9 The Impact of International Aid

The development literature has studied the role of international aid for more than half a century, starting with the seminal paper of Griffin and Enos (1970). However, the debate around aid effectiveness remains open due to contradictory shreds of evidence in macro and micro studies. On the one hand, cross-country estimations on macro development find a weak and ambiguous relationship between aid flows and development indicators (e.g., Rajan and Subramanian, 2008; Clemens et al., 2012). On the other hand, there are micro studies at the project and sector level showing more conclusive evidence in favour of positive impacts (e.g., Mosley, 1987; Dreher et al., 2008; Mishra and Newhouse, 2009). The lack of beneficial effects at the macro level is puzzling in the face of robust findings on the positive impacts of targeted aid projects. Presumably, this micro–macro paradox results from the complex causal chains linking the aid flow to aggregate development outcomes. This way of thinking is consistent with (Bourguignon and Sundberg, 2007), who suggest that the empirical literature has historically treated aid as a black box.

Complexity and causation are difficult to analyse through purely aggregate data-driven frameworks. For instance, naïve regressions and reduced-form approaches have a few troublesome issues such as unobserved heterogeneity, confounders, and reverse causality.[1] Moreover, due to causal chains linking different aggregation levels (i.e., from micro behaviours to macro indicators), studies that rely solely on macro–macro relations fail to produce a good enough description of the data-generating process. Another defining characteristic of this literature is its focus on economic growth as a leading proxy of development. This approach ignores

[1] See Brückner (2013) for more on the simultaneity bias in the aid–growth relationship.

https://doi.org/10.1017/9781009022910.013 Published online by Cambridge University Press

effects on intermediate outputs or other development objectives such as building up human capital (e.g., public education and health care). Therefore, numerous development scholars and practitioners have stressed the importance of adopting a multidimensional perspective on the study of aid and its impact (e.g., Tezanos, 2018; Kiselakova et al., 2020; Sterling et al., 2020).

This chapter overcomes several limitations in the existing literature on aid effectiveness. First, it establishes an explicit link between foreign aid inflows and development indicators classified in the multidimensional setting of the SDGs. Second, this linkage is not a black box as it takes advantage of PPI's causal model of policymaking describing budget allocations and indicator performance. Third, we explicitly incorporate explicitly salient features of the public governance of aid as part of the data-generating process, namely, fungibility and an imperfect (or even absent) rule of law.[2] Fourth, we consider the complex pattern of interactions across development dimensions through their network of conditional dependencies. Hence, the resulting response functions to aid changes are not necessarily linear, and their shape is context specific (non-linear impacts have previously been argued and documented in the literature). In other words, indicator changes triggered by aid flows may respond not only to the magnitude of such flows but also to many other factors, such as interdependencies between indicators, the quality of the rule of law, the amount of government spending (different from aid), and the country's initial level of development.

Our simulation strategy in this chapter consists of creating counterfactuals in which we remove aid flows. Given that

[2] The fungibility of aid inflows occurs when the financial assistance intended for a specific project ends up being used by the recipient government as a substitute for previously planned government expenditure on a similar policy issue (Devarajan and Swaroop, 2000). Thus, fungibility is a form of diversion of resources, with the difference that it is not necessarily an embezzlement by a government official, but a re-purposing of the funds. Such detour does not preclude a personal gain by the bureaucrat, as they may be motivated to gain political status by spending on 'more visible' policy issues.

government expenditure, aside from aid flows, present different types of volatility across countries and indicators, we can estimate aid impact and assess its statistical significance at the indicator or country levels during the first decade of the twenty-first century. Performing such disaggregate estimations allows us to deal with heterogeneity-related concerns and embed our empirical analysis into specific structural contexts. Moreover, we produce a validation exercise comparing our results with econometric evidence found in a well-known sector-level study (access and sanitation of water) using a subset of our data. In brief, we provide a detailed picture of the impact of aid on sustainable development that no previous work has documented.[3]

9.1 STUDIES ON AID EFFECTIVENESS

Much of the literature on aid effectiveness originates in mainstream macroeconomics and focuses on the cross-country estimation of aid impact on macro-development outcomes. Because of this, a single methodological approach became dominant in this field: *country-level regression analysis*. For the most part, regression-based studies tend to find a weak and ambiguous relationship between aid flows and the performance of the chosen development indicators (e.g., see Rajan and Subramanian (2008) for both a cross-sectional and panel analysis, and Clemens et al. (2012) for a thorough assessment of different studies). The lack of a clear positive effect at the macro level is puzzling, given the robust evidence on the beneficial impact of targeted aid projects (Mosley, 1987). This micro–macro paradox is mainly the result of complex (vertical) causal chains linking aid flows to aggregate development outcomes.

If a causal chain takes place across different aggregation levels (i.e., from micro behaviours to macro indicators), studies that rely only on macro–macro relations fail to represent the data-generating

[3] This chapter is based on our work in Guerrero et al. (2023). We are grateful to our co-author, Daniele Guariso, for his contributions and insights in this paper, which expanded our knowledge of the aid-effectiveness literature.

https://doi.org/10.1017/9781009022910.013 Published online by Cambridge University Press

process properly. For the most part, empirical research of this type tries to overcome endogeneity issues by using instrumental-variable techniques, an approach questioned by Deaton (2010), among others. Other methods, like naïve regressions and reduced-form approaches, suffer from problems related to unobserved heterogeneity (Papanek, 1973), misspecification (Burnside and Dollar, 2000), and reverse causation (Brückner, 2013).

Another defining characteristic of this literature is its focus on economic growth as a leading proxy of development. This view ignores effects on intermediate outputs (e.g., infrastructure and technology) or other development objectives such as poverty eradication education, health care, and environmental issues. In other words, most of these studies assume a unidimensional view of development. In contrast, numerous development scholars (from different subfields) and practitioners argue in favour of adopting a multidimensional perspective.[4]

Concerning the main independent variable in regressions, most studies operationalise aid with macro measures such as the ratio of net official development assistance (ODA) to government expenditure (or ODA-to-GDP). The reason behind employing the ODA-to-expenditure is to overcome the issue of foreign aid's fungibility. However, recent studies have shifted their focus to analyses at the sector level, showing more conclusive results (Dreher et al., 2008; Mishra and Newhouse, 2009). For instance, Gopalan and Rajan (2016) use a large panel of countries and find a positive association between aid disbursement and access to water supply and sanitation facilities. Nonetheless, there are some studies showing no effect or ambiguous results of sector-specific aid on development outcomes.[5]

[4] See, for instance, (Anand and Sen, 1997; Hicks, 1997; Sen, 1999; Anand and Sen, 2000; Alkire, 2002; Herrero et al., 2010; Alkire and Foster, 2011; Tezanos and Sumner, 2013; Tezanos, 2018; Bourguignon and Chakravarty, 2019; Kiselakova et al., 2020; Sterling et al., 2020).

[5] See, for instance, Wilson (2011) and Williamson (2008) for the case of health development assistance, Christensen et al. (2011) for educational foreign aid, Bain et al. (2013), and Wayland (2017) for water, sanitation, and hygiene.

An issue with focusing on targeted assistance in a single policy area is the potential neglect of spillover effects to other development dimensions. This is the main finding of Kotsadam et al. (2018) on the role of ODA in reducing infant mortality in Nigeria. This paper combines georeferenced aid data with several iterations of the Demographic and Health Survey. Thus, it provides one of the few empirical studies on aid effectiveness at the subnational level. Their results suggest that the positive effect of international assistance on a specific development outcome (e.g., infant mortality) might come from aid projects that are not directly related. An obvious limitation of this approach is the difficulty of testing indirect effects in a regression framework since one would need to account for numerous interaction terms. This scheme is unfeasible when considering many dimensions because of the coarse-grained nature of development indicators (typically annual observations and short-time series).

From a cross-country perspective, a study by Arndt et al. (2015) assesses the impact of aid on multiple development outcomes. They find a positive long-run effect on income growth, structural change, social indicators, and poverty reduction. These results stress the importance of evaluating aid effectiveness not only on economic growth but also on other development dimensions. In the same study, the authors create a directed acyclic graph model that explains the causal relationship between aid, intermediate, and aggregate outcomes. This specification provides a sound theoretical basis for using instrumental variables in empirical analyses. However, their estimates do not take into account the potential interactions and spillovers across multiple intermediate outcomes and development dimensions. In addition, their approach focuses on capturing the average static effect of aggregate aid, and it does not fully account for the heterogeneity of the impact across countries and sectors. It also imposes a linear relationship between variables while, as shown by Gopalan and Rajan (2016), the impact of foreign aid can have significant non-linear effects on development outcomes.

9.2 DATA

9.2.1 *Countries and Indicators*

The core dataset used in this chapter consists of a sub-sample of the SDR data that contains only aid-recipient countries and covers the 2000–2013 period. We establish the criteria to define this sub-sample through the characteristics of a complementary dataset on international aid flows that we explain ahead in Section 9.2.3. The new sample consists of 146 aid-recipient countries. Since most countries belonging to the West group are removed from this sub-sample, we rearrange the countries into the following groups: *Africa*, *Eastern Europe and Central Asia*, *East and South Asia*, *Latin America (LAC)*, *MENA*, and *OECD*. Figure 9.1 shows the map corresponding to the new grouping scheme.

In Table 9.1, we present a summary of the different levels of development experienced by each country group in the aid-recipients sample between 2000 and 2013. As expected, the worst overall performance corresponds to the average African country. However, the performance ranking of SDGs varies across groups. For instance, SDG

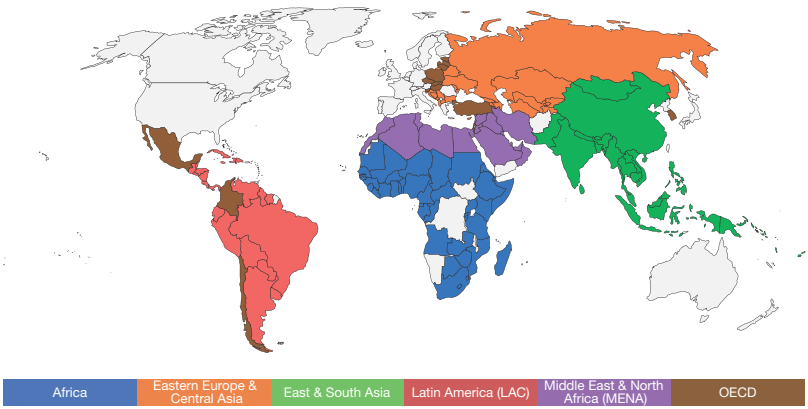

FIGURE 9.1 Countries recipients of SDG-classified aid between 2000 and 2013.

Notes: To clarify the presentation of our results, we classify the countries into six groups. However, we produce all our estimates at the country level.

Sources: Authors' grouping.

https://doi.org/10.1017/9781009022910.013 Published online by Cambridge University Press

Table 9.1 *Average indicator level by SDG and country group (2000–2013)*

SDG	Africa	E. Europe and C. Asia	East and South Asia	LAC	MENA	OECD
1	46.08	87.76	74.86	86.13	94.97	86.48
2	55.07	66.59	57.56	63.81	62.47	68.00
3	63.19	86.60	79.95	83.76	85.66	85.06
4	52.49	80.84	71.30	73.50	74.94	75.34
5	48.63	52.61	47.08	52.82	35.43	56.54
6	62.60	93.76	78.96	88.62	90.46	88.70
7	41.12	91.11	63.44	85.94	96.37	67.32
8	79.82	76.60	82.21	78.32	78.13	64.08
9	2.27	10.33	7.07	7.37	9.76	26.67
10	–	–	–	–	–	64.30
11	56.52	75.33	66.40	75.57	64.64	75.69
13	99.10	95.95	97.32	96.68	90.39	94.35
14	66.68	56.66	59.29	60.00	62.29	64.60
15	62.34	56.37	50.89	55.03	54.32	61.73
16	72.64	80.12	76.19	77.40	75.97	82.66
17	11.71	16.38	13.15	15.60	15.48	23.13
All	54.68	68.47	61.71	66.70	66.09	67.79

Notes: The table reports the indicators' average level within the same SDG and country group in percentage. In SDG 10, all groups but the OECD lack data. The SDR dataset lacks indicators for SDG 12 for the period under study.
Sources: Authors' calculations with data from the 2021 Sustainable Development Report.

13 has the worst performance in MENA and the best in Africa. Likewise, SDG 16 presents the most lagged performance in Africa, while the most advanced is in the OECD. These numbers provide a general, but not surprising, picture of the state of the Global South during the first decade of the twenty-first century. Next, let us explain the data covering aid flows and show additional statistics.

9.2.2 Government Expenditure

Countries with a high ratio of aid flow to government expenditure are more dependent on international aid, and more susceptible to its

https://doi.org/10.1017/9781009022910.013 Published online by Cambridge University Press

withdrawal. Thus, it is important to account for government expenditure in addition to aid flows when measuring funding impacts on development. We use the expenditure data presented in Section 3.1.3 for each country in the sample. Importantly, these data do not include aid flows, so we rule out the possibility of double-counting resources when combining both sources of information.

Notice that government expenditure data classified into the SDGs are not available; something that we have to do in order to combine it with the aid flows data in our simulations exercises. We assume that public funds are distributed, at the level of SDGs, in a similar way as international aid. In other words, we compute the distribution of all the aid data across the SDGs (pooling all countries due to the sparseness of aid data) and use it as a guiding distribution for the expenditure data.[6] While this assumes that all countries share the same expenditure pattern across SDGs, we consider this a better approach than adopting a uniform distribution.

Because public expenditure data classified into the SDGs exist only for a few countries, there is not a cross-country dataset that we could use for this study. Therefore, we exploit the fact that aid data partially reflect the structural demands for expenditure across SDGs. In any case, the bottom row of Table 9.2 presents the total government spending (in real per capita USD) observed during the 2000 to 2013 period across groups, which we show using the countries' average. As expected, the largest average of government expenditure corresponds to OECD countries, while the smallest corresponds to countries in the African group.

[6] Unfortunately, SDG budget tagging is not a common practice across governments around the world. However, there are several countries (especially in Latin America) that do link budgetary programs to specific SDGs. Hence, in order to validate our approach, we rely on labelled fiscal data from Uruguay and Mexico. We verify if the distribution across SDGs of public spending correlates with the distribution of the aid received. A positive and significant relationship between the two would provide evidence favouring the use of aid flows as a guiding distribution for budget allocation. In both cases, the correlation coefficients are positive and statistically significant at the 99% level, supporting our approach.

Table 9.2 *Total per capita aid flows and government expenditure per country group (2000–2013)*

SDG	Africa	E. Europe and C. Asia	East and South Asia	LAC	MENA	OECD
1	10.02	8.99	50.79	10.05	3.33	0.48
2	71.41	38.42	75.35	79.61	13.21	6.93
3	130.38	30.41	184.46	53.42	21.42	0.11
4	108.80	55.16	317.59	48.02	64.78	11.69
5	3.92	3.88	5.21	3.04	2.04	0.05
6	64.55	44.32	58.52	27.11	42.92	8.73
7	38.31	56.52	177.29	17.01	54.70	4.17
8	27.88	49.22	73.97	21.23	26.27	4.92
9	87.03	66.28	281.04	38.83	71.93	20.38
10	4.60	4.98	14.31	4.75	2.01	0.18
11	97.53	67.84	297.45	55.94	46.41	23.46
13	0.66	0.28	1.09	2.87	0.39	0.01
14	2.76	0.80	4.13	1.48	0.48	0.16
15	10.32	3.26	16.83	11.54	1.60	1.03
16	173.49	197.74	527.16	145.94	114.52	14.56
17	141.79	36.18	33.24	55.48	86.79	9.50
All aid	974.10	664.58	2118.79	576.49	552.80	106.36
Gov. exp.	4,552.13	14,167.87	15,556.58	12,607.99	29,378.77	104,56.63

Notes: The table reports the total aid flows channelled to each SDG of an average country in a specific group during 2000 to 2013 (in real per capita USD). Thus, for example, one should read the first entry in the Africa column in the following way: on average, a country in Africa received 10.02 USD per capita in aid flows for SDG 1 between 2000 and 2013. The bottom row denotes the total government expenditure (in real per capita USD) per country group.
Sources: Authors' calculations with data from the 2021 Sustainable Development Report.

9.2.3 *Aid Flows*

For the aid data, we employ a comprehensive dataset built in the Aid Data research laboratory at William and Mary University (Tierney et al., 2011). This team classified a large sample of aid projects that accounted for more than one million aid flows. The resources for projects granted between 2000 and 2013 were assigned to the multiple dimensions of development. Using this dataset, Sethi et al. (2017) provide a characterisation of aid in the context of the SDGs. Therefore, for the simulations in this chapter, we combine this information with government expenditure as the two main factors that drive the dynamics of development indicators. In Table 9.2, we present the aid distribution across SDGs and country groups.

From the second last row, one can observe that the average country in East and South Asia receives the largest donation of aid in per capita terms, whereas the smallest donation during the same period goes to the OECD with 2,118.7 and 106.3 USD in real units per capita, respectively. Notice also the high variability of aid flows for a particular SDG across country groups. For instance, while the average country in LAC receives 79.6 USD per capita in real terms for SDG 2, the average country in the OECD gets only 6.9.

Figure 9.2 shows the aid-to-government-expenditure ratio for all countries in each group. With data disaggregated at the country level, one can see that aid as a proportion of government expenditure is significant in many countries. For many of them, this proportion is greater than 25%, especially in countries located in Africa and East and South Asia. The reader should be aware that the calculations in this figure are qualitatively similar to the World Bank indicator on ODA as a percentage of government expenditure (see Guerrero et al., 2023). Hence, the main difference between the ODA indicator and the data used in Figure 9.2 is that Sethi et al. (2017) gather more sources of aid. Furthermore, our sample contains more countries than the World Bank's ODA dataset.

https://doi.org/10.1017/9781009022910.013 Published online by Cambridge University Press

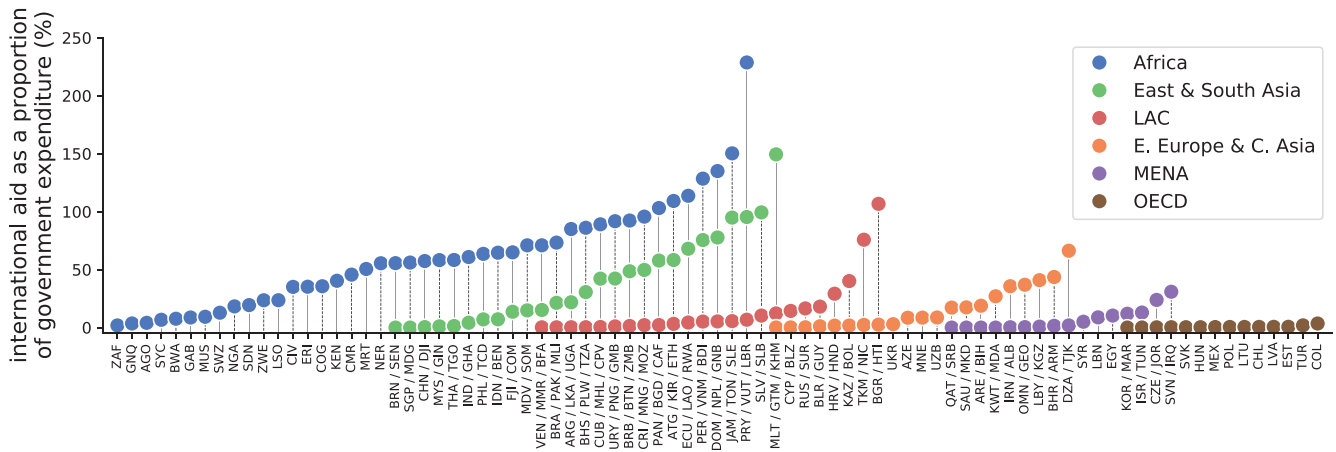

FIGURE 9.2 Total aid received as a fraction of government expenditure during 2000–2013.

Notes: The coloured dots indicate the aid-to-expenditure ratio in the vertical axis (in percentages), while the horizontal axis indicates the country codes, where the closest to the line corresponds to the highest dot.

Sources: Authors' calculations with data from Sethi et al. (2017).

9.3 SIMULATION STRATEGY

We assess the impact of aid flows by performing a counterfactual simulation that considers the removal of aid flows. This procedure allows us to generate data of a world where countries rely exclusively on the expenditure of their governments. Hence, our estimates capture the impact of aid flows in the context of each country while considering, in parallel, the spending capability of their central authorities. Let us summarise our simulation strategy in the following steps:

1. Calibrate the model for each country using government expenditure and aid together.
2. Perform simulations for the sample period using the fitted parameters.
3. Run counterfactual simulations with the fitted parameters but without aid flows.
4. Construct a statistic that quantifies how much development during the sample period is attributable to aid.
5. Assess the statistical significance of such a statistic.

Items 1 to 3 are straightforward as they only involve the removal of the aid component from the disbursement schedule to produce new indicator dynamics. The expected behaviour of this scenario is that the indicators will worsen since the probability of success should fall due to smaller contributions by the agents. Nevertheless, it should also be possible to observe occasional improvements due to unusual spillover structures or adaptations by the agents. For items 4 and 5, we need to construct a bespoke impact metric and a customised statistical test. These analytical tools should be flexible enough to be implemented at the level of indicators and countries while enabling their construction at higher levels of aggregation. Such flexibility is critical to assess the impact of aid across different policy dimensions and to make comparisons of the resulting effects at an aggregate level (e.g., SDGs or groups). This feature of our study is important since the aid-effectiveness literature has been unable to produce such disaggregated and comprehensive results. Furthermore, we validate our estimates by showing that they are consistent with the existing

https://doi.org/10.1017/9781009022910.013 Published online by Cambridge University Press

evidence at a more aggregate level. Next, we explain in detail how we implement our proposed simulation strategy.

9.3.1 Expenditure, Aid, and Counterfactuals

Recall Equation 4.7, in which we specify the disbursement schedule of the central authority. In that case, the behavioural component of the central authority determines, endogenously, the amount of funding allocated to each indicator. However, here, we consider a level of aggregation in the disbursement schedule in which the allocation of resources is conditioned by the data on SDG-level expenditure and aid flows. In Chapter 4, we explained that PPI is flexible to rely on the behavioural component of the government with various degrees of granularity depending on the quality of the data available. Such data quality is reflected in the disaggregation level of government expenditure and its linkage to development indicators. This application is the first in which we introduce more disaggregated expenditure data (not just on total government spending) and show how to condition the allocation of resources based on this information. Hence, at the level of SDGs, we can define an *aggregate disbursement schedule* described by a matrix

$$\mathbb{B}' = \begin{bmatrix} B_{1,1} + A_{1,1} & B_{1,2} + A_{1,2} & \dots & B_{1,T} + A_{1,T} \\ B_{2,1} + A_{2,1} & B_{2,2} + A_{2,2} & \dots & B_{2,T} + A_{2,T} \\ \vdots & \vdots & \dots & \vdots \\ B_{17,1} + A_{17,1} & B_{17,2} + A_{17,2} & \dots & B_{17,T} + A_{17,T} \end{bmatrix}, \tag{9.1}$$

where $B_{i,t}$ and $A_{i,t}$ correspond to government expenditure and aid flows, respectively. Matrix $\mathbb{B}'$ is fully exogenous, so the central authority decides how to allocate resources within each SDG. In Chapter 4 we explain that accounting for the inter-temporal variation of budgetary data involves certain technical complications that lie beyond the scope of this book.[7] Thus, in the same way, as we do with data on total expenditure, here we collapse the columns of $\mathbb{B}'$ by taking the inter-temporal mean of each row (SDG).

[7] See Guerrero et al. (2023) for a full treatment accounting for the inter-temporal variation of aid data.

https://doi.org/10.1017/9781009022910.013 Published online by Cambridge University Press

Hence, we reduce the aggregate disbursement schedule to a vector $\bar{\mathbb{B}} = \frac{1}{T}\sum_t^T (B_{1,t} + A_{1,t}), \ldots, \frac{1}{T}\sum_t^T (B_{17,t} + A_{17,t})$.[8]

We calibrate the model under this reduced specification of the aggregate disbursement schedule. For the counterfactual analysis, we simulate the dynamics of the indicators under a setting that excludes all the terms $A_{i,t}$ from $\mathbb{B}'$, so aid flows are absent, and the government can only redistribute its resources. After running baseline and counterfactual simulations, we generate data to estimate the impact of aid flows at the level of each indicator in the calibrated country. Next, we explain how we obtain such estimates.

9.3.2 *Impact Metric*

In our counterfactual exercise, we are considering an intervention that generates three possible outcomes: (1) continuous changes in development indicators that can close or widen the development gap during the period under analysis; (2) parallel changes in different endogenous variables as a consequence of processes working simultaneously (i.e., *ceteris paribus* assumptions do not apply); and (3) non-linear dynamics generated by interdependencies and spillover effects. Therefore, we need to design a statistical metric that can reflect the previous outcomes using a set of point estimates of the indicators across time. In this section, we build a metric that calculates the difference in simulated trajectories with and without interventions (i.e., baseline and counterfactual). Then, we consider different scenarios in the indicators' dynamic to explain intuitively how this metric works.

First, we produce a sample of Monte Carlo simulations of size M, with each yielding a specific level $I_{i,t}$ for indicator i in period t. The average level of indicator i in period t is the point estimate across all simulation runs $\bar{I}_{i,t} = \sum_m I_{i,t,m}/M$. Second, we calculate the historical performance of an indicator in period t through the difference between its estimated level and the lowest estimated

[8] Due to this temporal aggregation, the results presented in this chapter may differ slightly from those shown in Guerrero et al. (2023). However, for the most part, and qualitatively speaking, they are consistent.

https://doi.org/10.1017/9781009022910.013 Published online by Cambridge University Press

level across all periods. Note that performance could describe either continuous improvements (an indicator with a positive trend) or a worsening trough time (a negative trend). Third, we define the total historical performance as the sum of these differences: $\sum_{t=1}^{T}[\bar{I}_{i,t} - \min_t(\bar{I}_{i,1}, \ldots, \bar{I}_{i,t}, \ldots, \bar{I}_{i,T})]$. This quantity depicts the area under the expected trajectory curve; let us call these objects the baseline area and the baseline curve, respectively. As shown below, the proposed impact metric quantifies how much of the baseline area is due to aid flows.

Fourth, with the same calibrated parameters, we produce a new round of simulations with the counterfactual input and obtain the expected trajectory (point estimates) for each indicator. Fifth, we calculate the area between the baseline curve and the counterfactual curve according to $\sum_{t=1}^{T}[\bar{I}_{i,t} - \bar{I}'_{i,t}]$, where $\bar{I}'_{i,t}$ is a point estimate of the counterfactual curve. Let us call this residual the counterfactual area. The impact metric D_i of indicator i quantifies the fraction that the counterfactual area represents of the baseline area. Formally, it is described by

$$D_i = \frac{\sum_t^T [\bar{I}_{i,t} - \bar{I}'_{i,t}]}{\sum_t^T [\bar{I}_{i,t} - \min_t(\bar{I}_{i,1}, \ldots, \bar{I}_{i,T})]}, \tag{9.2}$$

which we usually present as a percentage (so we multiply it by 100 on our charts).

Most of the time, the counterfactual curve will be below the baseline curve because the removal of aid induces lower success probabilities $\gamma_{i,t}$.[9] This impact metric is equally valid across

[9] In rare cases, D_i could be negative. If it were, it could be due to (1) a particular configuration of the spillover network or (2) the behavioural component of the model. In case (1), it is possible that, by removing aid, one decreases the negative externalities received by an indicator and allows it to perform better, even under a lower budget. In case (2), a lower budget could shift the incentives of the agents in such a way that they become more proficient. For example, if the positive spillovers received by an indicator disappear, the agent's performance becomes more transparent to the central authority, so it can better identify and penalise inefficiencies. Through these penalties, the agent learns the convenience of becoming more proficient, and the indicator performs better under the counterfactual. As we show in our results, cases where $D_i < 0$ are unusual.

https://doi.org/10.1017/9781009022910.013 Published online by Cambridge University Press

indicators exhibiting positive or negative trends since the removal of aid affects the development performance in both scenarios. Another advantage of this metric is that we can build D_i using different aggregation levels, either across countries or SDGs. For example, suppose that we want to estimate the aggregate impact of aid for a group of countries $k = 1, \ldots, K$ and across a set of indicators $i = 1, \ldots, N$. Then, we only need to generalise Equation 9.2 with the following expression:

$$D = \frac{\sum_{k,i,t}^{K,N,T} [\bar{I}_{i,t,k} - \bar{I}'_{i,t,k}]}{\sum_{k,i,t}^{K,N,T} [\bar{I}_{i,t,k} - \min_t(\bar{I}_{i,1,k}, \ldots, \bar{I}_{i,T,k})]}. \tag{9.3}$$

Figure 9.3 shows the intuition behind the impact metric utilising three scenarios. In the first scenario (Figure 9.3a), we consider an indicator with a positive trend, in which the blue area quantifies the total historical performance (reproduced with our simulations). Under the intervention, the counterfactual curve is lower and slower

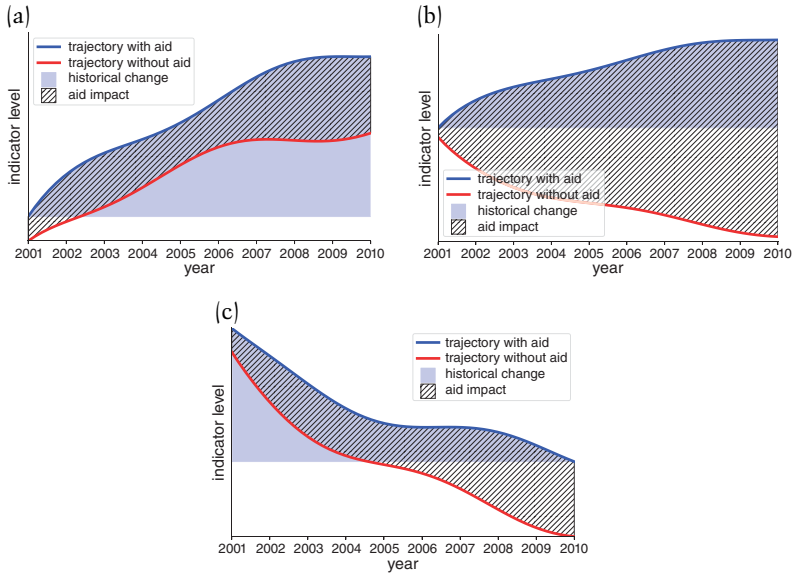

FIGURE 9.3 Hypothetical scenarios to illustrate the workings of the impact metric. (a) Scenario 1, (b) scenario 2, and (c) scenario 3.

https://doi.org/10.1017/9781009022910.013 Published online by Cambridge University Press

but retains the positive trend. Notice that the striped area describes the expected performance loss when removing aid. Therefore, our metric is the difference between these two areas, normalised by the total historical change in the baseline simulations. Hence, one can interpret this ratio as the potential performance loss in the absence of aid with respect to the overall improvements in the indicator. In this scenario, the metric is less than 1 because the striped area is smaller than the blue area. However, the metric is not bounded by 1, as shown in the next scenario.

Figure 9.3b illustrates the case of an indicator with a positive trend that becomes negative when, in the counterfactual simulations, we remove the aid flows. Under the counterfactual, the potential loss in performance is larger than the historical change; hence, the impact metric is greater than 1. Finally, in scenario 3 (Figure 9.3c), we illustrate the case of an indicator with a negative trend, whose performance worsens with the lack of aid. The metric has the same interpretation – a performance loss – because the counterfactual curve is, in general, lower than the baseline curve. Accordingly, the impact metric is agnostic to whether the historical performance exhibits a positive or negative trend.[10]

The proposed impact metric offers several advantages over more traditional approaches such as average effects from linear-regression coefficients. Let us discuss the main advantages. First, it accounts for non-linear dynamics by taking into account the complete trajectory of the indicators. Second, it can estimate the impact of counterfactuals with nuanced policy changes[11] by analysing the entire dynamic of the outcome variables. Third, it is easy to construct the metric at different levels of aggregation (and their respective hypothesis tests) by working with areas under trajectories. Fourth, it accounts for context-specific factors included in the model that may lead an indicator to grow more in one country than another. We do so by defining the impact in terms

[10] Likewise, the impact metric measures a performance loss when the sign is positive (almost always) or a win when the sign is negative (very rarely).

[11] Like the coordinated removal of aid across multiple SDGs.

of the historical change of country-specific indicators (i.e., simulated instead of observed). This procedure makes it feasible to compare interventions (e.g., aid effectiveness) across countries. Fifth, by using point estimates derived from Monte Carlo simulations, the metric ameliorates the potential influence of idiosyncratic shocks that may affect the empirical time series of the indicator – corresponding to a single realisation of the world.[12]

To the extent of our knowledge, no similar metric exists in the aid-effectiveness literature. Due to its flexibility and comprehensiveness, we also use this impact metric in the remaining chapters of this book. Next, we explain how to construct a suitable statistical test to assess the impact estimates obtained through this approach.

9.3.3 Statistical Significance

In Chapter 5, we discussed the importance of designing bespoke null models and statistical tests when working with computational models.[13] Formulating a suitable null model, in turn, involves thinking carefully about sources of uncertainty and data-generating mechanisms. In the case of this study, a key source of uncertainty comes from random changes in government expenditure resulting from idiosyncratic factors; something prevalent, especially, in unstable economies.[14] Such uncertainty makes the assessment of aid impact difficult because a random fluctuation in government expenditure could be as impactful as aid flows. In such a case, it would be unclear what is the actual effectiveness of the latter. In other words, we can say that the impact of aid is significant only when one can

[12] Because idiosyncratic shocks are unrelated to the intervention under consideration, the empirical time series should not be employed to build impact metrics.

[13] While statistical testing has not been a concern of the book's results so far, in this chapter, we would like to elaborate on it as it is an important component in establishing the effectiveness of aid.

[14] These random fluctuations could originate from contingent government changes, policy priority shifts, exogenous shocks like crises, or political instability, for example.

https://doi.org/10.1017/9781009022910.013 Published online by Cambridge University Press

distinguish, statistically, its influence from the idiosyncratic fluctuations of public spending.

To formalise this idea into a hypothesis test, we need to statistically model the fluctuations of government expenditure. Later on, we perform simulations under random realisations of the budget. Formally speaking, the impact metric D can be considered statistically significant for a particular country and indicator only if the impact is differentiated systematically from one produced using random budgets in the absence of aid. If we assume that a country experiences expenditure stability and sizeable aid flows (as a fraction of government spending), then D would be relatively large when removing aid flows from the model input. However, in a country experiencing a volatile government expenditure, it is more likely to generate equally large impact metrics in the absence of aid purely through the randomisation of the budget.

There are two steps in preparing the budgetary data for the null model: (1) generating a random realisation of the budget $\bar{\mathbb{B}}$ and (2) removing the aid component. With these data, we re-calibrate the model parameters. In this new parameterisation, we conceive a world in which aid does not exist, and we use this representation to establish control-treatment types of scenarios. That is to say, aid can be thought of as the treatment, and the null hypothesis should be the equivalent of a control group.[15] Then, we compute the impact metric *as if this randomised expenditure data were the counterfactuals* using the corresponding Monte Carlo simulations.

In other words, we preserve the original baseline and replace the counterfactual with the null model. We repeat this procedure for several randomised budgets and collect a large sample of impact metrics. By doing this, we are propagating the uncertainty of budgetary fluctuations (idiosyncratic to each country) to the impact metric. Therefore, we can build the distribution of the impact metrics obtained under

[15] This implies that randomised government expenditure works as a placebo in the sense that the dynamics of the indicators are explained only by government expenditure.

https://doi.org/10.1017/9781009022910.013 Published online by Cambridge University Press

the null (i.e., the random space of D under the null). The statistical test determines whether, with a given degree of confidence, the estimated impact metric would be an expected outcome under the null distribution.

For individual countries, we model the dynamics of government spending statistically. In this fashion, we can generate random realisations and produce ensembles of time series. With this aim in mind, we fit a Gaussian process to the time series of government expenditure of a given country.[16] Then, we generate a null time series of public spending by randomly drawing values from the point-specific estimated distributions of each country.[17] Then, we generate 1,000 expenditure realisations for each country to produce the corresponding distributions of null impact metrics and to establish the significance tests. Assessing statistical significance requires a single-tail test because we are interested in the outcomes with $D > 0$. Consequently, if the estimated D (i.e., whose counterfactual is removing aid) lies beyond the $95th$ percentile of the null distribution, we say that the impact of aid in such country/SDG/indicator is significant at the 95% confidence level.

9.4 RESULTS

As the reader will soon realise, the possibility of specifying the impact metric at different levels of aggregation is very convenient for detecting patterns across countries, SDGs, and indicators. The metric is aggregated at the country level when we measure the relative progress across all indicators within a country. We aggregate it at

[16] Gaussian processes have become widely popular in the machine-learning literature because of their high accuracy and flexibility in predicting time series. For instance, they have been used for the construction of composite environmental indicators (Becker et al., 2017). One of the virtues of Gaussian processes is that they permit inferring point-specific mean and variances, so they help to provide detailed uncertainty measures.

[17] Under this approach, it is assumed that each empirical observation is a realisation of its corresponding distribution. Therefore, the resulting null series corresponds to an alternative expenditure history that would be expected from the fluctuations implied by the data.

the SDG level when assessing its performance across all indicators within an SDG in a given country group. It is aggregated at the indicator level when we estimate the impact of aid flows for the same development indicator across different countries. In the following visualisations, we show the significance (99%, 95%, or not statistically significant) of each of the calculated metrics. Before discussing further details, we can establish that, overall, aid impact is significant in most countries, SDGs, and indicators. These results are not strictly associated with the amount of aid flows as a fraction of government expenditure.

In Figure 9.4, we rank the countries in terms of impact. The first outcome to notice is that most of them present an impact with at least 95% significance. This result speaks to the national-level relevance of aid flows in roughly two-thirds of the countries in the sample during the period under study. This analysis does not identify the exact reasons behind a lack of significant impact – at the national level – among the remaining aid recipients. However, we can conjecture that it could be a consequence of insufficient aid, the existence of idiosyncratic bottlenecks (bad implementation, wrong incentives, or logistic problems), or the presence of decreasing marginal returns in relatively well-developed policy issues.

Looking at specific countries in Figure 9.4, we detect several features worth of being highlighted. First, aid flows do not exhibit a significant impact in any of the emerging economies integrating the OECD (brown lines). Second, aid is statistically and economically effective in most countries in the Africa (blue lines) and East and South Asia (green lines) groups. Third, there is no strong association between the reception of aid – as a fraction of government spending – and its effectiveness. Fourth, large recipients of aid in relative terms (e.g., Liberia [LBR], Somalia [SOM], Cambodia [KHM], Guinea-Bissau [GNB], and the Central African Republic [CAF]) have a notorious impact metric. However, it is also the case that some countries receiving a low amount of aid experience statistically significant positive impacts (e.g., South Africa [ZAF], India [IND], and Togo

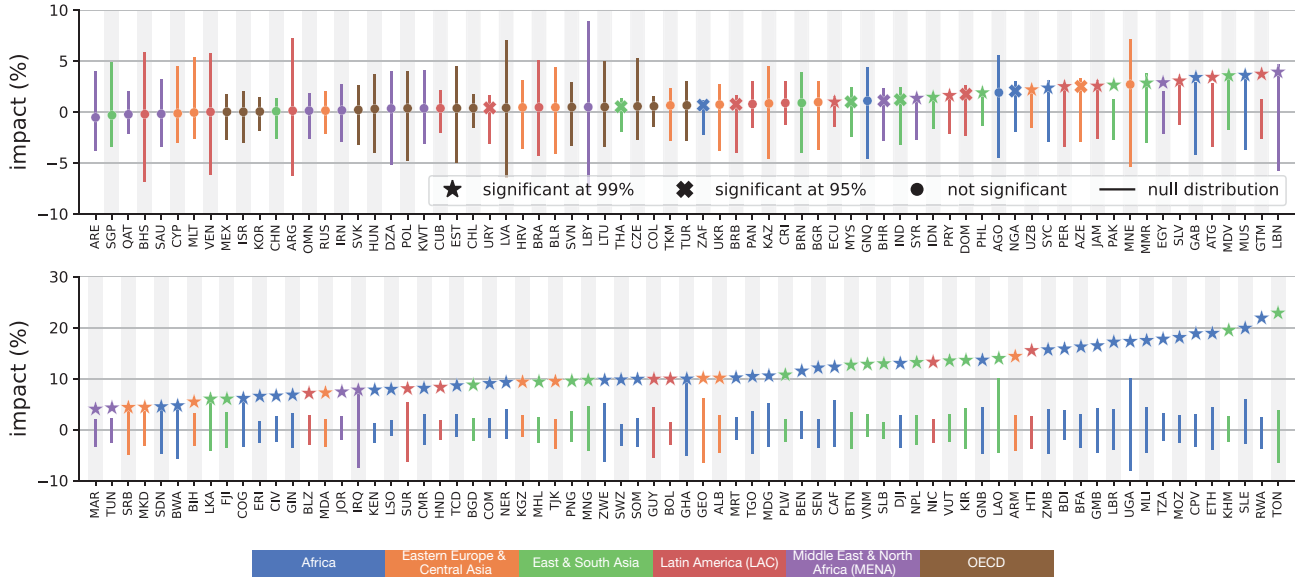

FIGURE 9.4 Country-level impact of international aid.

Notes: The markers (dot, cross, and star) indicate the statistical significance level of the impact metric. The coloured line in each column shows, for the corresponding country, the distribution range of D under the null model.

Sources: Authors' calculations.

[TGO]). Fifth, there are no countries from the Middle East and North Africa (MENA) among the top 50 in terms of aid effectiveness.

Next, let us reconstruct the impact metric to analyse aid effectiveness at the level of SDGs and country groups. In Figure 9.5, we notice that aid is effective across a large number of SDGs and country

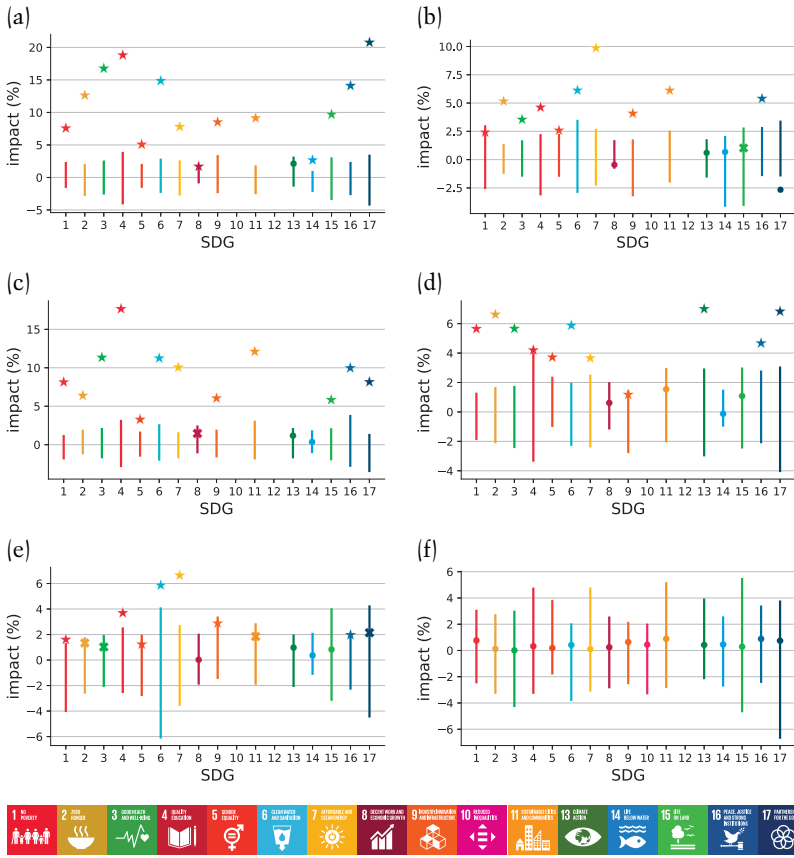

FIGURE 9.5 SDG-level impact of international aid by group. (a) Africa, (b) Eastern Europe and Central Asia, (c) East and South Asia, (d) LAC, (e) MENA, and (f) OECD.

Notes: The markers (dot, cross, and star) indicate the statistical significance level of the impact metric. Star = significant at 99%; cross = significant at 95%; dot = not significant; line = distribution range for D under the null model.

Sources: Authors' calculations.

https://doi.org/10.1017/9781009022910.013 Published online by Cambridge University Press

groups, except for emerging economies in the OECD (Figure 9.5f), where no influence is detected at this level of aggregation. On the contrary, we reject the null hypothesis for all SDGs but 13 (at the 99% level) in countries composing the Africa group. At this level of disaggregation, we can also observe heterogeneous impacts on different development indicators within and between groups. For instance, 'quality of education' (SDG 4) has an impact metric close to 20% in Africa and East and South Asia, but close to 4% in Eastern Europe and Central Asia, LAC, and MENA. Likewise, aid has its strongest impact on 'partnership for the goals' (SDG 17) in Africa, on 'affordable and clean energy' (SDG 7) in Eastern Europe and Central Asia and MENA, on 'quality of education' in East and South Asia, and on 'climate action' (SDG 13) in LAC.

We re-calculate the impact metric one more time, now at the level of each indicator, combining the data from all countries in the sample. In Figure 9.6, one can appreciate the impact that aid exerts on the different indicators. These indicator-level results are closer to those traditionally presented in the aid-effectiveness literature, where average effects come from cross-country pooled regressions. However, there is a critical difference in how both types of estimates are achieved. In regressions, the aggregation happens before the estimation, as these models need to exploit cross-country variation. In PPI, aggregation is an ex post step since we perform estimations for individual countries. Thus our aggregate results do not contain potential biases that may arise from pooled estimates.

Figure 9.6 suggests that, in 52 out of the 74 indicators, aid exerts a positive impact with at least a confidence of 95%. Furthermore, the impact metrics are relatively large (close to or above 10%) for several indicators in 'good health and well-being' (SDG 3), 'quality of education' (SDG 4), 'clean water and sanitation' (SDG 6), and 'partnerships for the goals' (SDG 17). Whereas in most SDGs there are impact disparities among their indicators, aid seems to be effective also on 'zero hunger' (SDG 2), 'affordable and clean energy' (SDG 7), and 'sustainable cities and communities' (SDG 11). In contrast, aid

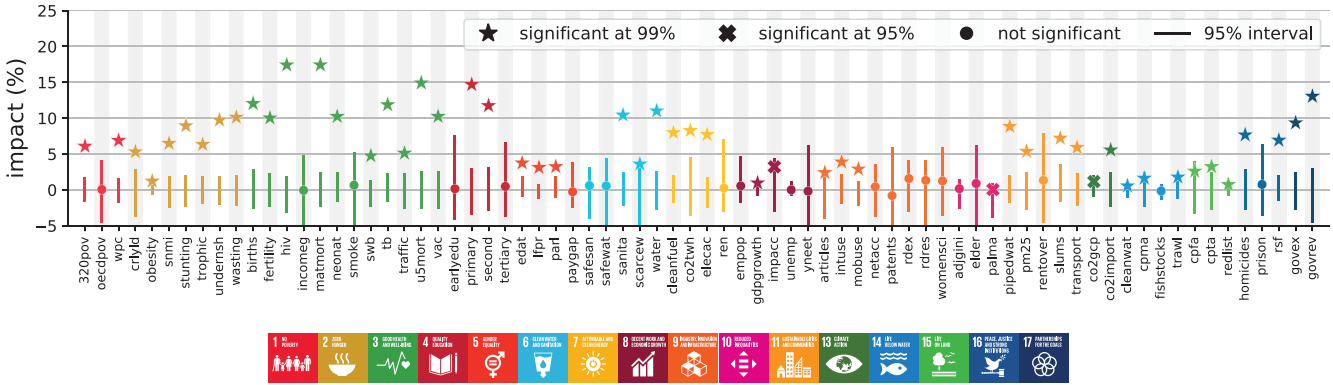

FIGURE 9.6 Indicator-level impact of international aid.

Notes: The markers (dot, cross, and star) indicate the statistical level of significance of the impact metric. The coloured line in each column shows, for the corresponding indicator, the distribution range of D under the null model.

Sources: Authors' calculations.

seems to have a modest influence on the progress of the following indicators: 'decent work and economic growth' (SDG 8); 'industry, innovation, and infrastructure' (SDG 9); 'reduced inequality' (SDG 10); 'life below water' (SDG 14); and 'life on land' (SDG 15).[18]

9.5 BEYOND CONVENTIONAL METHODOLOGIES

In Chapter 2, we argue that, whenever possible, it is recommendable to compare the outcomes of different quantitative methodologies that study the same problem. So far, the problems analysed in this book are relatively understudied, at least from a systemic and quantitative point of view. In contrast, aid effectiveness is a well-established field of inquiry, with plenty of quantitative evidence using similar data to that employed in this chapter. Hence, this is an opportunity to show the reader an exercise comparing previously published results in this domain with those we obtain with PPI.

The purpose of this exercise is twofold. On the one hand, to present a soft validation of our approach by checking if our results are consistent with the findings of a previous study that links aid flows to the performance of indicators using regression analyses. On the other hand, we aim to show that, with PPI, we can produce more disaggregated estimates (i.e., country specific) than those obtained with regression-based findings (i.e., sample specific), and that this has important implications in terms of policy recommendations. The econometric study by Gopalan and Rajan (2016) is an ideal choice for this exercise. First, it is one of the few sector-level studies (water access and sanitation) linking aid with indicators' performance using observational data. Second, its sample period is a sub-period of our dataset. Third, two of our indicators in SDG 6 are equivalent to their main dependent variables.

Gopalan and Rajan (2016) find a significant impact of official development assistance on water and sanitation indicators with a sample of approximately 80 countries (the impact size varies slightly

[18] Even though, in some cases, the associated metric is statistically significant.

with different models). They obtain these results with 99, 95 or 90% significance levels (depending on the specification) by performing panel regressions. The dependent variable is one of two key indicators: *Improved water source (% of the population with access)* and *Improved sanitation facilities (% of the population with access)*.[19] Their main independent variable is gross ODA disbursements destined for water and sanitation, which is a subset of the aid flows in our data.[20] Overall, Gopalan and Rajan's (2016) results are robust across different tests, disaggregations, and specifications (linear and non-linear).

Two of our indicators in SDG 6 can be mapped into their main dependent variables:

- *Sanita*: Population using at least basic sanitation services (percentage)
- *Water*: Population using at least basic drinking water services (percentage)

Each of these indicators has a coverage of 136 countries in our dataset, substantially more than in Gopalan and Rajan (2016). We are interested in (1) verifying if our cross-country estimates for *sanita* and *water* are consistent with Gopalan and Rajan (2016) and (2) learning whether aid impacts countries differently in these policy issues. The first part validates our approach since results consistent with Gopalan and Rajan (2016) would suggest that we can arrive at similar conclusions through our methodology when aggregating the impact metric. The second part demonstrates the benefit of our framework since, for example, policy recommendations based on an average impact across countries could be misleading when a large number of nations in the sample do not experience significant improvements.

[19] They also produce disaggregated versions of these indicators by separating rural and urban populations.

[20] Note that, like us, Gopalan and Rajan (2016) also consider governance indicators related to the rule of law and the quality of monitoring. The difference is that Gopalan and Rajan (2016) consider these indicators according to the dependence account of causation, whereas we adopt the production account to incorporate the role of public governance into the model.

To verify the validity of our results, we can look at Figure 9.6 and see that both indicators (in SDG 6) show positive and significant impacts. Thus, our methodology yields results that are consistent with Gopalan and Rajan (2016). Next, in Figures 9.7a and 9.7b, we show the impact metric of *sanita* and *water*, respectively, estimated for each country in the sample. First, we can see that many countries exhibit a significant impact metric. Second, most of these countries are in the Africa group. Third, the impact ranking is not the same for both indicators.

The disaggregated estimates presented in Figure 9.7 indicate a high degree of heterogeneity in aid effectiveness across countries. As suggested in Table 9.3, 60% (67.6%) of the countries show some level of significance in the impact metric of 'sanita' ('water'). This result is one of the main reasons why Gopalan and Rajan (2016) find a positive and significant impact of aid in water access and sanitation facilities when producing aggregate estimates. Notice that the number of countries exhibiting an impact varies across groups and confidence levels. Most groups present aid effectiveness in the majority of countries, except for OECD. Thus, in very general terms, one could say that Gopalan and Rajan's 2016 results are valid because they apply to most countries in the sample. However, even if the proportion of countries showing an impact within a group were above 50%, it is still unwise to produce group-wide recommendations based on evidence from average effects. For instance, many countries in groups like MENA and Eastern Europe and Central Asia clearly show a small impact at the individual level.

9.6 SUMMARY AND CONCLUSIONS

One of the longest-standing debates in the development literature is aid effectiveness. This debate remains unsolved due to the contrasting evidence offered by macro studies, which tend to exhibit ambiguous results regarding aid impacts on development, and micro studies showing positive results in aid-funded projects. We argue, in this chapter, that one of the main limitations of macro analyses is their

https://doi.org/10.1017/9781009022910.013 Published online by Cambridge University Press

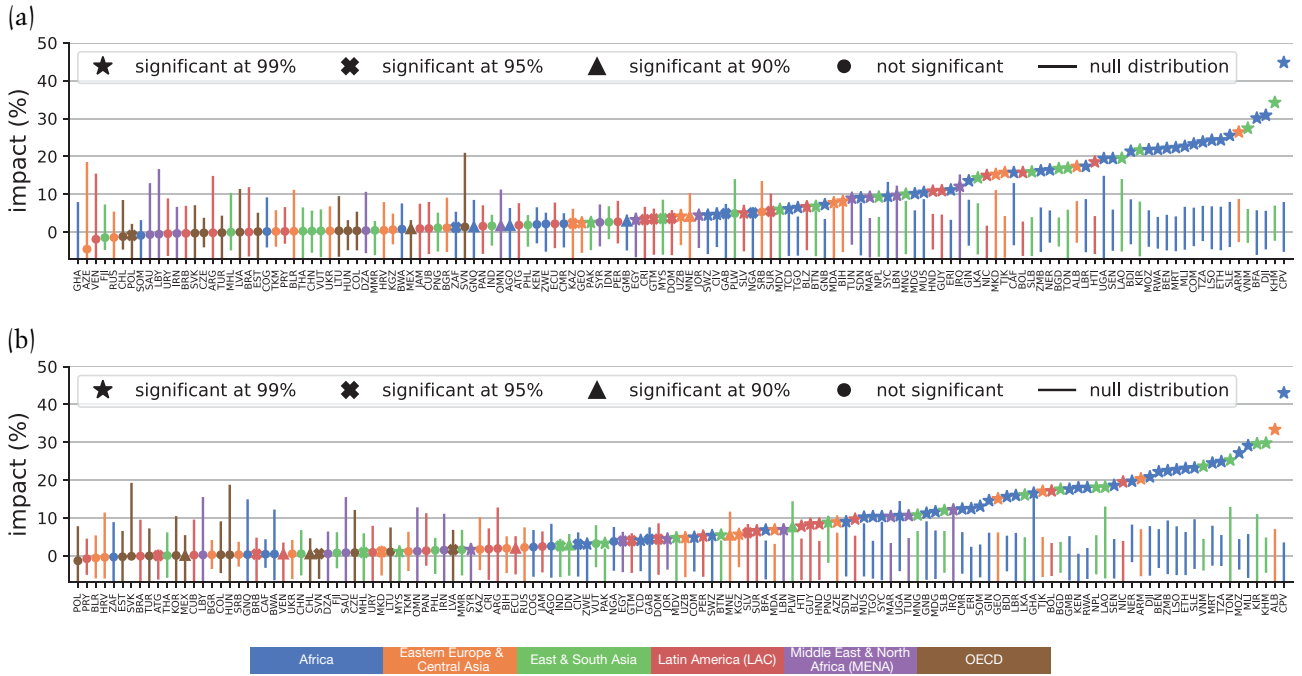

FIGURE 9.7 Disaggregated impact of aid related to access to basic sanitation services. (a) *Sanita*: Population using at least basic sanitation services (percentage) and (b) *Water*: Population using at least basic drinking water services (percentage).

Notes: The markers (dot, cross, and star) indicate the statistical significance level of the impact metric. The coloured line in each row shows, for the corresponding country, the distribution range of D under the null hypothesis. The countries have been sorted from lowest to highest impact metric. Following Gopalan and Rajan (2016), we report significance at 99, 95 and 90%.

Sources: Authors' calculations.

Table 9.3 *Percentage of countries with statistically significant aid impact in water and sanitation*

Group	N	Sanita	Water
Africa	44	70.45 (77.27) [84.09]	79.55 (84.09) [84.09]
E. Europe and C. Asia	19	42.11 (47.37) [52.63]	42.11 (47.37) [52.63]
East and South Asia	25	52.00 (56.00) [56.00]	60.00 (68.00) [76.00]
LAC	23	30.43 (47.83) [47.83]	34.78 (56.52) [65.22]
MENA	12	50.00 (50.00) [58.33]	50.00 (58.33) [58.33]
OECD	13	0.00 (8.33) [16.67]	0.00 (15.38) [30.77]
Total	136	48.15 (55.56) [60.00]	52.94 (62.50) [67.65]

Notes: Group: country group. N: number of countries in the group. Sanita: percentage of countries in the group with statistical significance in the indicator of sanitation services. Water: percentage of countries in the group with statistical significance in the indicator of drinking water services. Confidence levers are 99%, (95%), and [90%].
Sources: Authors' calculations.

inability to produce statistical inferences of country-specific impacts by pooling cross-national datasets. Therefore, we study aid effectiveness using PPI and devise a simulation strategy to circumvent this limitation. Through such a strategy, we produce indicator- and country-specific estimates that, later, we can aggregate as needed.

Through PPI, it is possible to establish a causal link between aid flows and the evolution of development indicators with relatively low levels of granularity. In contrast, neither econometric analyses nor machine learning methods can handle this task with currently available data; at least not in a setting with multiple dimensions and interdependencies. The former approach has problems tackling multiple interactions, endogenous relationships, and datasets with relatively short time series and many indicators. The latter cannot establish causal mechanisms to connect public spending with the indicators' performance. Consequently, the main contribution of this study is the application of a computational method that produces granular estimates of the impact of international aid across a broad range of recipient countries. In the data-generating process, our model

captures several features of the political economy of each country and the presence of spillovers between indicators.

Once we calibrate the model for each country using a large dataset of aid flows and development indicators, we infer that approximately two-thirds of the recipient countries are impacted favourably in their development during the 2000 to 2013 period. The countries benefiting the most belong, principally, to Africa and East and South Asia. We discover that a relatively large share of aid, as a proportion of government expenditures, makes this outcome more likely. However, our results also suggest that having a significant ratio is not strictly necessary for achieving progress. Moreover, concerning the nature of the induced development, the evidence from counterfactual simulations indicates that aid exerts a positive impact across several dimensions. Yet this impact is negligible, on average, in some SDGs and groups. These heterogeneous effects vary across SDGs and country groups and are practically nonexistent in OECD recipients. Altogether, this study supports the idea that aid matters for macro development and, in this manner, produces consistent results with study cases and regression analyses at the sector level.

This chapter concludes Part II of the book. It also provides the first application of disaggregated data, the development of a bespoke statistical test, and the combination of different data sources under the same framework. In the next part of the book, we dive deeper into more disaggregated data, levels of government, sectoral specificities, and even model tweaks. Thus, the remaining chapters display the flexibility of PPI to tackle a wider set of problems and provide an overview of the state of the art in the modelling of policy prioritisation and its empirical analysis.