3 Relevant Data and Empirical Challenges

Data and modelling are central elements of this book. Together, they help scientists and analysts to generate quantitative evidence about the effectiveness and impact of public policies. During scientific discovery, the ability to apply new modelling techniques tends to be constrained by the data at hand. This obstacle is not an exception in the study of sustainable development, where the most abundant data consist of aggregate development indicators. Typically, the temporal coverage of indicators is short, and their level of aggregation is coarse grained. These data restrict the nature of quantitative methods a researcher can apply to assess sustainable development from a systemic point of view (i.e., if one wishes to account for multidimensionality and complexity).[1] Thus, one of the challenges of capacity building for development is that governments and other organisations have to invest substantial resources in constructing more and better indicators.

Initially, one of the motivations for creating PPI was precisely providing a tool that can operate with data that policymakers have at their disposal. Accordingly, this book presents empirical applications of PPI using only publicly available data. In this chapter, we introduce the reader to the public datasets that we employ throughout the chapters in Part II, where we provide a worldwide view of the state of sustainable development and how it responds to government expenditure. In light of this global dataset on development indicators, we also describe the most popular analytic tools and their limitations. Finally, we reflect on the main empirical challenges that researchers

[1] Of course, some initiatives use higher-resolution data, for example from satellite imaging. Others resort to detailed information from field experiments. However, many of these applications are highly specific and remain in domain silos.

https://doi.org/10.1017/9781009022910.006 Published online by Cambridge University Press

face when studying sustainable development with these data and motivate the methodological proposal of the book.

3.1 A WORLDWIDE LOOK AT SUSTAINABLE DEVELOPMENT THROUGH DATA

When the United Nations member states launched the SDGs (UN General Assembly, 2015), they also set in motion a strategy for data construction. This strategy took advantage of previous efforts, such as those of the Millennium Development Goals Project, to assemble development indicators with the purpose of evaluating progress towards the 169 targets contained within the 17 SDGs. While such initiatives are commendable, they rely heavily on a political process since member states approve which 'official' indicators to add to the dataset. This process means that the inclusion or exclusion of an indicator may be subjected to unnecessary red tape and the whims of particular government administrations as, sometimes, these actors may perceive that certain data reflect negatively on their performance.

Thus, while the official SDG indicators dataset (United Nations, 2020) is the result of solid statistical procedures, its coverage is not ideal for many of the methodological frameworks discussed in this chapter. For example, many time series have poor temporal coverage or missing values for many countries (especially for lower-income countries). Thus, to overcome the limitations of the official SDG indicators, multiple organisations have created alternative datasets. For example, the World Bank has mapped several of its previously built indicators into the SDGs (World Bank, 2020), and the OECD has compiled a dataset to assess the distance to SDG targets (OECD, 2020). In this book, we employ an alternative dataset built by the Sustainable Development Solutions Network and the Bertelsmann Stiftung for the 2021 Sustainable Development Report (SDR) (Sachs et al., 2021).

3.1.1 SDGs and Indicators

The SDR data have become popular among researchers due to their comprehensive coverage of countries, themes, and years. In addition,

https://doi.org/10.1017/9781009022910.006 Published online by Cambridge University Press

we identify three main reasons why the SDR dataset has gained widespread acceptance. First, while alternative datasets may contain more indicators, they fail to provide consistently long time series (i.e., numerous indicators have only one or two observations). Second, the majority of the data sources of the SDR are recognised international (and intergovernmental) organisations; meanwhile, the others are scientifically sound products such as surveys from statistics bureaus, reputed NGOs, and academic institutions. Third, the construction of these data is less subjected to the discretionary decisions of governments, providing a more reliable source of information for policy evaluation.

When we started writing this book, the most updated SDR dataset corresponded to the 2021 Sustainable Development Report (covering the 2000–2021 period). A caveat of these data is that SDG 12 (responsible consumption and production) contains excessively short time frames that we cannot use in our analysis; hence, we drop these indicators for Part II of the book. The same problem prevails in all other SDG datasets, and it owes to the fact that the policy issues related to SDG 12 (such as waste management) have been quantified recently in a handful of countries. Therefore, we end up with 74 development indicators after preprocessing and cleaning the data.

First, let us introduce the reader to the 17 global goals defined by the 2030 Agenda, which we show in Figure 3.1. The 17 SDGs are the first global attempt to acknowledge the multidimensionality of development and, hence, one of the reasons why we are shifting from a socioeconomic view of development to a sustainability view. Another innovation of the SDGs is the acknowledgement of interconnections between its policy dimensions. It adds to the complexity of development and calls for new analytical frameworks for impact evaluation and prospective analyses. Addressing multidimensionality and complexity is challenging, especially if one tries to formulate a theory explaining micro-level mechanisms leading to macro-level outcomes. For this reason, the book aims to take the initial steps in this research agenda. We do this by introducing an analytical

FIGURE 3.1 The 17 Sustainable Development Goals.

framework that attempts to strike a balance between the use of aggregated data and the amount of complexity accounted for in the data-generating process. In other words, we do not want to fall into the trap of ending with an overly complicated model that remains essentially theoretical.

In Table 3.1, we provide the complete list of the indicators sampled from the SDR dataset, each classified into its SDG, and with a code label that we use to present some of our results in Part II. Moreover, we also include one column with the name *instrumental*. When we define an indicator as such, it implies that governments have at their disposal policy instruments or programmes specifically designed to impact the indicator's performance. While a government programme implementation may be deficient and ineffective, its mere existence is already informative about the availability of policy instruments. Thus, PPI does not assume ex ante that a government programme exists by the mere presence of an indicator.

Indicators not defined as instrumental (i.e., when a country has no relevant programme) are called *collateral*. There are several reasons why we classify an indicator as collateral instead of instrumental. One is the possibility that an indicator is too aggregate to realistically consider that any policy instrument can exert direct

https://doi.org/10.1017/9781009022910.006 Published online by Cambridge University Press

Table 3.1 *Indicators sampled from the Sustainable Development Report*

SDG	Indicator code	Instrumental	Indicator name
1	320pov	yes	Poverty headcount ratio at $3.20/day (%)
1	oecdpov	yes	Poverty rate after taxes and transfers (%)
1	wpc	yes	Poverty headcount ratio at $1.90/day (%)
2	crlyld	yes	Cereal yield (tonnes per hectare of harvested land)
2	obesity	yes	Prevalence of obesity, BMI ≥ 30 (% of adult population)
2	snmi	yes	Sustainable Nitrogen Management Index (best 0–1.41 worst)
2	stunting	yes	Prevalence of stunting in children under 5 years of age (%)
2	trophic	yes	Human Trophic Level (best 2–3 worst)
2	undernsh	yes	Prevalence of undernourishment (%)
2	wasting	yes	Prevalence of wasting in children under 5 years of age (%)
3	births	yes	Births attended by skilled health personnel (%)
3	fertility	yes	Adolescent fertility rate (births per 1,000 females aged 15–19)
3	hiv	yes	New HIV infections (per 1,000 uninfected population)
3	incomeg	yes	Gap in self-reported health status by income (percentage points)
3	matmort	yes	Maternal mortality rate (per 100,000 live births)
3	neonat	yes	Neonatal mortality rate (per 1,000 live births)
3	smoke	yes	Daily smokers (% of population aged 15 and over)
3	swb	yes	Subjective well-being (average ladder score worst 0–10 best)
3	tb	yes	Incidence of tuberculosis (per 100,000 population)
3	traffic	yes	Traffic deaths (per 100,000 population)

Table 3.1 *(cont)*

SDG	Indicator code	Instrumental	Indicator name
3	u5mort	yes	Mortality rate, under 5 years of age (per 1,000 live births)
3	vac	yes	Surviving infants who received 2 WHO-recommended vaccines (%)
4	earlyedu	yes	Participation rate in pre-primary organized learning (% of children aged 4–6)
4	primary	yes	Net primary enrollment rate (%)
4	second	yes	Lower secondary completion rate (%)
4	tertiary	yes	Tertiary educational attainment (% of population aged 25–34)
5	edat	yes	Ratio of female-to-male mean years of education received (%)
5	lfpr	yes	Ratio of female-to-male labor force participation rate (%)
5	parl	yes	Seats held by women in national parliament (%)
5	paygap	yes	Gender wage gap (% of male median wage)
6	safesan	yes	Population using safely managed sanitation services (%)
6	safewat	yes	Population using safely managed water services (%)
6	sanita	yes	Population using at least basic sanitation services (%)
6	scarcew	yes	Scarce water consumption embodied in imports (m^3/capita)
6	water	yes	Population using at least basic drinking water services (%)
7	cleanfuel	yes	Population with access to clean fuels and technology for cooking (%)
7	co2twh	yes	CO^2 emissions from fuel combustion for electricity and heating per total electricity output ($MtCO^2$/TWh)
7	elecac	yes	Population with access to electricity (%)
7	ren	yes	Share of renewable energy in total primary energy supply (%)

8	empop	no	Employment-to-population ratio (%)
8	gdpgrowth	no	GDP annual growth rate
8	impacc	yes	Fatal work-related accidents embodied in imports (per 100,000 population)
8	unemp	no	Unemployment rate (% of total labor force)
8	yneet	yes	Youth not in employment, education or training (NEET) (% of population aged 15 to 29)
9	articles	yes	Scientific and technical journal articles (per 1,000 population)
9	intuse	yes	Population using the internet (%)
9	mobuse	yes	Mobile broadband subscriptions (per 100 population)
9	netacc	yes	Gap in internet access by income (percentage points)
9	patents	yes	Triadic patent families filed (per million population)
9	rdex	yes	Expenditure on research and development (% of GDP)
9	rdres	yes	Researchers (per 1,000 employed population)
9	womensci	yes	Female share of graduates from STEM fields at the tertiary level (%)
10	adjgini	no	Gini coefficient adjusted for top income
10	elder	yes	Elderly poverty rate (% of population aged 66 or over)
10	palma	no	Palma ratio
11	pipedwat	yes	Access to improved water source, piped (% of urban population)
11	pm25	yes	Annual mean concentration of particulate matter of less than 2.5 μm in diameter (PM2.5) ($\mu g/m^3$)
11	rentover	yes	Population with rent overburden (%)
11	slums	yes	Proportion of urban population living in slums (%)

Table 3.1 *(cont)*

SDG	Indicator code	Instrumental	Indicator name
11	transport	yes	Satisfaction with public transport (%)
13	co2gcp	yes	CO_2 emissions from fossil fuel combustion and cement production (tCO2/capita)
13	co2import	yes	CO_2 emissions embodied in imports (tCO_2/capita)
14	cleanwat	yes	Ocean Health Index: Clean Waters score (worst 0–100 best)
14	cpma	yes	Mean area that is protected in marine sites important to biodiversity (%)
14	fishstocks	yes	Fish caught from overexploited or collapsed stocks (% of total catch)
14	trawl	yes	Fish caught by trawling or dredging (%)
15	cpfa	yes	Mean area that is protected in freshwater sites important to biodiversity (%)
15	cpta	yes	Mean area that is protected in terrestrial sites important to biodiversity (%)
15	redlist	yes	Red List Index of species survival (worst 0–1 best)
16	homicides	yes	Homicides (per 100,000 population)
16	prison	yes	Persons held in prison (per 100,000 population)
16	rsf	yes	Press Freedom Index (best 0–100 worst)
17	govex	yes	Government spending on health and education (% of GDP)
17	govrev	yes	Other countries: Government revenue excluding grants (% of GDP)
17	oda	yes	For high-income and all OECD DAC countries: International concessional public finance, including official development assistance (% of GNI)

Notes: The 'Instrumental' column denotes if, in the policy issue associated with the indicator, the government is likely to have programmes designed to directly impact the indicator.
Sources: 2021 Sustainable Development Report. The identification of instrumental indicators is provided by the authors.

impact in a reasonably controlled manner. A typical example of a collateral indicator is GDP, which is an aggregate description of how an economy performs. GDP, in turn, results from various processes taking place within and between several economic sectors. While many governments devise strategies and policies to (indirectly) promote economic growth, the intervention channels are different, work at a more granular level, and tend to be multidimensional.[2]

Another reason for classifying a variable as collateral could be that a policy issue is not considered relevant or existent (as a problem), so there is simply no government programme to address it. One example of this situation is extreme poverty in a highly developed country. Since these countries eradicated this form of poverty decades ago, some governments in this income group do not have policies to lift people out of extreme hardships or even to measure them. Alternatively, a programme may not exist because there is a lack of capacity in the government. So this indicator evolves as a function of non-governmental initiatives such as the private sector. An example is national cybersecurity, a problem prevalent in many developing countries whose states do not have the necessary technical capacity to attend. Here, private companies bear the responsibility of maintaining a level of safety in a country's digital ecosystem.

So far, we have not provided a clear motivation as to why it is important to classify indicators into instrumental and collateral. It will become evident in Chapter 4, where we present our computational model. For now, it is enough to mention that this distinction matters when assessing the effectiveness of government expenditure. That is to say, one cannot evaluate the impact of public expenditure in, say, alleviating poverty if there are no relevant government programmes in place. This point may be obvious for some readers, but analysts frequently omit this distinction in interpretations of empirical studies and policy recommendations.

[2] In some cases, such strategies succeed at generating economic growth but fail in many others. Hence, even with very comprehensive policies to promote economic growth, no government could assert reliable control over an indicator like GDP, which is why we classify it as collateral.

For example, a study may run a linear regression model explaining health outcomes from changes in poverty indicators. Then, interpret the findings in terms of successful anti-poverty policies while, in reality, there may be no relevant programmes in several countries in the sample. This practice is commonly used in the development literature and among consultants working for international organisations and governments. Of course, determining whether an indicator is instrumental or collateral can be highly dependent on the context. In Part II of the book, we take a global view and, in consequence, we classify as collateral those indicators only when we are almost certain (from our own experience) that no government has control through expenditure programmes. Then, as we shift to more focused studies in Part III, we refine this classification with the expert knowledge obtained through interactions with specialists, policymakers, and high-resolution data on expenditure programmes.

3.1.2 Pre-processing Indicators and Descriptive Statistics

Before describing the indicators' performance, the reader should know how we preprocess these data. As seen in Table 3.1, the indicators are diverse, which implies that they come in different units. While this may not be problematic for some analytic methods (including the method presented in this book), the interpretation of results could be confusing. Thus, a common practice adopted by researchers and development consultants is to transform indicator data into a normalised version. A normalisation is just a way to re-scale the units of an indicator. A procedure like this is typically done, first, by defining the best and worst possible levels that the indicator could take. Analysts also refer to these values as technical/theoretical bounds or limits.

In some cases, technical bounds are implied naturally by the indicator itself. For example, a country cannot have more individuals with diabetes than the size of its total population (and neither less than zero). In others, the bounds are not that obvious, so data providers indicate high and low levels depending on their feasibility

or according to some statistical criteria. Fortunately, the SDR dataset provides bounds in terms of optimal and worst levels (other data sources do not), which makes it easy to re-scale the indicators accordingly. Given an indicator i and the provided bounds $I_{i,\min}$ and $I_{i,\max}$, we normalise the observation I_i by using the formula

$$I_{i,\text{norm}} = \frac{I_i - I_{i,\min}}{I_{i,\max} - I_{i,\min}}. \tag{3.1}$$

Under this normalisation, the indicator values are bounded between 0 and 1. Often, we present these values in percentage, meaning that we multiply $I_{i,\text{norm}}$ by 100. This type of normalisation is known as the min-max criterion, and it is quite common in development studies employing different methods (e.g., benchmarking, regression, and machine learning). Once more, this step is not entirely necessary for some quantitative methods, but studies tend to adopt this procedure when comparing indicators that come in different units. In our case, it helps us present many of our results, as the normalised indicators convey information about the country's performance in a specific policy issue in relation to its potential. Accordingly, we employ this normalisation technique throughout the book.

Another preprocessing step that we perform for the indicators is harmonising their direction. There are indicators where a lower value signals a better outcome, such as the previous example of diabetes cases in a country. Thus, a common practice in the analysis of development data is reverting the direction of indicators like this so that higher values suggest better outcomes. This reversion step does not affect our empirical estimates but is motivated to ease the interpretation of results. Having to analyse many indicators simultaneously comes with the burden of keeping track of the direction they have to follow when advancing. To alleviate such a burden, analysts often harmonise directions through this reversion, so one always interprets higher values as better results.[3] In our particular

[3] The companion code of this book provides scripts with every preprocessing step, so the reader may be able to take the data from its raw form to the shape deployed in the analysis presented in this book.

https://doi.org/10.1017/9781009022910.006 Published online by Cambridge University Press

case, we perform this reversion on the normalised indicator i by using the next formula

$$I_{i,\text{rev}} = 1 - I_{i,\text{norm}}. \tag{3.2}$$

Let us continue by describing the performance of the indicators. Figure 3.2 presents the average level of the development indicators grouped by SDG and cluster.[4] For a given group of countries, the bars originate in the centre of the circle and expand outwards. The height of the solid segment of a bar denotes the average level the indicator had during the sample period. In contrast, the height of the translucent segment indicates the average goal set for such an SDG in 2030. It is important to highlight that these goals consist of specific values for each indicator, reflecting the aspirations expressed in the 2030 Agenda. The SDR data provide these values, yet each government may have particular aspirations. Hence, the goals could be different from those declared in international agendas. To provide a global view, we employ the goals from the SDR since obtaining specific numbers conveying governments' real aspirations is impossible for this number of indicators and countries.

The translucent segments in Figure 3.2 are intriguing because they have an interpretation as *development gaps* that countries must close by 2030 if they wish to achieve the SDGs. As we write this book, we are just coming out of the Covid-19 pandemic, and the feasibility of closing these gaps in less than a decade seems highly unlikely.[5] Nevertheless, they are useful metrics to provide information about how difficult it may be for some countries to develop in the coming years.

In general, we can observe that, on average, countries in Africa and Latin America exhibit the widest development gaps, while countries in the West, Eastern Europe, and Central Asia present the most narrow gaps. However, the best performance varies by SDG.

[4] In Table 3.2, we present the list of countries belonging to each cluster.

[5] And we will need to wait a few years to see the effects of the pandemic on a large set of development indicators.

https://doi.org/10.1017/9781009022910.006 Published online by Cambridge University Press

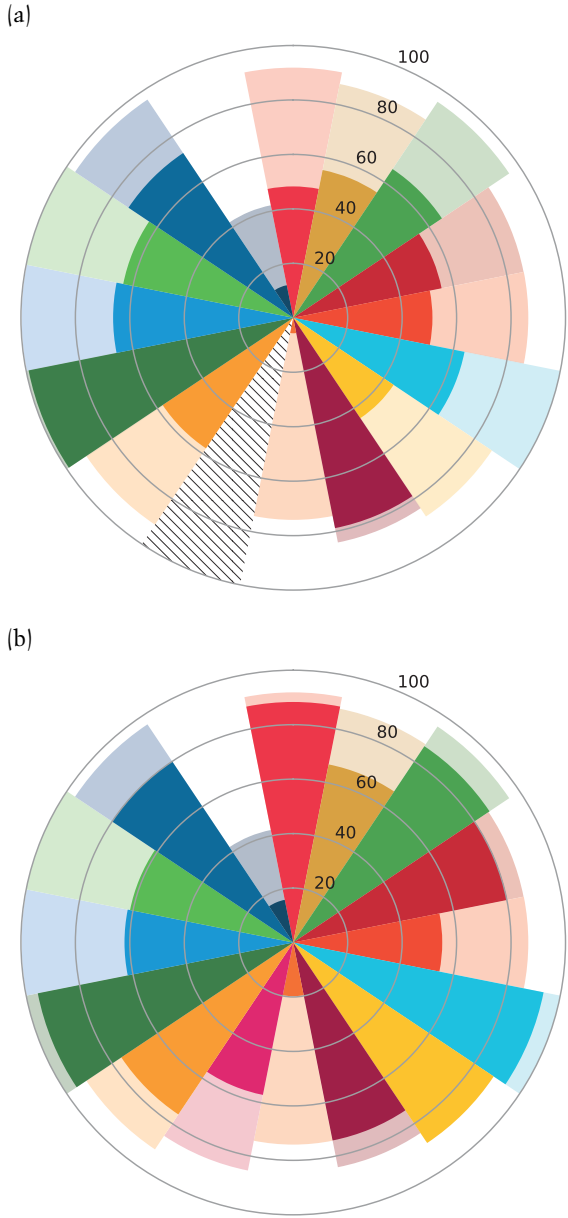


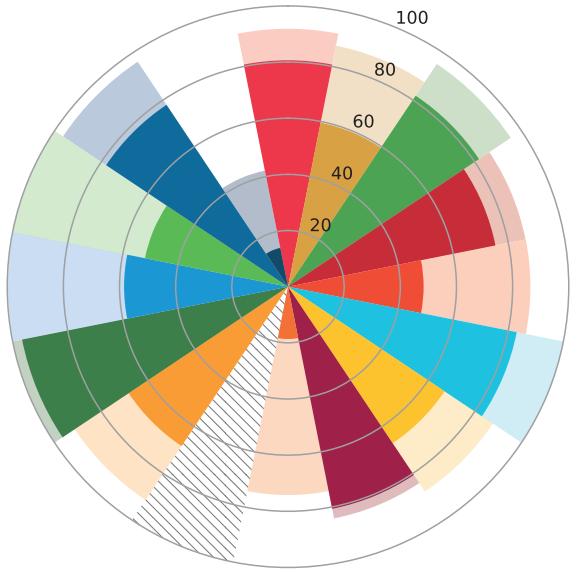
FIGURE 3.2 Average indicator levels and development gaps by country group. (a) Africa, (b) Eastern Europe and Central Asia, (c) East and South Asia, (d) LAC, (e) MENA, and (f) West.

Notes: The height of the solid bars indicates the average level of all indicators in a given SDG and country group. The height of the translucent segments denotes the average development goal that needs to be achieved according to the 2030 Agenda. The striped regions indicate that no data were available for an SDG in any country in the group. The units are percentages.

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

https://doi.org/10.1017/9781009022910.006 Published online by Cambridge University Press

(c)



(d)

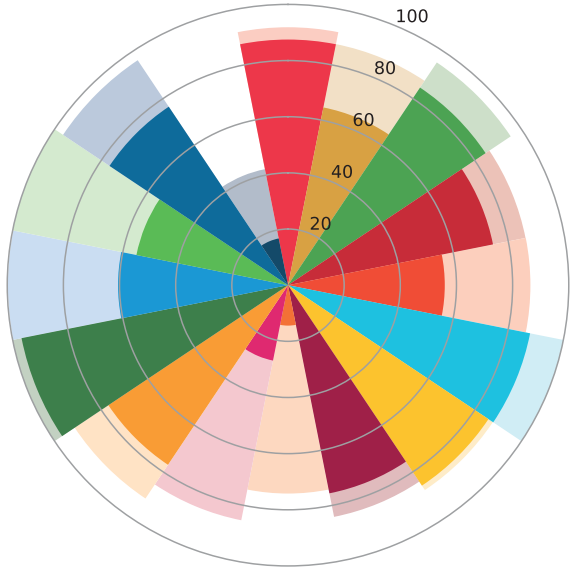
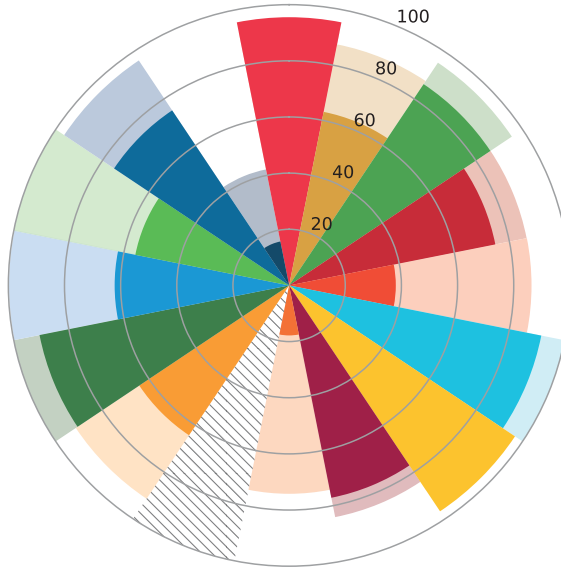


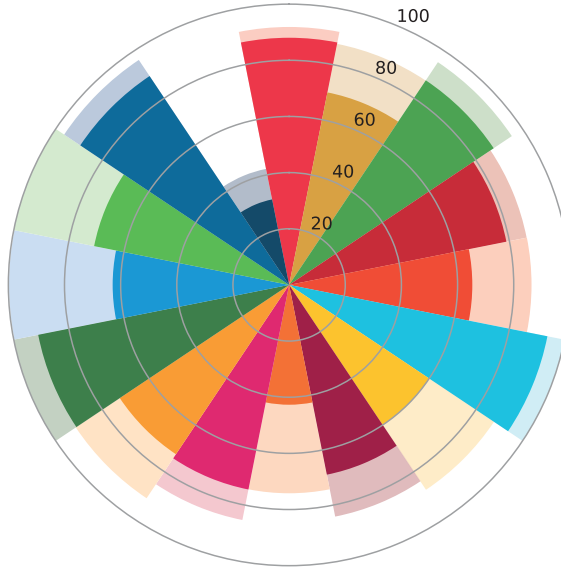
FIGURE 3.2 *(cont)*

https://doi.org/10.1017/9781009022910.006 Published online by Cambridge University Press

(e)



(f)



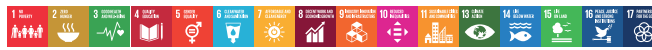


FIGURE 3.2 *(cont)*

https://doi.org/10.1017/9781009022910.006 Published online by Cambridge University Press

For instance, MENA countries, on average, have an outstanding performance in SDGs 1 and 7, the West in SDG 16, while African countries do so in SDG 13. It is also illustrative to observe that in several SDGs, like 15, there are no striking differences between country groups. Likewise, all countries present a large development gap in SDG 9. Notice also that the goals at the 100% level are established only for four SDGs (6, 13, 14, and 15).

3.1.3 Countries and Government Spending

Next, let us look at the country coverage of the SDR data. Throughout several book chapters, we provide global results by aggregating them into country groups or clusters. These clusters are partly defined by the SDR and further refined by us to reflect a combination of geographical and economic features shared, to some degree, by countries within a group. While these groups are a convenient way to communicate the analysis, the reader should be aware that we perform our calibrations and simulations at the level of individual nations. In total, the sample extracted from the SDR dataset covers 171 countries. Figure 3.3 presents the six country clusters that we define and use in multiple chapters. The geographical coverage is quite comprehensive, leaving a few countries, such as North Korea and the Democratic Republic of Congo, out of the sample due to data unavailability.

One of the motivations for developing PPI was to enable the evaluation of the government-spending impact on sustainable development in a multidimensional and complex context. From our experience, we notice that the empirical study of government expenditure is essentially limited to analysing highly specific situations (e.g., a cash-transfer programme in an indigenous community in Amazonian Peru). Although this type of analysis can be insightful to specialists and policymakers dealing with particular cases, their results are hardly generalisable to a multidimensional setting (Deaton, 2010; Stuart et al., 2015). Furthermore, quantitative tools employed in focalised studies fail to assess the relationship between expenditure and development when adopting a broader perspective.

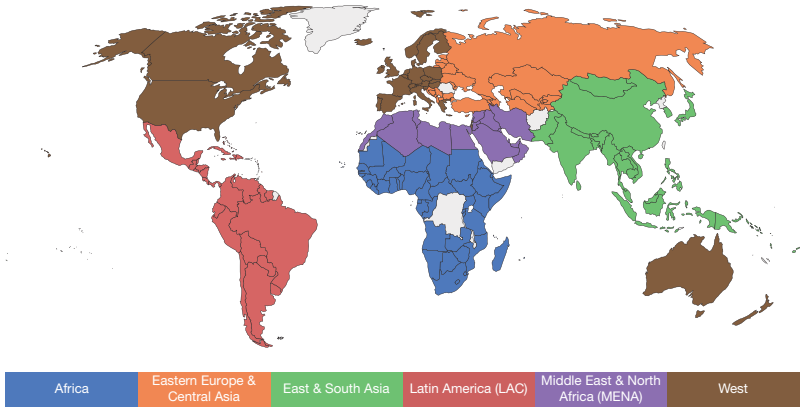


FIGURE 3.3 Countries and regions covered by the SDR dataset.
Notes: The country groups are the following: Blue: *Africa*. Orange: *E. Europe and C. Asia*. Green: *East and South Asia*. Red: *LAC*. Purple: *MENA*. Brown: *West*.
Sources: Authors' grouping with information from the 2021 Sustainable Development Report.

There are various reasons why this is the case. First, political-economy mechanisms intermediate between allocating the budget to a government programme and its successful implementation. For example, bureaucrats' lack of capacity and the diversion of public resources for personal use produce serious inefficiencies. Therefore, many disturbances in aggregate data prevent 'picking up' the right signals when treating the expenditure–development relationship as a black box through purely data-driven approaches. Micro studies help remove some of these disturbances by isolating the analysis of this relationship in specific issues. However, in the context of sustainable development, such isolation is impractical since, by definition, we are dealing with a systemic problem. Thus, we need to resort to a different approach: modelling spillover effects and the political economy explicitly.

A second reason for the absence of empirical studies linking expenditure and development in a multidimensional setting owes to the limited availability of government spending data. Historically, government expenditure has been undisclosed information in most countries. In the best-case scenario, governments would

https://doi.org/10.1017/9781009022910.006 Published online by Cambridge University Press

only reveal total expenditure information or budgets disaggregated into broad tranches. Such coarseness in open spending data prevents properly linking expenditure to indicators, so the best that researchers could do is gain access to specialised datasets on particular policies.

In recent years, open expenditure data have become better and more abundant. International organisations such as the Global Initiative for Fiscal Transparency, the International Budget Partnership, and the Open Government Partnership have championed the standardisation and publication of highly disaggregated datasets worldwide. To the extent of our knowledge, the rigorous use of such data remains scant. The reasons for this are plenty: it may still have poor documentation, be too large to handle by analysts without relevant computational skills, or have a dubious quality (since many governments still do not plan most of their budgets according to objectives), or because it simply exists in categories that are merely administrative and irrelevant from a development-outcomes point of view.

A third reason is that, even with highly disaggregated expenditure data categorised into policy issues, there is a gap between the taxonomies employed by budget specialists and development analysts. Linking expenditure programmes to indicators is a rare endeavour in most countries, as it requires substantial resources and the commitment of various governmental actors (because governments usually have thousands of expenditure programmes). We study this problem in detail in Guariso et al. (2023a), where we investigate the possibility of automated linkages via machine learning.

While machine learning is a promising avenue to generate expenditure–indicator-linked data, it will take some time to get there (even if using recently popularised large language models). As part of this process, several agendas, such as the SDGs, are currently being undertaken to improve the interface between budgetary categories and development indicators. Yet, even with such progress in creating better data, it is clear that connections between expenditure and development outcomes are intricate, obfuscating its usefulness and

rendering purely data-driven methods ineffective for the most part. We demonstrate such ineffectiveness in Guariso et al. (2023b).

In Part III, we present novel datasets with hundreds, or even thousands, of expenditure programmes linked to SDG targets and specific indicators. Nevertheless, one of the motivations for PPI was not to rely exclusively on such unique data but to deliver an analytic framework that is flexible enough to operate under different types of expenditure data and imperfect linkages to indicators.[6] This flexibility is necessary because most countries (developed and developing) still have coarse-grained government expenditures with poor or absent linkages to the SDGs. Therefore, before exploring these new datasets, we use data on government total spending in Part II. This approach provides a baseline for the usage of government budgetary data and gives us an informative picture of how heterogeneous is the expenditure capacity of countries around the world. We present these data in Figure 3.4.

We construct Figure 3.4 using data obtained from the World Bank through the indicator of *general government final consumption expenditure*. The source of this information is the World Bank National Accounts Data and the OECD National Accounts data files.[7] We normalise government expenditure in per capita real USD using the population variable of the SDR dataset. Each panel in Figure 3.4 contains the total government expenditure of each country in the associated cluster. We sort countries from lowest to highest. The black lines indicate the amount of expenditure in 2000, while the height of the bars denotes the level in 2021. Thus, the plot conveys the level of government spending and its overall change in the sample period.

The first thing to notice is that the budget increased in real per capita terms between 2000 and 2021 for almost all countries

[6] Naturally, the more aggregated the expenditure data are, the more limited the inferences to be drawn are.

[7] These data are in real USD, and we change the base year of its deflator to match that from the aid dataset (which is 2011).

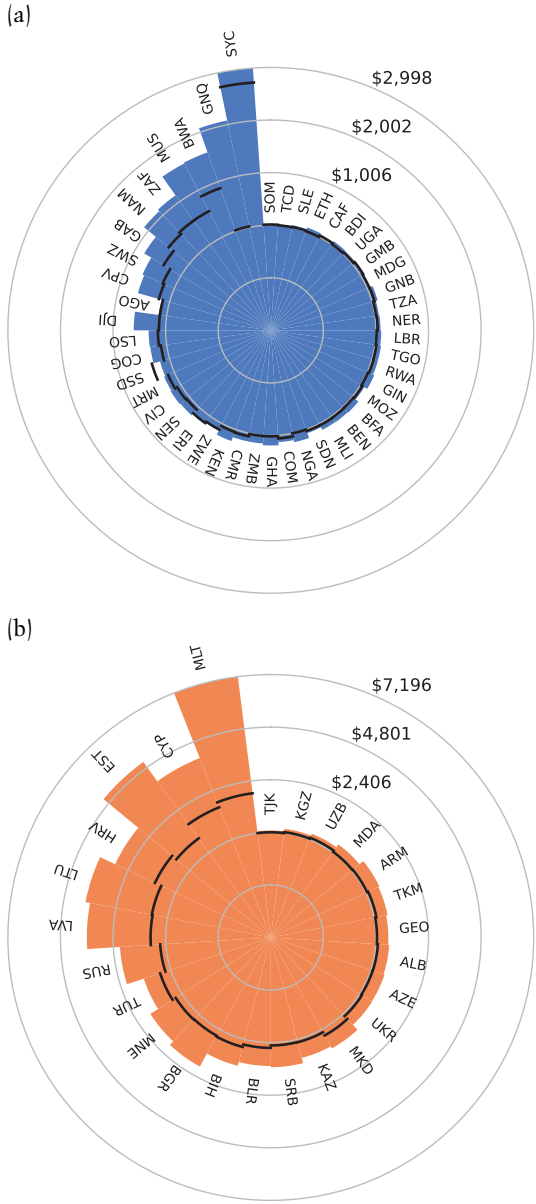


FIGURE 3.4 Government expenditure per capita (real USD). (a) Africa, (b) Eastern Europe and Central Asia, (c) East and South Asia, (d) LAC, (e) MENA, and (f) West.

Notes: Within each group, we sort countries according to their average (annual) government expenditure throughout the 2000–2021 period. The height of each bar indicates the expenditure level of 2021, while the black line refers to public expenditure in 2000. The units are 2010 constant USD per capita.

Sources: Authors' calculations with information from the World Bank's data on government final consumption expenditure.

https://doi.org/10.1017/9781009022910.006 Published online by Cambridge University Press

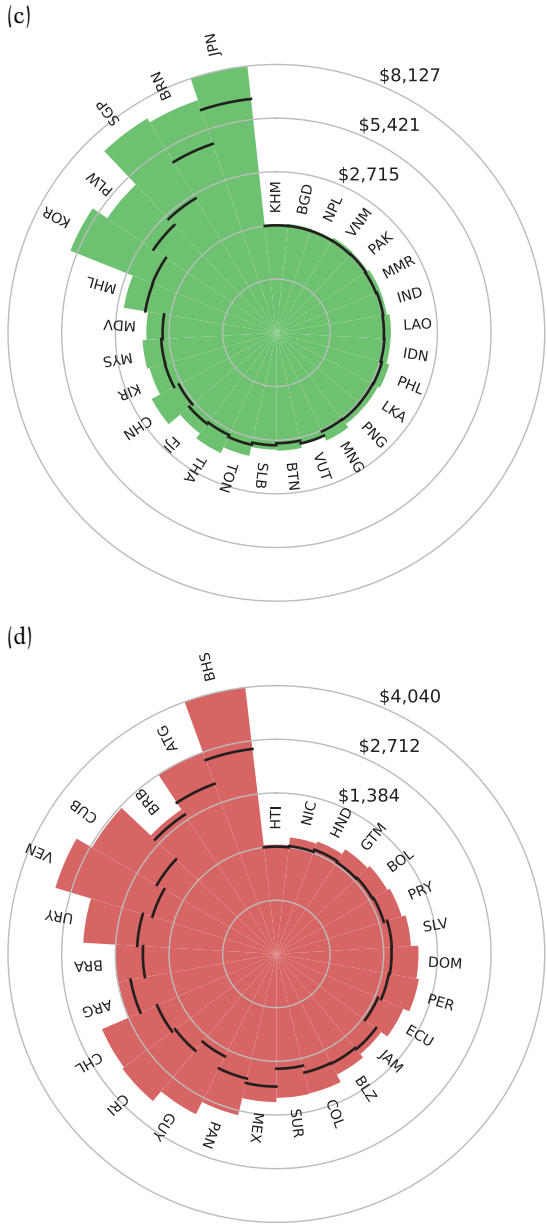
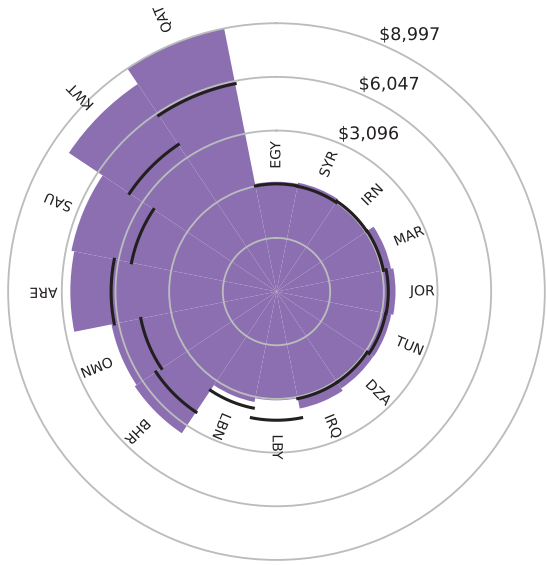


FIGURE 3.4 *(cont)*

https://doi.org/10.1017/9781009022910.006 Published online by Cambridge University Press

(e)



(f)

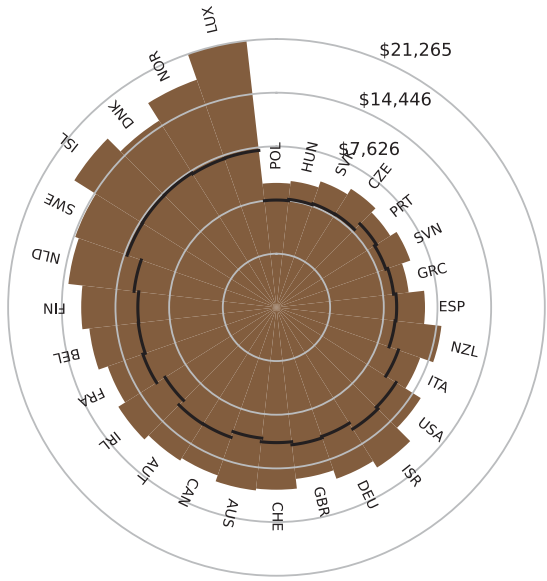


FIGURE 3.4 *(cont)*

https://doi.org/10.1017/9781009022910.006 Published online by Cambridge University Press

(i.e., there are few exceptions), although in diverse ways. Whereas budgetary capacity in several countries in Africa is the weakest in the world, in many countries in the West it is the strongest. The majority of countries in Africa present a yearly budget below the $1,000 per capita line. In contrast, the majority of countries in the West have budgets above the $7,500 line. Countries in Latin America and the Caribbean also have a weak budgetary capacity, with many of them below the $1,500 line. The situation in the other three groups (MENA, Eastern Europe and Central Asia, and East and South Asia) is grossly similar since none of these countries presents an annual budget, in 2021, above the $9,000 threshold. Finally, these panels show that countries in Africa and East and South Asia have more uniformity in their budgetary capacity, where only relatively few break away from the norm.

Table 3.2 presents the complete list of countries, with summary statistics about their average indicators (column: 'Development'), their level of government expenditure (column: 'Expenditure'), and other two indicators on public governance: *Quality of Monitoring* (column: 'Monitoring') and *Quality of the Rule of Law* (column: 'Rule of Law'). The last two indicators come from the Worldwide Governance Indicators dataset (for the same sample period), and we preprocess them in the same fashion as the SDR data. On the one hand, the quality of monitoring reflects the effectiveness of preventative measures against inefficiencies such as corruption in the public sector. On the other hand, the quality of the rule of law captures the effectiveness of the punitive or corrective measures taken by governments when public servants infringe on the law. Both indicators are used in PPI to account for the institutional framework shaping the political economy of the model. We elaborate on this in Chapters 4 and 8. For the time being, these variables offer a glimpse into the heterogeneity of global governance and reflect on their importance when studying the expenditure–development relationship.

The main observation one can infer from the last four columns in Table 3.2 is the presence of a wide heterogeneity among the

Table 3.2 *Countries sampled from the Sustainable Development Report*

Name	Code	Group	Development	Expenditure	Monitoring	Rule of Law
Angola	AGO	Africa	8.91	473.58	24.32	24.91
Albania	ALB	E. Europe and C. Asia	6.95	410.04	36.38	39.50
United Arab Emirates	ARE	MENA	65.95	4,054.25	67.33	63.27
Argentina	ARG	LAC	5.30	1,430.24	43.90	41.65
Armenia	ARM	E. Europe and C. Asia	6.09	359.65	38.72	44.17
Antigua and Barbuda	ATG	LAC	16.19	2,322.49	66.93	66.57
Australia	AUS	West	18.25	8,327.93	87.92	85.65
Austria	AUT	West	21.17	8,538.03	84.35	87.47
Azerbaijan	AZE	E. Europe and C. Asia	12.32	451.53	27.83	34.43
Burundi	BDI	Africa	19.65	49.32	27.85	26.40
Belgium	BEL	West	31.04	9,388.37	79.07	77.90
Benin	BEN	Africa	9.19	103.79	39.29	42.21
Burkina Faso	BFA	Africa	3.97	97.50	46.18	41.07
Bangladesh	BGD	East and South Asia	2.83	55.12	29.05	34.91
Bulgaria	BGR	E. Europe and C. Asia	9.48	1,099.06	46.23	48.81
Bahrain	BHR	MENA	14.03	3,108.37	54.78	59.44
Bahamas	BHS	LAC	12.34	3,505.11	75.82	67.77
Bosnia and Herzegovina	BIH	E. Europe and C. Asia	10.91	941.41	41.93	43.28
Belarus	BLR	E. Europe and C. Asia	8.86	812.08	40.78	31.11
Belize	BLZ	LAC	7.48	673.02	45.93	44.29

Bolivia	BOL	LAC	7.70	342.85	37.24	35.17
Brazil	BRA	LAC	8.27	1,510.25	47.82	47.10
Barbados	BRB	LAC	26.20	2,129.65	78.99	71.89
Brunei Darussalam	BRN	East and South Asia	16.70	6,763.99	61.64	62.71
Bhutan	BTN	East and South Asia	5.79	400.19	71.82	56.99
Botswana	BWA	Africa	10.93	1,243.65	67.76	62.84
Central African Republic	CAF	Africa	2.32	43.41	26.97	22.17
Canada	CAN	West	18.70	8,491.33	88.84	85.37
Switzerland	CHE	West	27.83	8,111.97	91.44	88.04
Chile	CHL	LAC	23.41	1,424.82	77.57	75.58
China	CHN	East and South Asia	3.64	835.59	42.58	42.58
Cote d'Ivoire	CIV	Africa	4.15	162.97	34.98	31.21
Cameroon	CMR	Africa	4.66	140.44	27.02	28.96
Congo	COG	Africa	8.88	340.90	26.94	27.24
Colombia	COL	LAC	6.40	765.53	43.92	40.98
Comoros	COM	Africa	8.08	129.40	32.00	31.16
Cabo Verde	CPV	Africa	7.93	517.53	65.55	62.13
Costa Rica	CRI	LAC	6.79	1,354.94	63.32	61.99
Cuba	CUB	LAC	67.04	1,956.07	54.61	35.70
Cyprus	CYP	E. Europe and C. Asia	9.24	3,323.96	69.69	69.88
Czech Republic	CZE	West	7.43	3,466.74	59.23	69.64
Germany	DEU	West	28.47	7,875.15	87.21	83.72
Djibouti	DJI	Africa	22.55	357.14	36.81	34.04

Table 3.2 *(cont)*

Name	Code	Group	Development	Expenditure	Monitoring	Rule of Law
Denmark	DNK	West	55.54	13,502.68	96.50	88.57
Dominican Republic	DOM	LAC	4.25	564.00	35.81	40.29
Algeria	DZA	MENA	11.45	650.76	36.86	34.43
Ecuador	ECU	LAC	18.63	594.83	35.82	35.01
Egypt	EGY	MENA	6.62	253.12	37.75	46.25
Eritrea	ERI	Africa	9.11	159.80	41.02	29.07
Spain	ESP	West	17.42	5,130.42	71.35	73.58
Estonia	EST	E. Europe and C. Asia	7.45	3,022.45	71.16	71.60
Ethiopia	ETH	Africa	2.92	42.92	38.52	37.26
Finland	FIN	West	28.96	9,898.58	95.71	90.12
Fiji	FJI	East and South Asia	7.18	779.71	53.72	45.71
France	FRA	West	27.62	8,998.56	76.92	78.96
Gabon	GAB	Africa	8.42	1,002.52	33.15	40.91
United Kingdom	GBR	West	36.37	8,051.75	87.12	84.13
Georgia	GEO	E. Europe and C. Asia	5.76	409.12	47.44	44.00
Ghana	GHA	Africa	5.06	131.55	46.77	51.00
Guinea	GIN	Africa	5.87	88.39	30.00	25.37
Gambia	GMB	Africa	5.86	60.27	39.31	43.80
Guinea-Bissau	GNB	Africa	14.50	68.63	24.82	24.17
Equatorial Guinea	GNQ	Africa	8.77	1,593.28	18.82	23.86

Greece	GRC	West	10.80	4,376.31	54.24	62.88
Guatemala	GTM	LAC	3.32	312.78	35.05	30.74
Guyana	GUY	LAC	25.52	1,152.85	41.35	42.02
Honduras	HND	LAC	6.20	284.80	32.90	32.44
Croatia	HRV	E. Europe and C. Asia	31.08	2,396.70	50.58	52.70
Haiti	HTI	LAC	6.12	83.92	24.04	24.89
Hungary	HUN	West	7.61	2,614.68	58.68	65.81
Indonesia	IDN	East and South Asia	4.39	236.37	35.75	39.50
India	IND	East and South Asia	3.54	138.58	42.72	52.55
Ireland	IRL	West	27.32	8,625.44	81.71	82.49
Iran	IRN	MENA	8.29	475.53	37.11	35.00
Iraq	IRQ	MENA	9.74	729.38	22.52	19.71
Iceland	ISL	West	15.24	12,999.77	90.10	85.60
Israel	ISR	West	5.47	7,640.97	69.10	70.59
Italy	ITA	West	13.47	6,258.53	57.22	61.61
Jamaica	JAM	LAC	7.87	660.76	46.71	44.78
Jordan	JOR	MENA	6.80	631.99	53.01	57.08
Japan	JPN	East and South Asia	15.12	7,439.49	77.28	78.05
Kazakhstan	KAZ	E. Europe and C. Asia	4.72	790.80	32.09	35.77
Kenya	KEN	Africa	23.05	147.31	30.47	35.80
Kyrgyz Republic	KGZ	E. Europe and C. Asia	26.00	159.02	28.33	31.72
Cambodia	KHM	East and South Asia	4.04	47.61	26.99	29.61
Kiribati	KIR	East and South Asia	16.14	919.36	50.14	59.31
Korea	KOR	East and South Asia	7.20	3,546.69	60.02	70.26

Table 3.2 *(cont)*

Name	Code	Group	Development	Expenditure	Monitoring	Rule of Law
Kuwait	KWT	MENA	21.70	6,632.68	55.17	59.60
Lao PDR	LAO	East and South Asia	4.02	185.96	29.49	32.14
Lebanon	LBN	MENA	5.61	846.45	33.74	39.64
Liberia	LBR	Africa	5.90	76.95	30.19	26.04
Libya	LBY	MENA	12.71	743.58	25.75	26.00
Sri Lanka	LKA	East and South Asia	4.01	263.33	44.78	52.00
Lesotho	LSO	Africa	10.64	351.96	50.37	48.32
Lithuania	LTU	E. Europe and C. Asia	6.06	2,273.72	58.38	65.45
Luxembourg	LUX	West	62.29	16,242.24	89.88	86.56
Latvia	LVA	E. Europe and C. Asia	4.84	2,252.72	55.51	64.00
Morocco	MAR	MENA	7.43	496.23	45.37	49.41
Moldova	MDA	E. Europe and C. Asia	7.93	267.25	36.17	43.72
Madagascar	MDG	Africa	2.81	68.31	38.60	39.18
Maldives	MDV	East and South Asia	6.76	1,064.09	39.77	45.68
Mexico	MEX	LAC	5.07	1,024.14	40.45	40.97
Marshall Islands	MHL	East and South Asia	7.34	1,737.67	43.23	50.15
North Macedonia	MKD	E. Europe and C. Asia	8.52	730.85	42.50	44.96
Mali	MLI	Africa	4.10	105.03	36.07	41.72
Malta	MLT	E. Europe and C. Asia	13.40	4,201.79	64.98	76.25
Myanmar	MMR	East and South Asia	3.48	128.43	25.65	23.53

Montenegro	MNE	E. Europe and C. Asia	65.91	1,212.46	48.66	48.43
Mongolia	MNG	East and South Asia	7.92	347.47	42.12	48.12
Mozambique	MOZ	Africa	5.68	95.39	38.33	36.09
Mauritania	MRT	Africa	8.00	194.69	37.63	37.54
Mauritius	MUS	Africa	17.24	1,146.51	57.15	69.17
Malaysia	MYS	East and South Asia	5.61	1,040.67	54.83	60.38
Namibia	NAM	Africa	9.36	1,043.09	57.59	55.33
Niger	NER	Africa	13.71	74.52	35.00	39.31
Nigeria	NGA	Africa	1.68	127.35	26.85	28.90
Nicaragua	NIC	LAC	4.38	211.77	34.22	35.70
Netherlands	NLD	West	48.43	11,326.98	90.82	86.19
Norway	NOR	West	64.13	16,006.50	92.53	89.29
Nepal	NPL	East and South Asia	4.70	57.07	36.29	40.30
New Zealand	NZL	West	17.50	6,189.67	95.26	87.98
Oman	OMN	MENA	20.98	3,346.75	58.82	60.47
Pakistan	PAK	East and South Asia	3.81	107.48	31.27	35.52
Panama	PAN	LAC	23.92	1,085.92	43.16	49.04
Peru	PER	LAC	5.50	580.14	42.85	39.68
Philippines	PHL	East and South Asia	4.08	253.66	39.10	43.54
Palau	PLW	East and South Asia	38.88	3,753.03	40.49	63.81
Papua New Guinea	PNG	East and South Asia	5.43	271.83	31.89	34.78
Poland	POL	West	5.95	2,087.11	61.47	63.51
Portugal	PRT	West	16.07	3,828.91	70.96	73.72
Paraguay	PRY	LAC	4.41	420.13	28.84	35.08

Table 3.2 *(cont)*

Name	Code	Group	Development	Expenditure	Monitoring	Rule of Law
Qatar	QAT	MENA	15.20	8,916.97	65.28	63.19
Russian Federation	RUS	E. Europe and C. Asia	7.85	1,623.21	30.97	34.52
Rwanda	RWA	Africa	18.35	82.44	50.63	40.70
Saudi Arabia	SAU	MENA	27.13	4,209.30	49.87	52.83
Sudan	SDN	Africa	4.32	105.14	24.23	24.17
Senegal	SEN	Africa	15.54	162.89	46.93	47.95
Singapore	SGP	East and South Asia	12.09	4,595.90	93.42	82.64
Solomon Islands	SLB	East and South Asia	8.23	412.51	45.71	43.19
Sierra Leone	SLE	Africa	3.47	42.69	34.13	31.51
El Salvador	SLV	LAC	6.15	488.57	40.08	37.96
Somalia	SOM	Africa	23.06	25.88	17.79	6.17
Serbia	SRB	E. Europe and C. Asia	9.68	811.53	38.90	39.61
South Sudan	SSD	Africa	57.86	253.93	33.16	27.03
Suriname	SUR	LAC	6.59	802.20	49.08	48.90
Slovak Republic	SVK	West	5.82	2,854.48	54.74	60.12
Slovenia	SVN	West	8.98	4,080.73	68.48	71.23
Sweden	SWE	West	65.28	12,546.65	93.86	87.90
Eswatini	SWZ	Africa	7.04	647.28	44.08	40.21
Seychelles	SYC	Africa	24.46	3,109.67	60.35	54.86
Syrian Arab Republic	SYR	MENA	20.18	274.81	26.65	30.50
Chad	TCD	Africa	8.21	39.14	22.77	25.04

Togo	TGO	Africa	4.81	81.65	32.80	35.20
Thailand	THA	East and South Asia	5.39	767.09	43.51	53.13
Tajikistan	TJK	E. Europe and C. Asia	15.57	77.41	25.46	26.35
Turkmenistan	TKM	E. Europe and C. Asia	11.85	365.16	22.44	22.34
Tonga	TON	East and South Asia	19.59	662.37	40.83	55.25
Tunisia	TUN	MENA	8.94	637.20	47.53	50.01
Turkey	TUR	E. Europe and C. Asia	8.78	1,236.42	47.10	50.05
Tanzania	TZA	Africa	3.19	70.87	38.16	43.28
Uganda	UGA	Africa	12.47	59.65	31.42	42.46
Ukraine	UKR	E. Europe and C. Asia	9.38	472.86	30.66	35.43
Uruguay	URY	LAC	28.32	1,531.54	73.73	63.22
United States	USA	West	11.28	7,528.30	79.08	81.59
Uzbekistan	UZB	E. Europe and C. Asia	7.00	210.11	27.21	26.64
Venezuela	VEN	LAC	27.26	1,630.82	26.67	20.52
Vietnam	VNM	East and South Asia	6.47	94.16	39.08	43.58
Vanuatu	VUT	East and South Asia	6.33	397.86	50.63	56.89
South Africa	ZAF	Africa	8.36	1,097.63	55.35	52.97
Zambia	ZMB	Africa	4.68	135.74	38.60	42.67
Zimbabwe	ZWE	Africa	5.11	152.53	26.69	22.84

Notes: The column 'Development' shows the average indicator level (in percentage) of each country. The column 'Expenditure' shows the average budgetary capacity of each country (in real USD per capita). The column 'Monitoring' presents the average level of the indicator on the quality of monitoring (in percentage). The column 'Rule of Law' denotes the average level in the indicator on the quality of the rule of law (in percentage).
Sources: Authors' calculations with data from the 2021 Sustainable Development Report, theWorld Bank's data on government final consumption expenditure, and Worldwide Governance Indicators.

countries in the dataset in terms of development, budgetary capacity, and public governance. For instance, while Somalia presents the lowest average in 'Rule of Law' (6.17%) during the sample period, Norway exhibits the highest (89.28%). Notice also that the budgetary capacity can be as weak as that of Somalia (USD $25.88), and as strong as that of Luxembourg (USD $16,242.24). With the descriptive data of these columns, one wonders if the relationship between budgetary capacity and development is straightforward or whether it presents noise that requires a process of cleansing mediated by the countries' public governance.

We have now introduced the data that we employ throughout the second part of the book. Next, we would like to discuss some of the main empirical challenges that, in our opinion, researchers and policy analysts face when studying the expenditure–development relationship through these data. However, to properly frame these challenges, it is convenient to provide a brief overview of popular quantitative methods that some analysts may see as potential options to model relationships between indicators. By explaining the intuition behind these methods, and their limitation, we aim to create clearer grounds to justify our view on the most pressing empirical challenges.

3.2 POPULAR MODELLING FRAMEWORKS AND THEIR LIMITATIONS

Broadly speaking, we can classify the most popular quantitative methods used in development studies according to the level of aggregation at which they operate. Micro-level approaches based on field experiments through randomised controlled trials have gained popularity recently among development scholars due to their ability to isolate and estimate highly specific causal effects. However, randomised controlled trials have also received substantial criticism due to their low explanatory power in terms of (absent) causal mechanisms, limited generalisability (also referred to as lack of external validity), high implementation costs (time- and money-wise), and inability to scale

up (economic- and logistic-wise) (Deaton, 2010; Kvangraven, 2020; Ogden, 2020; Pritchett, 2021). Given these limitations, micro-level experimental methods cannot be considered suitable for addressing systemic problems, so we focus our methodological review on the most commonly used macro-level tools. Having said that, the reader should be aware that PPI also touches on micro-level elements, yet, its computational model provides an explicit set of vertical mechanisms between both levels of analysis.

Macro-level analysis has a long tradition in development economics and other social sciences. Here, we would like to briefly discuss some of the most popular tools and their limitations in the study of sustainable development, especially when using development-indicator data. Thus, we focus on methods deployed for the simultaneous analysis of several indicators as those in the SDR dataset. In doing so, we aim to lay the grounds for a more thorough discussion on the empirical challenges that development scholars and consultants have to face in the near future. Such discussion also helps us to frame the contribution of our framework to overcoming these obstacles.

3.2.1 Benchmark Analysis

Benchmarking is a popular approach used by international organisations and governments to assess the performance of indicators (Huggins and Izushi, 2009; Huggins, 2010; Valor, 2012; Allen et al., 2018; Lamichhane et al., 2021). This method consists of comparisons in a large set of indicators across countries (or regions) to set some standards, sometimes by combining countries with structural affinities. For each indicator, analysts define the gap between the maximum observed level and the level of a particular country (similar to the development gaps presented in Figure 3.2). Typical results in these exercises are rankings, trend analysis, comparative analyses, and 'traffic light' assessments, such as those presented in the Sustainable Development Report (Sachs et al., 2021).

This method is commonly employed because it lends itself to the elaboration of intuitive visualisations, which can be convenient

for policymakers that do not have the time or the technical expertise necessary to understand more formal work. Yet, as an analytical device, benchmarking lacks substantive theoretical underpinnings and statistical rigour and, thus, exhibits several limitations for policy design. First, the analysis of each indicator is isolated from the others' influences, ruling out their potential interdependencies. Second, because benchmark analysis is mainly data-driven, it lacks a theoretical framework specifying causal links between budgetary allocations (or any other policy instrument) and outcomes. In other words, it treats the expenditure–development link as a black box. This warning is important because the evidence produced through this approach may be misguided and could encourage investments in ineffective government programmes.

The third major limitation is the absence of political economy elements. For example, omissions regarding the quality of the rule of law could lead to overly optimistic interpretations and recommendations. Fourth, the aim of benchmark analysis is, essentially, to support mainly ex post evaluation. Hence, it cannot generate rigorous prospective estimates and counterfactuals. This drawback limits the ability of benchmark users to assess the potential impact of an intervention such as budgetary reallocations across the SDGs. Hopefully, these remarks provide the readers with a glimpse of the shortcomings of benchmark analysis and the potential risks of decision making based only on this type of evidence.

3.2.2 Regression Analysis

Another common framework for studying development indicators is regression analysis. These models aim at predicting a single indicator (usually related to GDP). The most common regression model is linear, although there exist formulations that consider some types of non-linearities. An obvious limitation of these models is that, generally, they focus on a single dependent variable. This feature does not work well with the systemic nature of sustainable development, where one would like to quantify outcomes across multiple

development dimensions. While there are regression studies that consider multiple dependent variables through systems of equations, their data requirements are rarely compatible with existing indicator data.

In regression-based studies, analysts presuppose that the chosen "independent" variables can be directly intervened or manipulated. Consequently, in practice, they justify interventions in those policy issues if the associated regression coefficients are statistically significant. In reality, development indicators are the result of government actions that are partly motivated by the performance of the same indicators. Thus, the apparent "exogenous" change in a covariate might originate from spillovers emerging from the movements in the dependent variable or other covariates. This interdependency translates into endogeneity issues that are untreatable with regression macro models, even if analysts attempt to deal with causality issues through instrumental variables or structural equation models, which are also described through directed acyclical graphs. A proper econometric formulation would require a multidimensional and interdependent setting. However, this demands highly context-specific data, including policy instruments, which are rarely available.[8]

Another limitation of these models is that, by assuming linear relationships, most regression-based studies imply substitutability between public policies. Thus, they cannot be binding constraints on the countries' performance. In other words, there is no natural way to account for the bottlenecks generated from the scant performance of some indicators. Hence, only the statistical and economic significance of any variable justifies the usage of the associated policy. Unfortunately, the inability to incorporate bottlenecks leads analysts

[8] Regression studies considering SDG variables usually leave out policy instruments like government expenditure. In part, the omission owes to the scarcity of government spending data. While there are some instances of regression-based studies considering expenditure (Devarajan et al., 1996; Haque, 2004; Agénor and Neanidis, 2011; Bojanic, 2013; Neduziak and Correia, 2017; Yilmaz, 2018), the vast majority of this community of models focuses on associations between indicators.

to make indicator-focused recommendations disregarding the level of the other development dimensions in a particular country.

From a systemic perspective, modelling the interaction between indicators is central. A regression framework does not have the ability to consider a relatively dense interaction structure. Given that the collection of development indicators tends to be annual, the coarse-grained nature of the data prevents the possibility of estimating many interaction terms. The number of coefficients would be incredibly high for a dataset like the SDR (with roughly 70 indicators). Some economists may argue that spatial econometric models can account for such structure in less data-demanding models. However, these regression models also suffer from parameter identification problems (Anselin et al., 2008) that can only be solved by excluding multiple interaction terms (Elhorst, 2012).

There is another major problem with regression analysis that is critical for the translation of estimates into policy recommendations. Due to the short coverage of development indicators' time series, it is rare to find regression models with many covariates for a single country (even when removing interaction terms). Thus, this analysis resorts to panel regressions by pooling multiple countries' time series to estimate coefficients that capture 'average effects'. That is to say, the interpretation or recommendation that one can derive from such estimates will not be for a specific country but for a hypothetical country with the average characteristics of the sample (Rodrik, 2009). In other words, the context-specificity needed for policy advice vanishes.

3.2.3 *General Equilibrium Models*

A general equilibrium model is a system of equations (often non-linear) that describes economic behaviour once a price vector solves those equations. Typically, each equation represents an aggregate outcome of a stylised decision process of a representative agent playing a specific role in the economy (e.g., households and firms). Similarly, there is an equation representing the government's optimal

https://doi.org/10.1017/9781009022910.006 Published online by Cambridge University Press

budget allocation. For example, the equation of a household agent is the result of their optimal choice between work and leisure, incentivised by their wage (which enables consumption) and their consumption preferences (through a utility function). When including many elements in the economy (e.g., industries, foreign markets, or environmental damage functions), these models become increasingly difficult to solve as the equilibrium conditions restrict the solution space more. Generally speaking, models empirically usable in this literature require numerical methods to be solved, so they are known as computable general equilibrium (CGE) models. See, for example, Lofgren et al. (2002).

One of the benefits of CGE models is that they come with a theoretical backbone and, thus, they are more explicit about causal mechanisms. These models provide a structure of interdependencies between different economic sectors, which the economics literature has studied for a long time. However, expanding this structure to include novel dimensions such as gender inequality and marine life imposes a heavy theoretical burden. To implement such extensions, the researcher needs to, ex ante, specify how the new sectors would be modelled and interconnected to the rest of the system.[9] This may be straightforward in the context of economic production, but not necessarily when dealing with, for example, a social-justice issue. They also contain stringent assumptions like agent homogeneity, rationality, and equilibrium, which restrict the scenarios that these models can generate. While CGEs are a popular tool to study topics such as macroeconomics (McKitrick, 1998), international trade (Nijkamp et al., 2005), climate change (Lloyd and MacLaren, 2002), and natural-resource impacts (Calvin et al., 2019), they have severe theoretical and empirical limitations. These shortcomings preclude the possibility of including political economy complications.

[9] Demanding a full theoretical specification of the model can also be a handicap as it leaves no margin for a data-driven component in which one can infer some relationships that are not well understood yet.

A recent example of a CGE model dealing with SDGs is Basheer et al. (2022). These authors elaborate on a dynamic and recursive model for assessing the systemic impacts that alternative policy options exert in a small and open economy (Egypt). The model includes four economic agents: households, governments, producers and the rest of the world. Households are disaggregated into five income quintiles and two locations (rural or urban). Producers operate in fifteen different activities and use three types of inputs (labour, land and capital). The model considers a variety of policy instruments that can integrate different portfolios.[10] Likewise, the paper shows simulation results for seven different optimisation objectives associated with SDGs.[11]

The authors use the simulation outcomes of their CGE model in an algorithm of evolutionary optimisation to search for the best policy. Then, they deploy a machine learning technique (random forests) to establish the relative influence of policy instruments on sustainability targets. Aside from the potential difficulties of the model's calibration and validation processes,[12] this example illustrates some drawbacks of this highly detailed framework that a more parsimonious PPI circumvents. The following features of PPI address issues that CGE models currently do not. First, PPI incorporates an exogenous network of conditional dependencies between development indicators (see Chapter 4) to strive for a balance between theoretical content and data-driven inference. Second, it allows for the scalability of the model and, hence, the analysis of development

[10] Change in total government transfers to households; distribution of government transfers between household groups; absolute values of income taxes; absolute values of producer taxes/subsidies; absolute values of sales taxes/subsidies on commodities.

[11] The objectives are: discounted real GDP, discounted net real urban household income, discounted net real rural household income, mean overall Gini Index, mean urban Gini Index, mean rural Gini Index, and total CO2 emissions.

[12] Calibrating CGEs becomes more difficult as one increases the number of components or dimensions, a troublesome issue when dealing with the SDG systemic perspective. This limitation directly relates to the equilibrium-oriented nature of these models, as they require satisfying a large set of constraints. This requirement often leads to an absence of real-valued (non-negative) price vectors.

performance does not have to be constrained to a limited set of indicators (here just 12) and a few SDGs (here just four: 1, 8, 10 and 13). Third, includes political economy considerations which help to assess the impact of inefficiency in the use of public resources. Fourth, it addresses many other relevant questions, such as development accelerators and structural bottlenecks.[13]

3.2.4 System Dynamics

System dynamics are aggregate stock-flow models that try to capture relationships between the different components of a system (Meadows, 1972, 1994, 2004; Forrester, 1973). Generally speaking, they are more scalable than CGEs in the number of dimensions they can handle. This analytical advantage results from the fact that system dynamics models do not rely on equilibrium-based solutions. Nevertheless, such flexibility comes at the cost of weaker theoretical support. Furthermore, even with such flexibility, an ex ante specification of each potential relationship is still required, so formulating a full-blown system dynamics model conveys a theoretical challenge. Moreover, estimating the parameters of all the relationships between variables can still be empirically unfeasible due to the lack of granular data, even if such estimates come from analysing only relationships in pairs of variables. Often, these estimates are obtained through values from other studies or isolated regressions, which is not consistent with the idea of capturing the structural features of an entire system (also is not ideal in statistical terms, as one would like to estimate them directly and simultaneously from the data).

Today, system dynamics are often combined with regression models in what is known as integrated assessment frameworks (Hughes, 1999; Pedercini and Barney, 2010; Collste et al., 2017). These models have become popular among international organisations to provide a systemic understanding of global-scale problems such as

[13] A CGE model, however, can consider a large menu of policy options (i.e., taxes, subsidies, and transfers).

climate change. Beyond sustainability in general, there is extensive literature on system dynamics models and integrated assessment frameworks referring specifically to SDGs, as indicated in the review by Moallemi et al. (2021).

In terms of incorporating government expenditure data, some of these tools combine the stock-flow connections between indicators with input–output data and a social accounting matrix. These models can simulate how alternative budgetary allocations affect the indicators. Unfortunately, despite introducing conduits to explain how these resources become indicator improvements, these models have received a lot of criticism due to their weak internal and external validity (Nordhaus, 1973). Likewise, the focus of system dynamic studies is limited to relatively few goals and interactions; hence, their usefulness for policy prioritisation in budgetary allocations subsides. Lastly, formulating agents' behaviour, learning, and heterogeneity is complicated in a macro setting like this. Due to the difficulties of connecting decision making with SDGs' performance, the system-dynamics approach leaves aside considerations related to the discrepancies between the policy design and its implementation through government programmes, among other important things.

3.2.5 *Network Analysis*

In the last decade, network analysis has become a popular approach for studying associations between development indicators. Yet the term 'network analysis' is vague as it includes a broad set of modelling frameworks. We review this literature in Ospina-Forero et al. (2022), where we classify these methods into two generations in the context of the SDGs. The first generation appeared in the development literature a decade ago. According to their construction procedures, we can further classify them into two groups: (1) subjective studies that rely on qualitative information (e.g., the conceptual description of the variables) and (2) statistical studies that make use of panel data (countries through time).

In group 1, subjective opinions determine the existence of links between development indicators (targets or SDGs). These studies take either a brainstorming approach (e.g., expertise and stakeholders' knowledge) or a heuristic approach (e.g., informal text mining) (Blanc, 2015; Weitz et al., 2018; Allen et al., 2019). In contrast, group 2 applies quantitative techniques to development-indicator data. In these studies, each node corresponds to a specific indicator, while an edge (or link) denotes a relationship between nodes. Usually, these graphs have weighted edges, and sometimes these edges have directions. Some examples of these studies are Czyżewska and Mroczek (2014); Castañeda et al. (2017); Cinicioglu et al. (2017); Pradhan et al. (2017); El-Maghrabi et al. (2018); Ceriani and Gigliarano (2020); Dawes (2020).

The second generation of network-based studies uses much more sophisticated computational and statistical techniques. Within this generation, there is a variety of approaches. Examples of these include physics-inspired models, correlation-based models, Bayesian networks, Granger-causality-inspired models, and dynamical-systems-based models (not to be confused with system dynamics), among others. Going deeper into technical details on how these various methodologies differ lies beyond the scope of this book, so we refer the reader to our comprehensive review in Ospina-Forero et al. (2022). Nevertheless, it is convenient to highlight some commonalities between all network-based approaches and their limitations.

First, to the extent of our knowledge, all network-based studies focus on indicator–indicator associations. Thus, in contrast with most regression studies, GCEs, and system dynamics, this literature does not address expenditure–indicator relationships. Consequently, recommendations for schemes prioritising policies are based exclusively on topological network metrics such as node centrality.

Second, indicator–indicator networks represent aggregate associations or conditional dependencies between indicators and cannot be thought of as causal relationships involving policy instruments, even when using causal-inference methods. The

argument behind this statement is that network-estimation methods employ a dependence account of causation, not a production account (Hall, 2004; Manzo, 2022). In contrast, causal factors in a production account help generate or bring about specific outcomes, as explained in Chapter 2. In the context of the expenditure–development relationship, one would need to include budgetary decisions, adaptation processes, and political economy factors related to government spending. Nevertheless, it is out of the question to include these social mechanisms when working exclusively with aggregate associations.

Third, except for a couple of methods (Weitz et al., 2018; Aragam et al., 2019), most network-based frameworks suffer from the same data-related constraints as regression models: they require long time series. To solve this issue, most of these studies pool countries together, so context specificity is lost as the resulting networks capture the interdependencies of a hypothetical "average" nation. Fourth, because these are data-driven methods, they lack a theoretical backbone that allows accounting for the political economy and other social mechanisms. Thus, one should be careful when producing policy recommendations with this type of study.

Despite the shortcomings of network-based studies, it is worthwhile to mention that these frameworks help generate insights concerning the structural properties of a system. That is to say, the inferred linkages between indicators may only be correlations or conditional dependencies; however, they offer analytical benefits for two main reasons. First, they provide structural information about the system. Second, they remove the burden of specifying a theory for each potential interdependence if an encompassing model uses the network as an exogenous input. We adopt this strategy in PPI to achieve a balance between theory and data. Accordingly, from our perspective, network analysis for SDGs is a promising field where more methodological developments are needed, not to produce reliable intervention analysis by itself but to complement other quantitative frameworks, such as PPI.

3.3 EMPIRICAL CHALLENGES

Now that we have provided an overview of the modelling landscape that is relevant to the systemic analysis of sustainable development, we elaborate on the major empirical challenges that need to be addressed by researchers and consultants. Our purpose here is twofold: (1) to reflect on a set of empirical issues that require solutions to produce significant progress in the study of sustainable development, and (2) to make a call for action towards their investigation. The methodology developed in this book seeks to tackle several of these challenges. While we think that the book's framework moves in a promising direction, we do not claim that we have solved them entirely. Thus, further investigations, better data, and innovative techniques are required.

3.3.1 Adapting to Coarse-Grained Indicators

As we have explained, development indicator data tend to be aggregate, have short temporal coverage, and come with low-frequency observations. Unfortunately, most statistical methods use procedures derived from limit distributions, which perform correctly with relatively large samples. This assumption is no exception in most machine learning models, which also require many observations, and often big data, to train complex algorithms. As studies on sustainable development increase the number of dimensions under analysis and their interconnections, the data available become inconvenient for conventional quantitative methods.

Some readers may see this challenge as a mere data-construction problem. Hence, they may argue that it is just a matter of time before governments and scientists start collecting data on more development dimensions, especially with the digital technologies at our disposal. The reality is that development indicators have been a work in progress for a long time, and it seems unlikely that data providers (statistics agencies and NGOs) will change their collection protocols shortly due to quality, ethics, and representativeness

https://doi.org/10.1017/9781009022910.006 Published online by Cambridge University Press

considerations. Thus, the problem of coarse-grained development indicators is unlikely to go away anytime soon, which leaves us with the option of developing new methods designed to work with the data currently available.

While PPI is taking the right direction by operating with coarse-grained indicators, more efforts are indispensable to deliver modelling tools that apply these data in real-world systemic analyses of sustainable development. Models of special relevance will be those that: strike the right balance between theoretical content and statistical rigour; are estimated/calibrated with relative ease; avoid cross-country data pooling to provide country-specific policy recommendations; are capable of being scaled; and allow comprehensive studies involving a large number of dimensions.

3.3.2 *Moving beyond Associations*

A common assumption in policy advice based on indicator associations is that policymakers can manipulate indicators defined as independent variables and directly exert a causal impact on a dependent variable. This approach is a prevalent misinterpretation of statistical analyses that, to a certain extent, has been traditionally fostered in some academic work. To avoid misleading advice it is necessary to be more explicit about the variables that are susceptible to interventions and those that are not.[14] Furthermore, there is a need to work with data related to policy instruments, not just outcome variables.

There are a few instances in which indicators may be close to instruments. For example, some governments use the number of hospital beds to measure the country's health-related capacity. In this case, it is reasonable to assume that bed counts resemble an instrument since most governments can increase the number of beds if they wish to (given they have enough funding). A bed increment, in turn, may impact various health outcomes. However, if one looks at SDG-related datasets, it becomes clear that most

[14] This is why, in Section 3.1.1, we introduce the concepts of instrumental and collateral indicators.

https://doi.org/10.1017/9781009022910.006 Published online by Cambridge University Press

indicators measure development outcomes, not instruments, so there is a need for data that capture policy instruments. This requisite is particularly notorious when thinking about the environmental dimensions of the 2030 Agenda and the pressing need to decrease the adverse impact that human activity is having on our planet. For instance, indicators related to endangered species or harmful emissions are mainly measurements of outcomes. Most indicators do not offer insights about the intervention mechanisms through which governments can impact those outcomes.

Creating indicators on policy instruments would require a substantial effort by governments. So, in the best-case scenario, it is a long-term aspiration. An alternative is to employ government expenditure data as a proxy for intervention variables. However, as we argue in Section 3.1.3, when using statistical methods for studying multidimensional development, the empirical relationship between indicators and government spending is far from obvious. In a recent study using a large-scale expenditure dataset, we find that average effects and predictive accuracy of government expenditure on indicators (using different statistical methods) are negligible (Guariso et al., 2023b). This result contradicts the consensus among academics and policymakers, stating that government spending largely shapes development. We interpret this finding as indicative of the inadequacy of purely data-driven methods to study the expenditure–indicator relationship in a multidimensional setting and with existing datasets. Therefore, causal analysis at the macro level requires working with policy instruments and analytic frameworks capable of quantifying the expenditure–indicator relation. In other words, causal inference needs an explicit model of the social mechanisms underlying this relationship.

3.3.3 Handling Complex Expenditure Linkages

If we embrace using government spending data, we will soon face the challenge of classifying expenditure categories into development-relevant policy issues. This functional classification is indispensable

https://doi.org/10.1017/9781009022910.006 Published online by Cambridge University Press

for establishing an explicit link between indicators (the outcome variables) and government programmes (the intervention variables). However, as we have discussed in Section 3.1.3, not only there are insufficient expenditure datasets with comprehensive multidimensional coverage, but the few that exist exhibit imperfect linkages to development indicators. Scenarios like this occur in many situations: an expenditure category may be vaguely defined and, thus, it may cover multiple programmes and indicators; there may be several programmes that are relevant to only one indicator, or several indicators that are relevant to a specific programme; or many government programmes may be short-lived. Anew, there is a data-creation side and a methodological-development side to this issue.

From the data creation point of view, multiple initiatives are working towards the standardisation, linkage, and publication of government expenditure data. International initiatives such as Open Spending and the SDGs have enabled governments and international organisations to work towards the creation of such datasets. For example, the Mexican Ministry of the Treasury publishes annual federal budgets at a highly disaggregate level.[15] With the advent of the SDGs, the Mexican government developed a framework to map thousands of budgetary programmes into the 169 targets of the SDGs (SHCP, 2017). This dataset was the first of its kind. Another example comes from the UN and its initiative to implement globally 'Integrated National Financing Frameworks' that can align different financing sources to the SDGs, leading to the 'Budgeting for SDGs' paradigm (Palacios et al., 2022). These efforts are being accompanied by non-governmental organisations, such as the Global Initiative for Fiscal Transparency, through the elaboration of linkage procedures. In Guariso et al. (2023a), we provide a thorough overview of this and other initiatives.

While national and international agendas to create SDG-linked data are commendable, there is a long way to generating

[15] Through its fiscal transparency portal: *Transparencia Presupuestaria*.

https://doi.org/10.1017/9781009022910.006 Published online by Cambridge University Press

internationally comprehensive datasets. Besides, there is no guarantee that one would obtain a one-to-one linkage structure that can be useful in traditional statistical methods. Therefore, the methodological side of this challenge requires innovative methods that can handle the nuances of highly structured budgeting data and are able to estimate the impact of expenditure on sustainable development, despite government programmes and indicators having different temporal coverage. For these methods to have wide acceptance among scholars and analysts, they need to be flexible enough to accommodate various non-ideal data settings.

First, one should expect a sparse temporal structure, as many government programmes are short-lived. Besides, when accounting for a temporal structure, one has to consider the overlapped existence of multiple programmes with similar objectives. Second, as with the short time series of development indicators, new quantitative methods need to operate with short expenditure time series and disparities between indicators and expenditure in their reporting frequencies. The latter is not rare since most indicators have annual observations, while some expenditure data are reported quarterly (like in Mexico) or even daily (like in Argentina).

Third, it is unlikely that the linkage between expenditure programmes and development indicators is one-to-one. In Figure 12.4 we show, for the Mexican case, the complex network structure of government programmes and indicators of socioeconomic deprivation. Fourth, because the complexity of expenditure–indicator linkages may vary depending on the level of aggregation of the available data, it is necessary to develop methods that can work with different aggregation degrees. This flexibility is especially critical for real-world applications, as the nature of expenditure data varies considerably between governments.

3.3.4 Embedding Vertical Mechanisms

Even with proper expenditure/indicator data, one cannot easily estimate the effect of government programmes on development in a

https://doi.org/10.1017/9781009022910.006 Published online by Cambridge University Press

multidimensional setting. Partly, this is because expenditure allocations do not always translate into effective spending. When a government prioritises a specific policy issue and allocates more funding as a consequence, there are political-economy factors that mediate the transition from financial resources to policy outcomes. In many cases, such mediation renders the additional funds ineffective, for example, if the agency implementing the relevant programme has limited technical capacity or when corruption prevails in the public administration. Unfortunately, much of this political economy is non-observable, at least in a way that could be quantified in detail (at best, there are aggregate public governance indicators built from perception-based surveys).

One of the reasons for a limited ability to quantify political-economy mechanisms is that they take place at different scales. One could argue that once a government allocates funding to a programme, there is a micro-level process through which policymakers transform these resources into policy implementation. However, we can only observe and quantify the outcomes of such budgetary expenses through aggregate indicators. Furthermore, depending on these aggregate outcomes, governments usually revise their priorities and reallocate resources. We already elaborated on these vertical mechanisms in Chapter 2, which connect micro decisions to macroscopic behaviour and vice-versa. Here we only want to highlight the need to specify a theory with these ubiquitous processes to complement the information provided by coarse-grained data and to build a model to estimate impacts.

Traditionally, when employing aggregate models like those revised in this section, all inference is built on macro–macro associations in the data. However, if we really want to understand the expenditure–development relationship, we need methodologies embedding vertical causal mechanisms (i.e., social mechanisms connecting different scales) in an artificial data-generating process. As mentioned in Chapter 2, the inclusion of social mechanisms is part of the production account of causation. Importantly, despite the

lack of data on vertical mechanisms (at least with the dimensionality and coverage of the SDGs), it is possible to formulate an empirical model where we can embed these theoretical channels through an efficient and robust method of indirect calibration. This approach has the added benefit of creating proper counterfactuals to produce causal inferences.

3.3.5 *Estimating Interdependency Networks*

As we have explained previously, understanding sustainable development and providing policy advice require analysing the interconnectedness of its many dimensions. We have argued that accounting for all possible interactions between policy dimensions can become extremely difficult. On the one hand, development-indicator data are not big enough to achieve this through most statistical-learning methods. On the other hand, in theoretically motivated methods such as CGEs and system dynamics, researchers need to identify and specify these many relationships. Hence, empirical methods using data to quantify such interdependencies require a balance between theoretical richness and data-driven content. From our point of view, a promising way to achieve this is by combining a model emphasising social mechanisms with empirically estimated structural information, such as networks of interdependencies between policy dimensions. We adopt such an approach in this book, acknowledging that much work still needs to be done for estimating such networks.

In Section 3.2.5, we discussed the limitations of SDG networks when used as policy tools on their own. We have also highlighted their importance in providing insightful structural information when they are used to complement other methods. Currently, various disciplines work actively in developing new methods to infer network structures from time series data; we review many of them in Ospina-Forero et al. (2022). With the coarse-grained resolution of SDG indicators and their high dimensionality, most of these methods are nowadays unfit for purpose.

https://doi.org/10.1017/9781009022910.006 Published online by Cambridge University Press

Consequently, there is a pressing need to elaborate network-estimation methods that are adequate for such data. In our framework, we adopt the method of Aragam et al. (2019) because it works well on short time series. However, like any modelling framework, this approach has several assumptions. For instance, because it is part of the Bayesian-causal-inference tradition, the method assumes that all inferred networks are acyclical. Hence, two-way interdependencies cannot be estimated. This assumption does not hold in most socioeconomic systems, so overcoming this obstacle while operating with coarse-grained data is a challenge that social and network scientists should tackle. The quantitative analyses of sustainable development would greatly benefit from these efforts.

3.4 SUMMARY AND CONCLUSIONS

The systemic study of sustainable development has seen notorious progress with the creation of data and analytical tools. More and better development indicators have become available through various sources, offering extended coverage of policy dimensions and geography. In terms of policy instrumentation, some initiatives are pushing for the publication of open expenditure data to be linked to development indicators. As part of this endeavour, the SDGs have become an interface that facilitates matching expenditure categories into relevant development indicators. In this chapter, we provide an overview of some of these datasets. As we advance over the chapters of this book, we introduce additional datasets that may be more specific to a country, topic, or expenditure source.

While data improvements represent an essential contribution, there remains a wide gap between the capabilities of existing quantitative methods and the qualities of these data. To understand why such a gap exists, we review some of the most popular modelling techniques employed in studying development and highlight some of their shortcomings when working with these data from a systemic perspective. Building on this, we discuss five empirical challenges that need solutions to make significant progress in the field.

These are (1) developing methods that operate with coarse-grained data, (2) shifting the focus from indicator–indicator associations to instrument–indicator relationships, (3) improving empirical methods to accommodate expenditure–indicator-linked data with varying degrees of structural detail, (4) embedding vertical causal mechanisms in analytic frameworks, and (5) creating network-estimation techniques that are compatible with existing indicator data. Overall, these challenges provide a general picture of the current analytic limits in the sustainable development literature, so we call for action to make progress on all these fronts.

Tackling these challenges is likely to require interdisciplinary approaches and an open mind to learn what other disciplines have achieved in their respective fields. For instance, social scientists need to be re-educated in quantitative methods to incorporate many appealing insights stemming from computer science, physics, and applied mathematics. Likewise, technocrats with skills in quantitative methodologies need a better understanding of socioeconomic systems; for example, how incentives operate, how institutions shape incentives, and how social norms emerge.