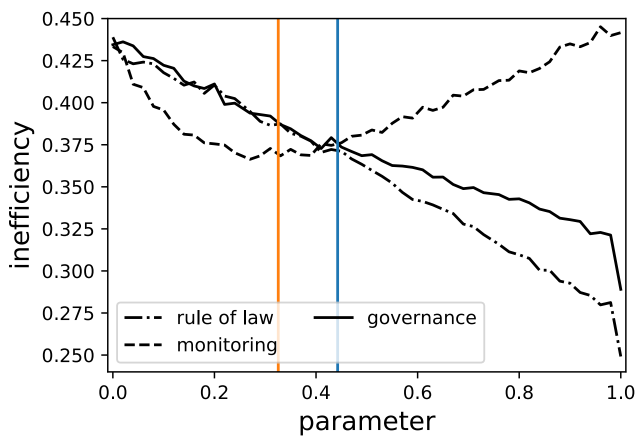


The second exercise illustrates the aggregate consequences of these micro-level dynamics. In particular, we show how different configurations of φ and τ can generate counter-intuitive aggregate relations between the level of inefficiency and the quality of the rule of law. When analyzing these results in detail, these apparent contradictions are explained by the emergence of social norms in the model.[26] Figure 20 shows the relationship between the aggregate level of inefficiency and the quality of governance. The two vertical lines denote the empirical values of Mexico's indicators on the strength of the monitoring efforts (φ) and the quality of the rule of law (τ). Let us first focus on the solid line. This relationship was produced by setting $\varphi = \tau$ and running PPI to obtain the total level of inefficiency

[26] As explained in the main text, social norms are one of the factors that explain why many developing countries have failed in curbing corruption through isolated (non-systemic) improvements to the rule of law.

for all values of the parameters in [0,1]. This exercise means that both aspects of public governance are marginally improved in the same amounts and that the relationship between the quality of governance and inefficiencies is negative; as predicted by most theories. Nothing is surprising from this result since clearly improving monitoring and the rule of law decreases the agent's incentives to be inefficient.
Next, let us focus on the dashed-dotted line, which also shows a negative relationship. Here, we fixed the quality of the monitoring efforts in Mexico's empirical level and varied τ from 0 to 1. More effective penalties decrease aggregate inefficiency. However, when the strength of the rule of law surpasses the quality of the monitoring efforts, its impact on inefficiencies is more effective than when improving both parameters in unison. This is an intriguing result because, according to the principal-agent perspective, agents should respond to both a shrinkage in the space of opportunities to be inefficient and to increments in the penalties. This result, on the other hand, suggests that it is more effective to leave some space of opportunity for inefficiency while focusing mainly on the rule of law. Our third experiment shows why this happens.
Now, note that the dashed line presents a U-shaped relationship between the quality of monitoring efforts and aggregate inefficiency. In this exercise, we fix the strength of the rule of law to its empirical level and manipulate φ. Why do we obtain an increase in total inefficiency when the quality of monitoring surpasses that of the rule of law? The answer lies in the micro-level dynamics presented in Figure 19; in particular, in the middle-right panel. A society with frequent monitoring but non-credible punishments learns that those penalties are the price to pay for being inefficient. The more frequent the monitoring is, the quicker the agents learn about the poor rule of law, so society gets locked into a higher inefficiency social norm; despite the government's best intentions. This is precisely the case of countries in which high corruption prevails, while the government frequently prosecutes corrupt officials. Eventually, the cases of such prosecutions are dropped or lost because the government was never serious about enforcing effective punishments (usually after the corruption scandal stops receiving media attention). Here, a complicit central authority fakes that it is fighting corruption, but society has already learned that high corruption is tolerable.

Figure 20: Inefficiencies and public governance



H. On validation

Given the policy relevance of PPI, a natural question that arises is how can it be validated. To answer this, it is important to first acknowledge that the meaning of validation varies across fields and methodologies. When it comes to computational models, a classic reference is a paper of Carley (1996), which speaks of at least eight validation levels. By today's standards, Carley's validation levels can be classified into external and internal validation. Therefore, we discuss these ideas, perform two tests, and elaborate on additional validation strategies (soft validation, stakeholder validation, and cross-validation).

H.1. External validation

External validation in agent-computing modeling typically means replicating one or more precise (formally measurable) quantitative stylized facts (distributions, moments, or correlations) by generating them from bottom-up. Importantly, the matched stylized facts should not come from statistical regularities used during the calibration procedure; this establishes a clear distinction between fitting and validation. Whenever possible, the validated stylized 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.

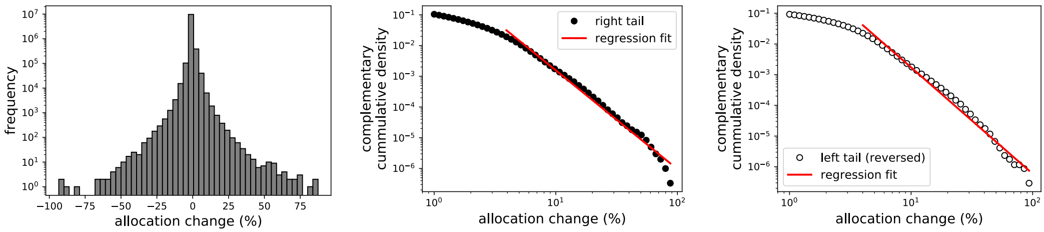
Early work on the CCG model (Castañeda et al., 2018) has been validated by replicating the cross-country distribution of corruption using an independent indicator on diversion of public funds that equates to the endogenous variable $P - C$ from the model. Such validation matched multiple features of this indicator: (1) a negative correlation between corruption and economic performance, (2) a large variation in corruption levels among middle-income countries, (3) no observations of low-income countries with low levels of corruption, and (4) no overlaps in corruption levels between low-income and high-income countries. This is relevant for the current paper because this model shares the same microeconomic specification, and is based on the same data on public governance. It is not a trivial result because while, indeed, the differences in governance parameters contribute to higher or lower levels of corruption, the spillover networks could have topologies that prevent this stylized fact from emerging, for example, because the number of synergies would encourage more corruption. Since this validation exercise is done at the international level, it requires a balanced cross-national dataset of indicators.

In this paper, we provide a new test for external validation that does not require cross-national data. In the political science literature, a power-law tail has been extensively documented in the distribution of budgetary changes (Jones et al., 2009). Then, it is straightforward to generate Monte Carlo simulations with random parameterizations in order to evaluate whether the emerging distribution of priority changes also follows a power-law in the tails. The procedure to perform Monte Carlo simulations considers the following:

- A random number of indicators (between 10 and 100);
- A spillover network with random weights drawn from a uniform distribution in (0,1) (weights are positive to be consistent with the theoretical considerations of the CCG model);
- Random governance parameters φ and τ drawn from a uniform distribution in (0,1);
- A random determination of which indicators are instrumental and which ones are collateral;
- A vector of random goals T drawn from a uniform distribution in (0,1);
- A vector of random initial values I_0 drawn from a uniform distribution in (0,1) such that $I_{i,0} < T_i$;
- A vector of random growth factors α drawn from a uniform distribution in (0,1).

We perform 10,000 simulations and pool the data on changes $P_{i,t} - P_{i,t-1}$ (the same results hold for single runs in general). Figure 21 shows our results, replicating the main plots of (Jones et al., 2009). Following the Jones et al.'s approach of testing for power-law tails through linear regression, we confirm that PPI generates this empirical regularity endogenously, providing strong evidence of external validity.

Figure 21: Internal validation test



0

H.2. Internal validation

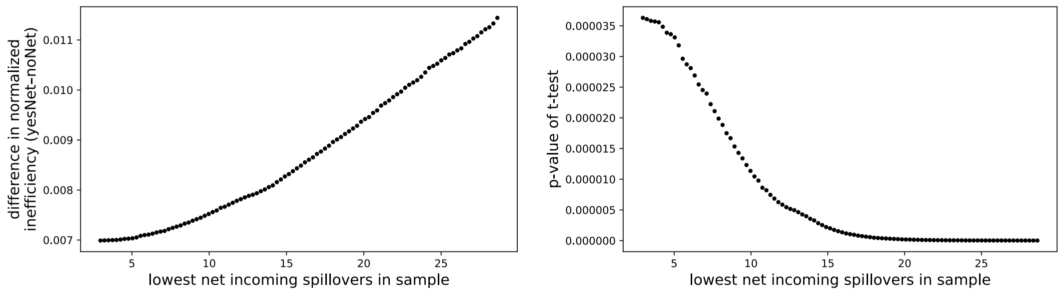
Internal validation consists of showing that certain theoretically-expected outcomes (whether externally validated or not) are sensitive to the social and behavioral mechanisms specified in the model. On the one hand, internal validation can show that the micro-mechanisms are theoretically consistent. On the other, it suggests that these mechanisms bring new 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 level.

In Castañeda et al. (2018), internal validation of the CCG model suggests that, in a system with positive spillovers, the network contributes to incentivize agents to divert more resources. Here, we replicate this exercise for PPI and show that the spillover network is important to determine inefficiencies. To provide the most-general possible internal validation, we do not use the Mexican data, but rather a theoretical specification that is generated at random. Thus, we consider a data specification identical to the one described in the previous sub-section.

The model with such specification is run 100 times, so the average normalized inefficiency $\omega_i = \frac{1}{L}\sum_t^L (P_{i,t} - C_{i,t})/P_{i,t}$ of each instrumental indicator is stored. Then, a second set of 100 simulations is run with the same specification, but removing the network. The average normalized inefficiencies $\omega_i' = \frac{1}{L'}\sum_t^{L'} (P'_{i,t} - C'_{i,t})/P'_{i,t}$ are saved for each instrumental indicator. Then, the difference in normalized inefficiencies $\omega_i - \omega_i'$ is computed for each instrumental node. If the difference is positive, is means that, indeed, the network contributes to incentivize the agents to be more inefficient. This procedure is repeated 1,000 times in a Monte Carlo fashion, in order to remove the potential influence from the other variables (because everything is randomized).

Figure 22 shows the result of our internal validation exercise. The left-most point corresponds to the average $\omega_i - \omega_i'$ from the 1,000 Monte Carlo simulation pairs across all indicators. The left panel shows the value of $\omega_i - \omega_i'$, while the right panel shows its corresponding p-value for a mean-equality t-test. Since $\omega_i - \omega_i'$ is positive and the p-value is lower than 1%, it suggests that the model produces significantly higher levels of inefficiency when spillovers are present than when they are absent. Furthermore, when conditioning the sample sets to nodes with at least a certain total amount of incoming spillovers, we can test if the difference in means increases. In both panels, the remaining points are the result of this exercise, showing the difference $\omega_i - \omega_i'$ and the corresponding p-values after filtering out indicators with a minimum amount of incoming spillovers. This confirms that, not only the network contributes to more aggregate inefficiency, but that it affects micro-level incentives and behaviors.

Figure 22: Internal validation test



H.3. Soft validation

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

In studying policy coherence with the CCG model, Guerrero and Castañeda (2020) provide a 'soft' validation exercise. This consists of estimating the 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 (2020), 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.

H.4. Stakeholder validation

In the literature of participatory modeling (Guyot and Honiden, 2006), researchers seek to involve the stakeholders of a problem in the modeling process, and this may be through role-playing games, experiments, consultations, workshops, and feedback activities, for example. 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.

During the project conducted in collaboration with the UNDP, 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.

Once presented with the results, the stakeholders acknowledged their intuitive nature. For example, the low priority given to SDG 16a, *peace and justice*, (see Figure 5) by the past administrations made complete sense to the participants, while the low feasibility of reaching the goals proposed for SDG 12, *responsible production and consumption*, was acknowledged by expert stakeholders. Therefore, PPI has also achieves some level of stakeholder validation.

H.5. Cross-validation

The machine-learning literature has popularized the concept of cross-validation. Roughly speaking, this consists of separating the data used for calibration from a sample used to test the predictive power of an algorithm. This separation is done multiple times at random, leading to robust measures of how good is the predictive performance of a model. It is considered validation because it tries to make sure that the estimates are not just the result of the sampling data, but that the model is able to find underlying relations that describe the generation of out-of-sample datasets. Clearly, cross-validation requires extensive data, reason why it is so popular in machine learning, and not in more traditional approaches such a growth regressions.

In principle, PPI could be subjected to these exercises. For this, a longer time-series of development indicators would be required. For instance, one could try to evaluate whether the convergence times under prospective estimations are close to the observed ones. This would require splitting the time series into a retrospective training set and a prospective testing one. At the moment, development indicator data does not allow such kind of testing. However, the fact that PPI would allow for it is a favourable feature that could be exploited in the future.

I. Further details on discovering accelerators

As explained in section 6.6, exploring the entire space of all possible combinations of indicators that would receive additional budget is computationally unfeasible. For this kind of problems, computer scientists have developed various heuristic optimization methods, some designed to work better than others under specific problems. One family of heuristic methods is known as evolutionary computing and, within this class, we can find genetic algorithms (Holland, 1975).

The first principle behind a genetic algorithm is that a generic solution to the optimization problem can be represented in a binary string, for example $\{1, 0, 0, 1, 0, 1, 0, \ldots, 0\}$. Each element in the string represents a control variable of the problem, while its binary value captures the state of the variable. In the case presented in section 6.6, the task is to select a subset of indicators from our sample, not to determine how much they will receive (that has been set ex-ante as an equal share of 10% of the budget). Hence, since the problem involves selecting or not selecting, each element in the string can represent an instrumental indicator, while its value indicates whether the indicator has been selected (1) or not (0). Thus, any solution to our problem can be represented by a binary string of length n (the number of instrumental indicators).

In this exercise, discovering accelerators means finding the binary string that minimizes convergence time. Thus, convergence time is what in this literature is known as the 'fitness' of the solution. The next element in this algorithm consists of initializing a population of random solutions and measuring their fitness. This means that, for every proposed solution, Monte Carlo simulations of PPI have to be run. Once every solution has a fitness, the population is ranked. Then, worst-performing half of the solutions are discarded (this is the selection step of evolution). Next, the remaining solutions are matched in pairs in order to create new solutions; their 'offspring'. A new solution emerges from the combination of two binary strings. That is to say, by determining a random i^{th} indicator for a pair of solutions, the algorithm takes the bits $1, 2, \ldots, i$ from the first solution, and concatenates them with the bits $i + 1, i + 2, \ldots, N$ of the second. This represents the passing of DNA material from the parent solutions to the offspring. The new solutions replace the discarded ones. Finally, as in evolution, there is mutation. This means that, one bit of each offspring solution will be randomly switched with a given probability (typically 5%). In this way, the genetic algorithm tries to maintain diversity in the population of solutions to prevent 'getting stuck' in a local optima.

This exact algorithm is performed for 200 generations (more than that does not improve fitness significantly). For robustness, we perform the algorithm multiple times since it is possible that the solution of one particular run is a local optima. Hence, the reported accelerators are those with value 1 in the binary string that presents the best fitness across all these computations.