5 Calibration and Validation

Now that we have discussed the theoretical foundations of PPI, we move to empirical matters regarding its usability with real-world data. For an agent-computing approach to help in providing policy advice, it is critical to analyse interventions in terms of an empirically grounded model. Due to the flexibility of a computational framework such as agent computing, where one can specify practically any mechanism, researchers often fall prey to the desire of 'wanting to account for everything possible' despite limited data availability. This trap creates identification-related problems because of having too many free parameters and few target metrics to calibrate them.[1] While the ABM community has developed many strategies to deal with problems related to model specification, selection, and overfitting, these strategies usually render agent-computing frameworks very complicated and difficult to scale or assimilate by non-experts.

One of our goals when developing PPI is to avoid falling into this trap; thus, a methodological prerequisite is to specify as many free parameters as the target metrics that we could identify. For scaling purposes, we aim for a model where calibration is possible in a high-dimensional parameter space that could be efficient, direct (without using surrogate-model strategies[2]), and simultaneous (so all parameters are assessed in each evaluation). PPI achieves all these

[1] Often, the target metric to calibrate a model is an aggregate quantity or stylised fact.

[2] A model is called a surrogate when it has been implemented to understand the behaviour of another model. For example, many ABMs often have a large parameter space and are extremely difficult to calibrate directly. In these cases, trying all possible parameter combinations to calibrate the model is computationally unfeasible. Hence, researchers often use the strategy of randomly sampling the parameter space to generate a large output dataset. Then, they deploy a statistical or machine-learning model to predict, using the parameters and output data, what would be the optimal parameter combination; effectively calibrating the ABM (see

https://doi.org/10.1017/9781009022910.008 Published online by Cambridge University Press

objectives since the micro-level parameters, such as adaptation steps and learning rates (usually the most difficult to calibrate), are endogenous. Leaving only free parameters at the macro level: $\boldsymbol{\alpha}$, $\boldsymbol{\alpha}'$, and $\boldsymbol{\beta}$. In this section, we explain how to calculate them and show – once calibrated – how to validate the model through various procedures. The calibration framework provided in this chapter is the same for all the models described throughout the book. This calibration strategy allows for PPI to be easily scaled and contributes to its appeal among stakeholders, given its practicality.

5.1 CALIBRATION STRATEGY

First, let us discuss how to specify the model's free parameters. We need to determine a total of $3N$ parameters (3 per indicator) from the data. 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 to be minimised when calibrating $\boldsymbol{\alpha}$ and $\boldsymbol{\alpha}'$ is the difference between the final value of the empirical indicators and the corresponding average final value of the simulated indicators. For $\boldsymbol{\beta}$, we seek to minimise the difference between the empirical success rate of the indicators (the number of times observing positive growth as a rate of the total number of periods-changes) and the average estimated probability of success $\gamma_{i,t}$. While both objective functions seem straightforward, there is more than meets the eye.

One of the challenges of calibrating the model relates to the agents' interdependencies and the presence of the spillover network (see Figure 4.3). In this setting, the indicator's dynamic is sensitive to the evolution of the other indicators. For instance, suppose we increase α_i and β_i while keeping all other parameters constant. If an indicator i conditions an indicator j through the spillover network ($\mathbb{A}_{i,j} \neq 0$), then j's dynamic will also change. For example, j would accelerate if i sends positive spillovers, meaning that we would need to readjust α_j and β_j in the opposite direction; likewise, this would

Carrella (2021) for a comprehensive review). Here, the statistical or machine-learning model is considered the surrogate.

https://doi.org/10.1017/9781009022910.008 Published online by Cambridge University Press

produce disarray in other indicators linked to j (or even in i). In earlier versions of PPI (Castañeda et al., 2018; Guerrero and Castañeda, 2020), we developed calibration algorithms assuming *ceteris paribus* conditions, where we only examined one parameter in each model evaluation. However, this procedure became computationally unfeasible when exploring more comprehensive models – parameter-wise – even if we were not experiencing issues of overfitting. Hence, as the PPI research evolved, we developed better algorithms (e.g., Guerrero and Castañeda, 2022) until we arrived at the one proposed in Guerrero et al. (2023), which we employ in this book.[3]

For computing the relevant statistics, our strategy consists of performing Monte Carlo simulations for the same set of parameter values. For example, suppose that we are trying to calibrate $\boldsymbol{\alpha}$. A single simulation may yield a final value for indicator i close to the empirical indicator, but another run may generate a very different estimate. This possibility is due to the stochastic components of the model and the potential presence of path dependence created by learning and social norms, something quite common in complex systems.[4] Once the simulations are run, we construct indicator-level statistics and errors that allow sensitivity to changes in indicator-specific parameters (overcoming the problem of having too many parameters and a single aggregate error).

Therefore, our optimisation algorithm uses a multi-objective function, which prevents the loss of indicator-specific error information, maintaining sensitivity to each parameter. Second, it readjusts

[3] A warning note for the reader is in order. If the spillover network is too dense and has very large weights, it is possible that PPI's calibration may become unfeasible due to the highly sensitive interdependency between the indicators. However, the density and weights required to break the calibration procedure would be far beyond what is typically observed in real-world data; hence, one could consider those networks degenerate. From our experience testing several methods to estimate networks from empirical time series of indicators, we have not found one that prevents us from calibrating PPI. Thus, when the calibration procedure fails, it is a warning signal to the users of PPI that they are considering a misspecified adjacency matrix $\mathbb{A}$.

[4] In a path-dependent process, the current outcome depends on prior decisions of agents and contingencies. See, for instance, Arthur (1994); Crouch and Farrell (2004).

https://doi.org/10.1017/9781009022910.008 Published online by Cambridge University Press

the parameters simultaneously, so it is significantly efficient. Third, it uses a normalised gradient-descend rule to perform direct optimisation (as opposed to indirect inference through surrogate models). Fourth, it considers hyper-parameters to improve its efficiency. We also want to highlight the scalability of our algorithm since its performance does not deteriorate exponentially with the dimensionality of the parameter space (i.e., with the number of indicators or policy issues involved). Furthermore, the method achieves high precision levels with enough Monte Carlo simulations in each evaluation. Next, we provide all the relevant details.

5.2 OPTIMISATION ALGORITHM

Let M denote a given number of independent Monte Carlo simulations; $I_{i,-1}$ is the empirical final value of indicator i; and $\bar{I}_{i,-1,m}$ its simulated final value in the m^{th} model run. Then, the expected final value of a simulated indicator i is

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

Then, the α-error of indicator i is

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

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 periods, 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.[5]

[5] The reader may be concerned about how representative can r_i be if a time series is too short. Indeed, this could be problematic for very short time series. In this book,

https://doi.org/10.1017/9781009022910.008 Published online by Cambridge University Press

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 generated with 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 observed initial value of the indicator, so $I_{i,-1} - I_{i,0}$ represents the gap that was closed during the sample period. The intuition behind normalising the α-error is that, in the empirical data, simulating indicators that closed bigger gaps (in the same amount of time) introduced more volatility. Thus, the normalisation helps obtain more stable statistics. This scenario does not happen for the β-error because the associated feature (the probability of success) is always bound to (0,1), so it is not necessary to normalise the corresponding error terms.

Note that both $\hat{e}_{\alpha_i}$ and e_{β_i} can be positive or negative. This setting 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

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

most of our time series have more than 20 observations so, while not ideal, they provide a reasonable sample of successful events. If one would not have such data, a solution would be to group indicators by category, e.g., by the SDG to which they belong. Arguably, the governance and nature of government programmes are not too dissimilar within the same development class, so one could collate all the observed events of success (across indicators in the same category) to construct a more reliable r_i. This estimate would be the same for all the indicators in the same group. While this sacrifices heterogeneity within groups in terms of success rates, it is a reasonable price to pay since we still have the parameter vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\alpha}'$ to account for structural differences between indicators in the same category.

https://doi.org/10.1017/9781009022910.008 Published online by Cambridge University Press

There are two readjustment cases because the direction of adaptation depends on the sign of the error forecast. Similarly, we define the readjustment of α_i' as

$$\alpha_i' = \begin{cases} \alpha_i' \times \max(0.99 - |\hat{e}_{\alpha_i}|, 0.25), & \hat{e}_{\alpha_i} > 0 \\ \alpha_i' \times \min(1 + |\hat{e}_{\alpha_i}|, 1.5), & \hat{e}_{\alpha_i} < 0 \end{cases}, \tag{5.9}$$

and for the β-error is

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

The principle behind these readjustment rules is twofold: (1) penalising deviations and (2) adapting the penalty size as the error shrinks. For instance, in the case of a positive trend, $\hat{e}_{\alpha_i} > 0$ means that the simulated indicator was slower than its empirical value since it ended at a lower value. Therefore, the adjustment is to increase α_i by a fraction not larger than 0.5. As this process continues, the fraction becomes lower than 0.5 because the error decreases, so the size of the readjustment is $|\hat{e}_{\alpha_i}|$. In contrast, for the case of a negative trend, $\hat{e}_{\alpha_i} < 0$ implies that the indicator's expected simulated value did not decrease as much as the observed indicator did during the sample period. Under these circumstances, the calibration procedure reduces α_i and increases α_i' in the next iteration, and in this manner, the reduction of the simulated values becomes more likely. Bounding the adjustments to factors of 0.25 and 1.5 is simply a heuristic rule that allows accelerating the optimisation with respect to an unbounded version of the algorithm. The more technical user could design a hyperparameter optimisation procedure to define the best bounds in a particular application. Putting together these elements, we construct an optimisation algorithm that iterates until reaching a tolerance threshold. We provide its pseudocode in Algorithm 2.

The threshold criterion is a choice variable. From our work, we have found that a criterion that achieves high goodness of fit is to stop the calibration once the worse-performing parameter attains the minimum goodness of fit according to the metrics defined in

https://doi.org/10.1017/9781009022910.008 Published online by Cambridge University Press

Algorithm 2 Calibration pseudocode

```
initialise vectors α, α′, and β with random values;
while a tolerance threshold is not met do
    run M Monte Carlo simulations;
    compute the errors ê_{α_1}, ..., ê_{α_N} and e_{β_1}, ..., e_{β_N};
    foreach indicator i do
        adapt parameters according to Equation 5.8,
        Equation 5.9, and Equation 5.10;
```

Section 5.3. The computational cost is determined partly by the number of Monte Carlo simulation runs in an evaluation: M. How many simulations should one run to calibrate the model? It depends on how conservative one wants to be concerning the average error threshold. The stricter the threshold, the more precision is required, and more precision demands a larger M for obtaining 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 an increase in precision 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 to a higher level after a few iterations. Nevertheless, we can see that, as the algorithm iterates further, all the indicator-specific errors decrease. Furthermore, with higher precision, the error decay becomes smoother.

Increasing the number of Monte Carlo simulations helps achieve lower errors, yet this is at the expense of higher computational costs. To mitigate this cost, we introduce three hyperparameters and a routine that allows setting M automatically as the optimisation proceeds. The procedure follows the idea that errors tend to be high during the first evaluations (as the parameters

https://doi.org/10.1017/9781009022910.008 Published online by Cambridge University Press

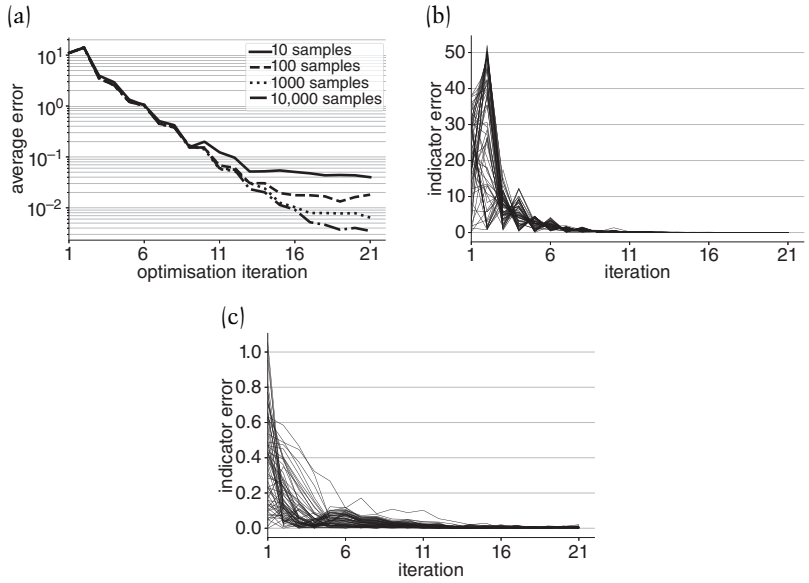

FIGURE 5.1 Error minimisation behaviour. (a) Average error, (b) α-errors, and (c) β-errors.

Notes: Figures 5.1b and 5.1c display the errors associated with each indicator, calculated at the indicator level.

Sources: Authors' calculations using Mexican data from the 2021 Sustainable Development Report.

are initially set at random). Thus, a small M is enough to generate coherent responses when initially readjusting the parameters. The hyperparameter routine consists of making a few Monte Carlo simulations and then increasing M after a certain number of evaluations. From our experience, an initial $M = 10$ for 100 evaluations is enough to drop the average error substantially. The number of evaluations using a low M is the first hyperparameter. Then, the routine increases M periodically. We have found that increments of 1,000 Monte Carlo simulations with every evaluation are a good balance between error reduction and computational cost for the applications presented in this book. The size of the increments and how frequent they are define the second and third hyperparameters, respectively. Here, we have determined the values of the three hyperparameters by building experience through trial and error. This approach has been enough

https://doi.org/10.1017/9781009022910.008 Published online by Cambridge University Press

to calibrate the model for a country in a few seconds. However, one could also design more sophisticated routines to optimise the hyper-parameters. Something that we leave for the technical enthusiast.

5.3 GOODNESS OF FIT

We have seen that our calibration procedure is effective in minimising the different errors in the model besides being computationally efficient. However, how good is this optimisation for fitting the empirical features of interest? Any quantitative method requires goodness of fit or accuracy metrics to address this question. The construction of such metrics usually obeys particular characteristics of the problem at hand. For example, linear regressions use the R^2 to get a sense of how much the independent variables included in the model contribute to explaining the variance observed in the dependent variable. Similarly, the ratio of correct predictions to input samples is commonly 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 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 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 low fitness values.[6]

[6] When testing alternative calibration methods, we find that they yield several indicators displaying negative values for Ψ_{α_i}. This scenario does not happen in our algorithm.

https://doi.org/10.1017/9781009022910.008 Published online by Cambridge University Press

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 Equation 5.4.

To show the high goodness of fit obtained from our calibration procedure, we calculate the Ψ_{α_i} and Ψ_{β_i} of each indicator and every country in the SDR dataset. For illustration purposes, we bin them into different levels and plot their frequencies in Figures 5.2a and 5.2b,

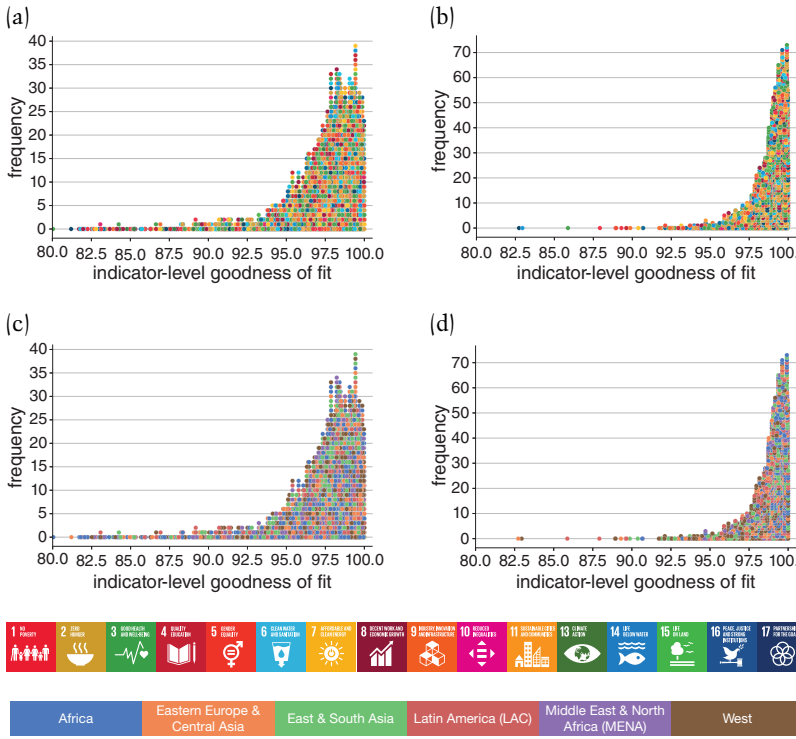

FIGURE 5.2 Distribution of goodness-of-fit metric by SDG and country group. (a) Ψ_{α_i} by SDG, (b) Ψ_{β_i} by SDG, (c) Ψ_{α_i} by country group, and (d) Ψ_{β_i} by country group.

Notes: The goodness of fit is in percentage.

Sources: Authors' calculations with data from the 2021 Sustainable Development Report.

https://doi.org/10.1017/9781009022910.008 Published online by Cambridge University Press

colouring them by SDG. Likewise, Figures 5.2c and 5.2d show the same information but coloured by country group. In these plots, we have set the threshold criterion at 80%. In other words, the calibration stops when the worst goodness of fit is 80%. Importantly, we observe that the overwhelming majority of parameters (i.e., indicators) exhibit goodness of fit above the 95% threshold.

5.4 ON STATISTICAL CONFIDENCE AND TESTING

Now that we have discussed the model's calibration procedure, it is important to clarify the distinction that exists in the literature between this concept and that of estimation. In the old days of simulation practices, the objective of calibrating a model was merely to produce a plausible representation of reality. Hence, in the past, no attempts were made to produce statistical inferences concerning the model's parameters or the simulation outcomes (Gatti et al., 2018). In contrast, in econometric practices, the estimation objective has traditionally been to infer the correct specification of the stochastic process generating real data. In other words, the idea is that the model's parameters resemble their true values, at least when using large samples of data (i.e., that the estimated parameters are consistent). Nowadays, this distinction has become somehow blurred with the new and more sophisticated analytical devices for calibrating the parameters of computational models through simulations.

On the one hand, model-specific goodness-of-fit metrics are now more commonly used to specify the quality of the adjustment to the data. Then, this and other distance metrics are employed to study the statistical relevance of the models' results and their capability to replicate stylised facts and statistical regularities observed in the data (e.g., distribution moments). On the other hand, by invoking specific sources of uncertainty, such as measurement problems in the data, one can produce confidence intervals for many metrics through Monte Carlo simulations. These metrics reflect different outcomes of the model that might be of interest for explaining causal mechanisms and inferring the impact of policy interventions. In fact, in Chapter 9,

we demonstrate how to construct confidence intervals that capture the uncertainty of government expenditure.

In this section, we unpack and discuss several basic concepts conflated with the teachings of statistics in the social sciences, such as confidence intervals and hypothesis testing. However, these concepts have to be reassessed in simulation methodologies to be meaningful since they rely on building null models for producing statistical inferences that are also theoretically insightful (not just relying on the statistical analysis of random variables). In other words, instead of contrasting inferred parameters with zero-value null hypotheses, one has to compare a metric using simulation outcomes under a null model (e.g., a benchmark setting) with the corresponding metric in a counterfactual simulation.

5.4.1 Confidence Intervals

Constructing confidence intervals for a metric or statistic of interest refers to quantifying the uncertainty of the estimation of such metric. With this aim in mind, the modellers have to assume the source of such uncertainty. Then, they need to formulate a device that 'propagates' these uncertainty effects throughout the simulation exercise up to the statistic of interest. Usually, one can obtain a distribution of the metric or statistic as a result of propagating this uncertainty. In computational modelling, constructing confidence intervals is not always straightforward. The interdependencies of complex systems make it extremely difficult to track the propagation from the source of uncertainty to the statistic.[7] Accordingly, in computational models like that described here, researchers need to pay attention to the following issues: (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 concluded that one of the most

[7] Contrary to the linear models of econometric analyses.

https://doi.org/10.1017/9781009022910.008 Published online by Cambridge University Press

common sources of uncertainty is the quality of the indicators (i.e., measurement uncertainty). In this respect, an approach is to propagate the uncertainty behind observed indicators through an ensemble of calibrations. Hypothetically speaking, one could randomise the original indicator data with the suggested probability distribution.[8] Then, in the next step, it is possible to generate one calibration for each randomised dataset. Through this procedure and Monte Carlo simulations, one could generate distributions for the model's free parameters requiring calibration. Then, for each of them, one could produce the metrics of interest to describe the statistical significance of the estimation using the empirical data.

As experienced readers using development-indicator data are aware, this method for quantifying uncertainty is generally unfeasible due to a limited understanding and characterisation of the source of uncertainty. Thus, in a more realistic setting, one needs to model the uncertainty in the data.[9] For example, in Guerrero and Castañeda (2022), we use the inter-temporal volatility of each indicator as a proxy for the non-observed difficulties associated with the collection of such indicators. The logic behind this argument is that least-developed countries exhibit more volatility in their indicators. This trait is, to some extent, a consequence of having more fragile infrastructure and methods for collecting data than those of developed nations.[10]

Another example of this quantification is in our work in Guariso et al. (2023b), where we model each indicator time series through a Gaussian process. In that work, we employ indicator-specific models to generate randomised synthetic indicators that we later use to create the calibration ensembles. Finally, in Guerrero and Castañeda (2020), we analyse model uncertainty by implementing different model specifications for the government heuristic. While this exercise focuses on

[8] Something that data providers rarely supply.

[9] As done when assuming a normally distributed error in a regression analysis.

[10] Moreover, their policies and government programmes are more erratic and, as a consequence, their effect on indicators is fickler. Hence, besides measurement problems, such volatility captures a policymaking uncertainty.

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 concerned with the statistical confidence of the parameter vectors $\boldsymbol{\alpha}$, $\boldsymbol{\alpha}'$, and $\boldsymbol{\beta}$. This perspective contrasts with the regression framework to which quantitative social scientists are more accustomed. The estimated coefficients of regressions usually carry an explicit meaning in terms of the average impact of explanatory variables and, hence, ought to be statistically tested against the null hypothesis of being zero valued. On the contrary, in PPI and other areas of computational social sciences, often, testing for the significance of the model's parameters is not something worthwhile to pursue. Under this framework, parameters do not always carry a meaning conducive to explaining social phenomena or guiding policymaking. Thus, defining a null hypothesis is a model-specific task. Therefore, depending on the problem at hand, the main interest lies in testing the statistical relevance of more meaningful metrics such as development gaps, time savings, and efficiency gains.

In a computational framework like PPI, building a null hypothesis for these statistics requires a deeper understanding of the theory behind the model, its simulation capabilities, and the domain of application. Instead of testing the validity of a zero-value statistic, one can explore the observed empirical value of a metric and then check if such an outcome also happens in a world without specific mechanisms or the usage of different intervention policies. This way of thinking – creating a null model or simulation instead of deriving the distribution of parameters – is common in several fields of study, such as network science, statistical mechanics, and computational biology. In essence, a null model is a generative model of the phenomenon under study, but where the causal factor of interest or the mechanism being tested is 'missing' or 'deactivated'. In the natural sciences, null models are usually created via simple stochastic processes. In social sciences, creating null models is not so straightforward as,

https://doi.org/10.1017/9781009022910.008 Published online by Cambridge University Press

even when removing the causal mechanism of interest, it is necessary to account for other detailed social mechanisms and behavioural elements that are likely to generate the empirical data;[11] otherwise, a null model based on purely random behaviour would tend to overstate the significance of an estimate. This is a very common problem in fields such as social physics and econophysics. Because of the nature of null models and the challenges to implementing them, it is no coincidence that social scientists whose only exposure to quantitative methods has been through the tradition of statistical and econometric analyses often find this approach more difficult to comprehend.

In general, the type of statistical tests that will be of most interest in this book has to do with performing counterfactual simulations. In principle, such a scheme requires producing two variants of runs: one generating a set of benchmark simulations and another simulating the model with the assumed counterfactual. For example, as we show in Section 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 sets of simulations with and without a spillover network. Beyond visual inspections, in which one compares empirical and artificial data, the testing for the statistical significance of the results has a twofold interpretation. First, through simulations that only employ the observed data as input, one can test using non-parametric methods whether the model can replicate certain statistical regularities. Second, by including uncertainty in the input and its propagation through the null model, one can produce distributions and confidence intervals of the metric under study. This approach allows us to compare the null hypothesis with the metric generated through a counterfactual exercise.

Note that there is no one-size-fits-all method to formulate statistical tests. Each problem requires carefully thinking about the meaning of the benchmark and the counterfactual and how to use the information derived from simulated distributions (e.g., a difference-

[11] This is why agent computing is so well suited for these problems.

https://doi.org/10.1017/9781009022910.008 Published online by Cambridge University Press

in-means test, a paired test, or a custom-built non-parametric test). Throughout the book, we present some examples in which we have devised different strategies; thus, all application chapters come with a section named 'simulation strategy'. 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 quantitative social sciences.

5.5 VALIDATION

The topic of validation is critical for discriminating between competing models and generating knowledge. In the computer simulation literature, the meaning of validation varies among fields and authors. Sometimes it refers to methods for checking the model's theoretical consistency, and on other occasions to procedures for testing the model's reliability (i.e., reality checks, generalisability, or robustness). Furthermore, a model's validation can be framed and tested using multiple methodologies. Hence, in this section, we first identify the concepts of validation that are relevant in an agent-computing context. Something that we have previously done in the framework of PPI (Guerrero and Castañeda, 2020), and which we discuss here.

In computational modelling (not just ABMs), the validation process is done through several schemes that have evolved as more data and new methods have become available (Fagiolo et al., 2019). Perhaps one of the pioneering works in categorising several of these conceptions is Carley (1996), which identifies up to eight levels of validation. By today's standards, Carley's validation levels can be classified hierarchically, with external and internal validations at the core of the taxonomy and different varieties 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 ABMs typically means replicating one or more quantitative statistical regularities or stylised facts (e.g., distributions,

https://doi.org/10.1017/9781009022910.008 Published online by Cambridge University Press

moments, or correlations) by generating them from the bottom up. Importantly the reader needs to be aware that matching a stylised fact when validating the model should be disconnected from the calibration exercise. Otherwise, the validation is trivial. Therefore, the stylised facts to replicate should come from a dataset (testing set) independent of that used to calibrate the model (training set). In our early work (Castañeda et al., 2018; Guerrero and Castañeda, 2020, 2021a), we externally validate earlier versions of the PPI's model by replicating two well-known statistical patterns: (1) the skewed distribution of budgetary changes and (2) the negative relationship between development and corruption. Here, we would like to revisit those validation strategies briefly and show that they still hold with this new model and data.

First, we consider the distribution of budgetary changes. A large body of literature in political science and public administration has documented non-normal tails in the distribution of changes in the government budgets (total changes and disaggregated into policy issues) (Jones et al., 1998; John and Margetts, 2003; Jones and Baumgartner, 2005; Jones et al., 2009). This evidence is not convincing enough to suggest any particular distribution for generating the budgetary data. Nevertheless, an indisputable feature is that changes in government expenditure do not follow distributions with exponentially decaying tails like in a normal distribution. So, the question is whether our model can generate simulated budgetary changes that exhibit similar tails without the influence of empirical data on budgets.

To demonstrate that this is the case, we perform 10,000 Monte Carlo simulations with fully randomised data. That is, in each simulation, we generate (1) a random number of indicators (between 50 and 200; besides 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 presents a log-log scale, suggesting that the tails are non-normal because they show a linear-decaying pattern. We build the graphic by

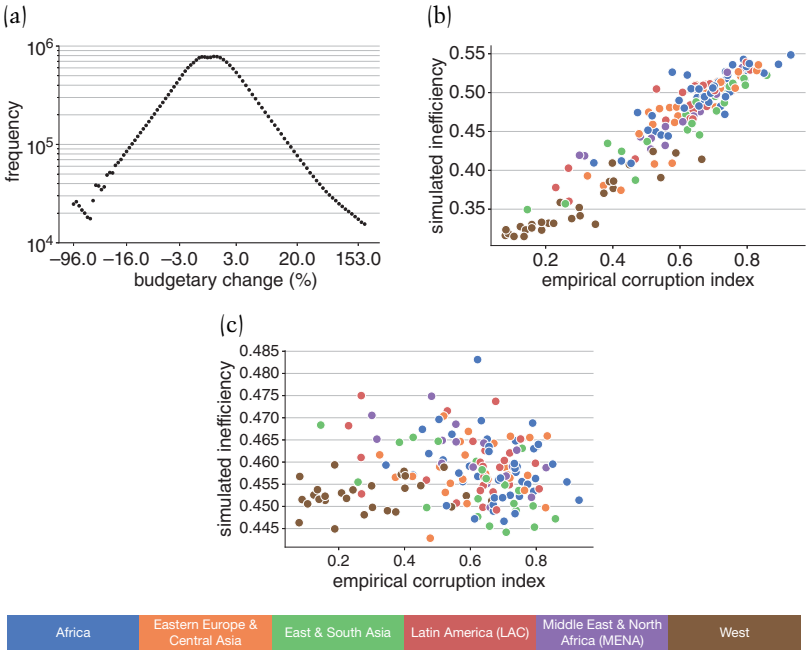

FIGURE 5.3 External validation. (a) Distribution of budgetary changes, (b) differentiated governance, and (c) homogeneous governance.
Notes: Figure 5.3a shows the simulated distribution of budgetary changes at the level of expenditure programmes. Figure 5.3b compares Transparency International's corruption index against the model's endogenous level of corruption (they have a linear correlation larger than 93%). Figure 5.3c shows the association between Transparency International's corruption index and the model's corruption level under a counterfactual where the governance parameters φ_i and τ_i equal 0.5 for every country.
Sources: Authors' calculations.

taking all the changes of each 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 in a bottom-up fashion from the model.

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

https://doi.org/10.1017/9781009022910.008 Published online by Cambridge University Press

redundant with the model's endogenous variable of inefficiency $P_{i,t} - C_{i,t}$, which one can interpret as corruption (at least partly). Thus, while the corruption index is left out of the study, we can use it to validate the model by assessing how well the endogenous inefficiency variable matches this 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 lost in inefficiencies. Besides, we invert the directionality of the corruption index so that higher values denote less corruption.

In this way, if the model's emergent inefficiency across countries exhibits a positive correlation with the corruption index, we can validate PPI's public governance mechanisms. Precisely, Figure 5.3b shows a strong association between the model's inefficiency and the corruption index. Their linear correlation is larger 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 vanishes because the public servants' responses, in terms of their contributions, are not distinguishable across countries with the same quality of public procurement.

5.5.2 *Internal Validation*

Internal validation tests attempt to show that the theoretically expected outcomes (whether externally validated or not) are sensitive to the social and behavioural mechanisms specified in the model. That is to say, internal validation checks whether the assumed micro and systemic mechanisms have a theoretical meaning in the performance of the computational model.[12] Accordingly, when

[12] A related but different concept, often confused with internal validation, is verification. The verification of a computational model relates to revising whether there are programming errors (bugs) or artefacts (for further details on these issues, Castaneda (2021b, Ch. 20)). Artefacts consist of implications in the model's outcomes produced by assumptions considered auxiliary at the moment of its creation. In

https://doi.org/10.1017/9781009022910.008 Published online by Cambridge University Press

internal and external validation procedures deliver positive results, the model's causal mechanisms offer insightful information to explain the social phenomenon under study.[13] Chapter 4 provides one example of internal validation in Figure 4.2 by showing that the bureaucrats' learning, as a response to public governance, is consistent with expected real-world behaviour. Here, we provide further evidence validating our model internally.

In Castañeda et al. (2018) and Guerrero and Castañeda (2020), we internally validate earlier versions of the PPI 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 functionary because they would be able to 'disguise' their inefficiencies through the 'inflated' indicators. Thus, theoretically, one should expect that the more positive spillovers received by public functionaries, the less efficient they will be (i.e., the size of their contributions will be lower). Notice that the connection between spillovers and efficiency is not easy to trace mathematically due to the complexity arising from the interactions embedded in the model and its behavioural components. However, by employing simulations, we can confirm that these mechanisms exert the expected influence on the functionaries' behaviour and the system. To demonstrate that this behaviour takes place in our model, we provide two alternative tests to validate its occurrence.

First, we validate, at the micro level, the spillover→efficiency mechanisms. For that, we show that public servants who receive more positive spillovers tend to be less efficient. For functionary i, the average amount of spillovers received across M simulations is

contrast to substantive assumptions, auxiliary assumptions do not attempt to describe reality. Their only purpose is to close up the model in a simplified manner.

[13] In an equation-based modelling approach, internal validation corresponds to checking whether the equations (or theorems) are properly solved (or proved), while verification refers to clarifying if the model uses 'the right' equations to explain the phenomenon (Midgley et al., 2007).

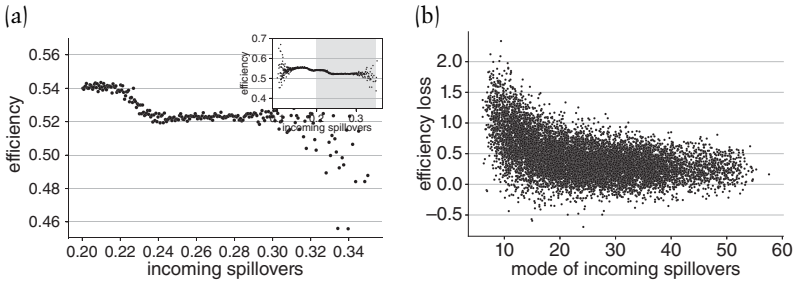

FIGURE 5.4 Internal validation. (a) Functionary level and (b) aggregate level.

Notes: Figure 5.4a shows inefficiencies at the agent level. Figure 5.4b indicates the aggregate efficiency loss.

Sources: Authors' calculations.

$\sum_{t,m}^{T,M} S_{i,t,m}/(T \times M)/N$.[14] We compute this quantity for each functionary across 10,000 sets of $M = 100$ Monte Carlo simulations each. Each set of Monte Carlo simulations employs the same parameter values. The latter are randomly generated in the same fashion as we do for the external validation procedure mentioned above. At the same time, we compute the functionary-level average efficiency $\sum_{t,m}^{T,M}(C_{i,t,m}/P_{i,t,m})/(T \times M)$. Figure 5.4a shows that our model elicits inefficiencies through spillover effects. The plot suggests a general negative non-linear relationship between the spillovers received and the efficiency level (they exhibit correlations above −75% within each set of simulations). The inset figure shows the complete relationship, with large volatility at extreme spillover levels. Thus, we can claim that, for the most part, our model is internally valid concerning the functionary's incentives and their response to network effects.

Second, we also validate the spillover→efficiency relationship through aggregate evidence. For this, we use the same simulation as before but calculate the aggregate levels of incoming spillovers and efficiency with the following expressions: $\sum_{i,t,m}^{n,T,M} S_{i,t,m}/(n \times T \times M)$

[14] We divide the spillovers over N because, to properly randomise the Monte Carlo simulations, we compare simulations with different numbers of indicators, which affects how many spillovers occur in the system.

https://doi.org/10.1017/9781009022910.008 Published online by Cambridge University Press

and $\sum_{i,t,m}^{n,T,M} C_{i,t,m}/(B \times M)$, respectively. A distinctive feature of this exercise is that, for each set of $M = 100$ Monte Carlo simulations, we assemble another one with the same parameters but without the network. In this manner, no spillovers occur in the counterfactual simulations. Once we have the aggregate efficiency in both sets, we can compute the difference between the efficiency in a setting without a network and the efficiency when there is one. A positive difference means that the economy behaves more efficiently without positive spillovers because functionaries have fewer incentives to engage in inefficient activities. We call this difference an efficiency loss. Figure 5.4b shows the result of this exercise. The results are striking. Not only because an overwhelming majority of the efficiency losses are positive, but also because there is a clear negative association between the size of the loss and the average level of spillovers received by the public servants. The latter outcome indicates that when positive spillover effects tend to be large, for most of the functionaries, there are fewer incentives to misbehave since they face more homogeneous conditions and, thus, it is more difficult for them to hide inefficiencies. Hence, with these results, we provide evidence that validates our model internally at different levels of aggregation.

5.5.3 *Soft Validation*

Soft validation is probably the most common form of testing models in the agent-computing literature, as it involves a qualitative assessment of an observed pattern. These procedures differ from other external validation methods because the soft assessments do not use a formal metric but a qualitative judgement. This approach is common when attempting to make a 'proof of concept' through a simulation. Seminal examples of this kind can be found in Schelling's model, where aggregate segregation patterns emerge from tolerant individuals (Schelling, 1971), or in Axelrod's cultural model, in which a polarisation pattern ensues despite individuals exhibiting deep social interactions (Axelrod, 1997b). This validation criterion is common in toy models attempting to describe empirical patterns but

https://doi.org/10.1017/9781009022910.008 Published online by Cambridge University Press

without referring to statistical tests to compare means or distributions. Some artificial stock market models exemplify the use of this validation approach since simulated time series of returns replicate observed patterns in real data, such as extreme values and clustered volatility.

When studying policy coherence in the context of policy priorities (Guerrero and Castañeda, 2021b), we provide a 'soft' validation exercise for a variant of our model. This validation consists of estimating an index of policy coherence for countries known to have been coherent with their governments' official discourse when attempting to emulate the development pattern of specific nations through budgetary prioritisation. For example, the case of Korea following the steps of Japan, or Estonia adopting the Nordic development model. When the coherence index is consistent with the qualitative narrative of successful emulations of more developed economies, the model's outcomes provide further evidence favouring PPI. Such exercise requires a balanced cross-national panel of development indicators and a verifiable narrative as to why such a qualitative pattern is likely to emerge. In Guerrero and Castañeda (2021b), such a narrative is provided by Akamatsu's flying geese description of changing development patterns (Aikman et al., 2019), scholarly work on the countries under study, and the public discourse of government officials.

5.5.4 Stakeholder Validation

In the literature on participatory modelling (Becu et al., 2003; Gurung et al., 2006; Guyot and Honiden, 2006; Barnaud et al., 2013; Barreteau et al., 2014), researchers involve the stakeholders of a problem in the modelling process. This engagement is done through role-playing games, experiments, consultations, workshops, and feedback activities, to mention a few possibilities. The idea is that stakeholders can help to determine the nature of the data, which mechanisms 'actually' take place in decision making, and verify that the model's features make sense in general.

Our work with policymakers in various consulting projects (Castañeda and Guerrero, 2020a,b,c; Guerrero and Castañeda, 2020; Sulmont et al., 2021; Castañeda and Guerrero, 2022a,b) has allowed us to build a certain level of stakeholder validation. For instance, during a collaboration with the UNDP-Mexico, several stakeholders from the federal- and state-level governments and NGOs participated in multiple workshops. In these events, we presented and discussed, in detail, the model's methodology, data, and results. The stakeholders took part in exercises to classify the indicator dataset into instrumental or collateral and to reach a consensus in terms of the model's specification. In this and other projects, stakeholders also expressed their opinions on the degree of flexibility in budgetary allocations since fiscal rigidities are something that occurs quite often in public administration. Moreover, the expert opinion of public officials from treasuries and ministries of finance on how to conceive development plans was also helpful insight to specify the behavioural components of PPI. Hence, stakeholders provided early feedback on the use of the data available and the configuration of the model.

Besides their early involvement, stakeholder validation also took place during the elaboration of policy experiments and while writing our policy reports. Through a review process provided by the participants of our workshops, we refined some analyses and interpreted their results attending to the relevant problems of these policymakers. An example of this process is the publication of several reports on the application of PPI to different contexts (Castañeda and Guerrero, 2020a,b,c; Gobierno del Estado de México, 2020; Sulmont et al., 2021).

5.6 STATISTICAL BEHAVIOUR

As we have explained, agent-computing epistemology differs from that underpinning more traditional statistical methods. The intricate data-generating processes that one can specify in an ABM may generate complex dynamics requiring extensive Monte Carlo simulations to produce outcomes with a proper characterisation. This feature has

implications for parameter calibration and the estimation of counterfactuals. Thus, depending on the statistical behaviour of the model's variables, one needs to select an appropriate simulation strategy, as discussed in Chapter 2. While providing an exhaustive account of the model's statistical properties is beyond the scope of this book, we would like to discuss two features that provide solid statistical grounds for the Monte Carlo strategy pursued in this book.

Overall, we can perform independent runs of a model when we produce Monte Carlo simulations with the same set of parameter values. In this section, we would like to demonstrate how, with this approach, (1) we obtain consistent impact estimates and (2) we can recover the true parameters by calibrating the model against simulated data. While these two features are not the only way to demonstrate the "good" statistical behaviour of PPI, they are some of the most discussed among quantitative researchers when it comes to arguing in favour of the adequacy of their methods. Thus, with these elements, we provide further arguments for the reliability of PPI in supporting evidence-based policy guidelines (see the appendix of Guerrero and Castañeda (2022) for additional analyses).

5.6.1 Testing for Synthetic Counterfactuals

As mentioned in Section 2.3.3, for an ABM counterfactual to be valid in measuring a causal impact, it must describe a system that behaves similarly to that of the baseline. In complex systems, random initial conditions or a sequence of random factors in the interaction process might create non-linear effects that move the system in opposite directions. These two types of randomness might affect the results in the intervened and baseline simulations. In particular, they could produce divergent paths that are not comparable for measuring causal impacts when they come from different systems in a random sense. In the former case, uncontrolled randomness could produce distributions of the impact metric with two or more modes due to the sensitivity of the model's initial conditions. In the latter case, the realisation of opposite random shocks might produce extreme impacts

when comparing two simulation runs, one for each variant (activated or deactivated interventions). Therefore, to test for the presence of these complications in PPI, we produce Monte Carlo distributions of the average differences in the final values of intervened and baseline simulations for a set of development indicators.

Figure 5.5a shows bell-shaped distributions for seven synthetic indicators.[15] These distributions do not suggest random non-linearities affecting the intervened and baseline populations. Figure 5.5b shows the results of Monte Carlo simulations in which we establish random seeds in pairwise comparisons between the two types of synthetic trials. As expected, in this alternative procedure, no complications appear when looking at the distributions of our impact metric. Moreover, irrespective of using a fixed random seed or not, we generate similar mean impact metrics in both procedures and, most importantly, very close Monte Carlo distributions. Accordingly, we can argue that the synthetic counterfactuals we implement throughout the book are valid instruments to assess causal impacts.

5.6.2 *Parameter Recovery*

A common concern addressed in statistical models is whether an estimation procedure can recover the true parameters of the underlying stochastic process. We must highlight that this way of assessing the quality of a model and its estimation method comes from a very particular way of thinking about models; usually, one in which the production account of causation is dominant. In the ABM literature, recovering parameters is not always straightforward; because the dynamics in an ABM emerge in a bottom-up fashion from the micro-level interaction of agents. Consequently, micro-level parameters are mapped into macro-level outcomes, which are empirically observable and used for determining error functions.

[15] We choose to illustrate this behaviour with only seven indicators as it is easy to appreciate their distributions in a single plot. However, these results also apply to a larger set of indicators.

https://doi.org/10.1017/9781009022910.008 Published online by Cambridge University Press

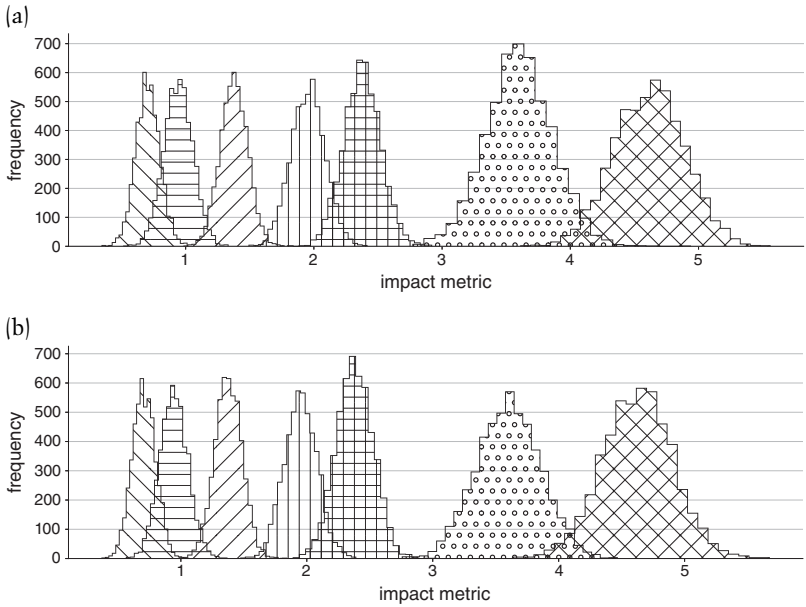

FIGURE 5.5 Well-behaved impact estimates. (a) Impact under fully randomised trials and (b) impact under pairwise randomised seeds.
Notes: We present the results from simulating indicators and calculating the impact metric of counterfactuals. All parameters are randomly determined and fixed across all simulations. The counterfactual consists of duplicating government expenditure. The impact metric consists of the difference between the indicator's average final value from 500 Monte Carlo simulations under the counterfactual and the mean final value under the baseline. We generate the set of impact metrics by repeating the previous calculation 5,000 times. Figure 5.5a corresponds to simulations where the seed is fully randomised. Figure 5.5b refers to simulations in which the seed is the same for a pair of baseline and counterfactual estimates but different between different pairs.
Sources: Authors' calculations.

Furthermore, if the system exhibits qualitative changes (such as transition phases and discontinuities) under certain parameter combinations, the probability of recovering true parameters could be low. For this reason, researchers in the agent-computing literature often rely on direct microdata imputation to explore how the model's simulations respond in the parameter space. Often, what a statistician may interpret as an overfitting problem (because different combinations of parameters may yield the same outcome) could be, for a

https://doi.org/10.1017/9781009022910.008 Published online by Cambridge University Press

computational modeller, a sign of equally valid states of the world (so the relevant type of analysis is deriving the probability distribution over these states). Therefore, while the exercise of parameter recovery would be a desirable feature in an agent-computing model, its absence does not necessarily speak of a poor specification or a weakness in the estimation method.

Aware of the differences between the traditional statistical and the agent-computing points of view, we intentionally developed PPI to avoid free parameters at the micro level. Conciliating both views would require deciphering the mapping of these unknowns into relevant errors, a task that is not always easy to carry on. Thus, by making all behavioural parameters endogenous, we bring our model closer to the statistical tradition of a more direct mapping between outcomes and parameters. Next, we would like to illustrate how one can perform the parameter recovery to demonstrate the empirical strengths of PPI. Notice that $\boldsymbol{\alpha}$ and $\boldsymbol{\alpha}'$ are associated with the same α-error; thus, there is an identification issue. However, this is not a drawback for us, as it would be in regression analyses. We have previously discussed that, in agent-computing, the aim is not to interpret parameters because they do not necessarily convey information such as average effects. Instead, agent-computing researchers calibrate these parameters and, later on, perform counterfactual analyses. Thus, instead of focusing on the true α_i and α_i' of indicator i, we are concerned about recovering their difference $\alpha_i - \alpha_i'$, as this is the key parameter determining the trend feature of the data.[16]

The procedure is straightforward. We randomly define the number of indicators, their initial conditions, those that are instrumental, the governance parameters, a spillover network, and the parameter vectors $\boldsymbol{\alpha}$, $\boldsymbol{\alpha}'$, and $\boldsymbol{\beta}$. We are interested in vectors $\boldsymbol{\alpha} - \boldsymbol{\alpha}'$ and $\boldsymbol{\beta}$, so we

[16] In earlier versions of PPI, the model only generated positive trends, so α_i' did not exist. Under this specification, the identification issue is not a problem as the α-error is only associated with parameter α_i. When generalising PPI to account for negative trends in the indicators, it is necessary to introduce α_i', so the identification shifts to its difference with respect to α_i.

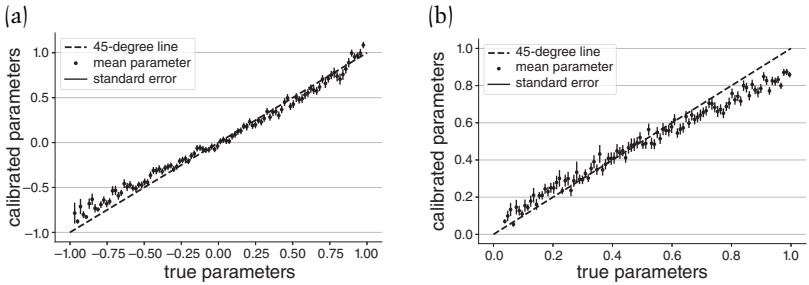

FIGURE 5.6 Parameter recovery. (a) Parameter difference $\alpha_i - \alpha_i'$ and (b) Parameter β_i.

Notes: The 45-degree dashed line indicates a perfect match. Each dot corresponds to the average calibrated parameter in each bin of the true parameters. The vertical lines denote the standard error in each bin.

Sources: Authors' calculations.

say that they are the 'true' parameters to be recovered. Next, using these true parameters and the other random data, we run a single simulation to generate one-time series with a similar length as those from empirical data. Then, with these synthetic data, we calculate the success rate γ_i, the indicators' final values, and calibrate the model using our optimisation algorithm. Once calibrated, we compare the parameters that we obtain against the true parameters. If we make a scatter plot and most dots lie close to the 45-degree line, the result indicates that we recovered the true parameters reasonably well.[17]

Figure 5.6 shows the results of repeating the above procedure for 100 different random parameterisations. This setting implies that, in total, we collect approximately 10,000 pairs of true and calibrated values for each parameter. The plot presents the results by binning

[17] Since we generate only one set of simulated time series for a given collection of random parameters, one should not expect a perfect match between the true and calibrated parameters. The reason is that a single realisation of the model may carry idiosyncratic fluctuations that are not representative of the average behaviour of the model. One could, instead, generate multiple sets of time series and compute average success rates and average final values. However, this procedure defeats the purpose of the exercise as, in the real world, we only see one realisation of the underlying mechanisms. Thus, if the overall pattern across indicators reflects a consistency between calibrated and true parameters while using a single set of time series, it is possible to argue that there is strong evidence in favour of the model's robustness.

https://doi.org/10.1017/9781009022910.008 Published online by Cambridge University Press

the true parameters and displaying the average calibrated parameter in each bin. Figure 5.6a shows the case of the differences $\alpha_i - \alpha'_i$, whereas Figure 5.6b displays the matches for β_i. In both cases, most of the scatter plot lies near the 45-degree line, meaning that our calibration procedure recovers the true parameters of the model reasonably well. The vertical lines depict the standard error intervals, so their absence in most dots denotes a high matching density between true and calibrated parameters. This result provides statistical validity to PPI and increases our confidence in the inferences drawn from this toolkit.

5.6.3 *Overfitting*

A typical concern of any reader facing a model containing a large number of parameters is the potential problem of overfitting. Overfitting means different things in different communities. In machine learning, for example, it usually refers to the poor capacity for making accurate predictions outside of the sample used to train a model. In agent computing, overfitting commonly relates to the insensitivity of the outcome variables to changes in specific parameters. Intuitively, if a model has an 'excess' of parameters, some may be redundant and contribute with no new information to generate the same outcomes. This scenario can occur when the amount of parameters exceeds the number of error functions used to calibrate or estimate a model. In PPI, this logic could apply to the parameters α_i and α'_i, since we calibrated them against the same error. Alternatively, because of the stochasticity of the model and the calibration method, one may say that different calibration runs may lead to very different parameter combinations.

In terms of parameter insensitivity, it is easy to see that changing α_i and α'_i induces major changes in the associated indicator, as their difference determines the trend component (see Figure 4.1). Regarding the possibility of multiple parameter combinations yielding the same fitting, we can argue that this scenario is unlikely. Because of the interdependencies and the model's vertical mechanisms, there is a set of implicit constraints preventing numerous heterogeneous combinations of α_i and α'_i from being feasible parameters. To test this, we could

https://doi.org/10.1017/9781009022910.008 Published online by Cambridge University Press

verify if, across different calibrations, the distribution of a parameter is bell shaped. Intuitively, a high concentration of estimations around the mean implies a tendency to produce a particular parameter value.

Let us demonstrate this point by doing a simple exercise that involves performing a large set of independent calibrations for a synthetic dataset. We randomly create a set of 100 artificial indicators, a network, and governance variables, and perform 1,000 independent calibrations. Then, for each indicator, we test if the distribution of α_i is unimodal. We do the same for α_i' and β_i. A suitable test for this exercise can be found in Siffer et al. (2018), which yields a simple metric called the 'folding statistic'. When the folding statistic is greater than 1, we are in the presence of a unimodal distribution. In Figure 5.7, we show all the mean parameter values and their corresponding folding statistics. Notice that the vast majority of the

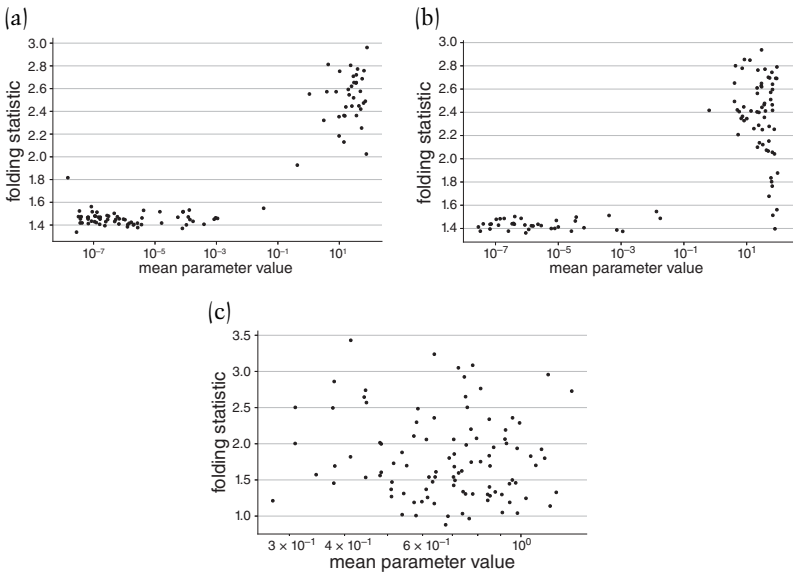

FIGURE 5.7 Overfitting detection ('folding statistic'). (a) Parameters α_i, (b) Parameters α_i', and (c) Parameters β_i.
Notes: Each panel contains 100 values (one per indicator) for the mean parameter and the folding statistic.
Sources: Authors' calculations.

https://doi.org/10.1017/9781009022910.008 Published online by Cambridge University Press

parameters exhibit unimodal distributions, suggesting that potential overfitting is unlikely.

5.6.4 *Time Equivalence*

A couple of clarifications regarding parameter T are in order. While T represents the number of simulation periods, the reader may be interested in producing simulations with the temporal equivalence of calendar time. This equivalence is straightforward since we can establish it using only the information about the coverage time of the sample period. For example, if $T = 20$ and the data cover 10 years, then each computational period represents 6 calendar months. Another clarification is that T should not be too small. The reason for this is the learning process of the agents. For PPI to produce consistent simulations in a Monte Carlo setting, agents need to emerge a social norm of inefficiency. Accordingly, we need enough simulation periods for this emergent property to be a likely outcome. From our experience, agents learn, and social norms stabilise after $T = 25$. Furthermore, in the appendix of Guerrero and Castañeda (2022), we show that PPI's results are robust to different values of $T \geq 25$. Throughout this book, we use $T = 50$.

5.7 ON INTERDEPENDENCY NETWORKS

The last issue to be discussed in this chapter is the network of interdependencies between indicators. In Chapter 3, we argued that one of the empirical challenges to be overcome in the future relates to the ability to produce reliable network estimates of interdependencies between indicators. This challenge is especially troublesome given the coarse-grained nature of development-indicator data. We also explained that, to PPI, these networks are an exogenous input as they reflect long-term structural relationships or conditional dependencies. Hence, we do not expect the empirical configuration of these networks to change considerably when looking at data covering two decades. Thus, PPI is agnostic of the particular method of choice to obtain these networks. It is up to the researcher to devise

https://doi.org/10.1017/9781009022910.008 Published online by Cambridge University Press

empirical strategies to make a proper selection with the available data. Nevertheless, readers would surely appreciate some initial pointers regarding potential methods to construct such networks. Thus, we describe a Bayesian method in this section that we have employed in the past.[18]

Notice that, although our preferred choice is to specify a quantitative network structure, there exist qualitative approaches – based on expert opinions – that may be suitable when the data available is of poor quality (Weitz et al., 2018). Of course, qualitative frameworks suffer from scalability issues as it is neither cheap nor logistically feasible to gather experts in hundreds of different development dimensions. In addition, these experts are not readily available, especially when there is an urgent need for evidence-based policy prioritisation. Throughout the development of PPI, we have been fortunate enough to have data suitable for quantitative methods. Furthermore, we combine quantitative estimates with a qualitative validation/correction approach in multiple projects. In particular, we ask stakeholders to identify links between indicators that could be false positives or point out missing links that could be false negatives. This methodological approach is an example of how to combine PPI with other assessment tools.

The network-estimation method that we have employed the most is known as Sparse Gaussian Bayesian Networks, which was developed by Aragam et al. (2019) (and is known as sparsebn). This procedure has the distinct advantages of working well with high-dimensional datasets, even if they have short series, and producing adjacency matrices that try to minimise the number of links that may be false positives (hence the "sparse" term in the name). Without becoming too technical, it is important to mention that this method assumes that each observation in a time series comes from an independent random draw from a normal distribution. To increase the

[18] In Ospina-Forero et al. (2022), we conduct a thorough review of quantitative methods to be deployed when estimating networks of interdependencies between development indicators.

https://doi.org/10.1017/9781009022910.008 Published online by Cambridge University Press

chances of complying with such an assumption, one can transform the time series of an indicator into a vector of first differences (the difference between an observed value in period t and $t - 1$). Thus, a data-prepossessing step consists of transforming the normalised time series into vectors of first differences.

Another critical assumption of sparsebn that the reader should be aware of is that it can only estimate what is known as directed acyclic graphs. This assumption means that, in an estimated network, it is not possible to find a structure that would generate cycles. While this may be important in some contexts, it is not a crucial feature in PPI as the network is just an exogenous component generating spillover effects; that is to say, it is not the leading causal mechanism. Thus, we choose to leave aside network cycles and, instead, gain the ability to use more robust network estimates. Again, this is not a feature of PPI but a personal modelling choice. Another user could easily opt for a different network estimation method that they consider more suitable for their particular context and data.

We find two key benefits of using this method. First, we do not need long time series, an unfeasible requirement in multidimensional sustainable development. This issue is very important because one can produce country-specific networks, so no cross-national pooling is required. Arguably, one can capture much of the context of a nation by specifying how its different development dimensions are interconnected. Thus, by estimating country-specific networks, we allow PPI to preserve context specificity. Second, because this is a Bayesian framework, it can consider prior information about the potential structure of the network. That is, if a stakeholder knows, according to their experience, that certain links should be present in the network and others should be absent, sparsebn can consider this information through 'white' and 'black' lists. Thus, this method also facilitates the incorporation of qualitative insights from expert knowledge.

The reader should also be aware that once we produce an estimate of the interdependency network, we have to remove those

links with extreme weights. Removing outliers helps us to clean the network from cases that are highly likely the result of correlated idiosyncratic shocks. In this procedure, we estimate the distribution of weights in a network; then, we remove those edges with weights below the 2.5 percentile and above the 97.5 percentile. As for the choice of the network estimation method, this step is not a feature of PPI but a choice informed by our experience from conducting several studies in this field. Currently, there is no gold standard for estimating these networks in the context of sustainable development. Thus, PPI is not tied to any particular method, as the estimated network is just one more input and is optional.

Finally, from our experience in this field, we have noticed that the analysis of SDG networks has led to several misinterpretations about what one can directly learn from these objects.[19] We have explained (in Chapter 3) that SDG networks cannot convey causal information and, as such, one should rather use them as stylised facts that inform other modelling frameworks. Contrary to this, we have seen numerous academic and policy studies arriving at conclusions and recommendations based on correlation networks and other similar analyses. Clearly, such interpretations do not hold under the scrutiny of more rigorous impact-evaluation frameworks. To advert this kind of mistake, we avoid emphasising any particular network in the book. It is sufficient to know that, in every chapter, we estimate networks according to the procedure described in this section.[20]

5.8 SUMMARY AND CONCLUSIONS

This chapter has covered several empirical aspects of PPI, with a certain degree of technicality. It provides the more methodologically oriented reader with various statistical elements of our framework. These statistical assessments make the inferences presented

[19] For this reason, we do not present, throughout the book, any analysis or figures related to networks of interdependencies.

[20] The reader can access the data and code for estimating these networks from the companion depository.

https://doi.org/10.1017/9781009022910.008 Published online by Cambridge University Press

throughout the book more reliable. Some of these elements cover the calibration method, assessing goodness of fit, performing statistical tests, validating the model, and characterising its statistical behaviour. With this, we conclude the first part of the book. Then, we proceed to demonstrate the many different insights to be obtained through the PPI framework. We divide the remainder of the book into two more parts, one with various analyses at a global scale and another with more specific and nuanced studies (e.g., country specific, at a subnational level, or focusing on particular topics).