Data-Driven Approaches to Assist Achieving Sustainable Development for Nations

by

Chenyang Shuai

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Environment and Sustainability and Scientific Computing)
in The University of Michigan
2021

Doctoral Committee:

Professor Shelie Miller, Co-Chair
Professor Ming Xu, Co-Chair
Professor Maria Carmen Lemos
Professor Ji Zhu

Chenyang Shuai

cyshuai@umich.edu

ORCID ID: 0000-0001-8789-0111

Acknowledgements

My Ph.D. study was truly a wonderful and life-changing experience, which was not possible without the support and guidance that I received from many people.

First, I would like to express my sincere appreciation to my advisor Professor Ming Xu for his patience, guidance, encouragement and support. Since the day I received my Ph.D. offer, I have been much honored to become a student of Professor Xu. After I started my Ph.D. with him, Professor Xu has been very patient in teaching me about the fundamentals of scientific research. As I needed, he inspired me on my research ideas and frameworks on global sustainable development. For research outputs, he encouraged me to set high goals and never settle for less. Since I started looking for an academic position this year, he has provided me with lots of recommendations and supports without reservation. I would also thank my advisor Professor Shelie Miller for her priceless guidance and feedbacks on my research. She is a great researcher integrating theory and practice. Under her guidance, I have gradually undertaken research towards positive social impacts. I feel very privileged for having the opportunity to learn from her. I appreciate Professor Maria Carmen Lemos for her kindness and guidance on my research. During Christmas, she and her husband kindly invited us to her house to have wonderful food, wine and cake. In her research seminar course, she inspired me to conduct policy-friendly research based on my research interest. I also thank Professor Ji Zhu for providing valuable advice on the statistical part of my research. His statistics course was so interesting, practical and accessible, which introduced me data science. His guidance on my statistical logics and skills improved the robustness of my dissertation.

I also want to thank my dear friends at UM: Bu Zhao, Weicheng Huang and Mingyan Tian. I still remember we burnt the midnight oil to work on *EECS 505* and *CMPLXSYS 500* in G.G. Brown and walked home together with heavy snow. I am also grateful for the fellows in Professor Ming Xu's lab and those at the Center for Sustainable Systems for their invaluable encouragement and support: Bu Zhao, Ping Hou, Morteza Taiebat, Shen Qu and Yabin Dong, among many others.

I am also very grateful for my family for their forever support and kindness. I deeply appreciate my parents, Yukang Shuai and Ling Chen, for raising me up with thirty years' long efforts and dedication. Other than the love they give me, my father particularly cultivated me with good habits and the way of thinking. At last, I thank my life partner and lovely wife, Dr. Xi Chen, for her precious love, encouragement and company. The days we spent together are filled with joy and love.

Table of Contents

List of Tables

List of Figures

List of Appendices

Abstract

The United Nations set 17 Sustainable Development Goals (SDGs) in 2015 as a universal call aiming to end poverty, protect the planet, and ensure that all people enjoy peace and prosperity by 2030. These 17 SDGs are committed by 193 countries and regions to transform the conventional development agenda for sustainable development. My dissertation focuses on using data-driven approaches to address some of the challenges in SDG implementation for nations, including challenges in data collection, performance comparison, and prediction.

To monitor the progress towards achieving SDGs, the 17 goals are underpinned by 169 targets which are measured by an even larger number of SDG indicators. The sheer number of SDG indicators makes data collection a critical challenge. My dissertation begins with identifying the principal indicators, the changes of which can represent the variations of the majority of SDG indicators with the lowest difficulty of data collection. Integrating principal component analysis and multiple regression, I identify 147 principal indicators that can explain at least 90% of the annual variation of 351 SDG indicators. My results can guide future investment in the data infrastructure for SDG monitoring by giving priorities to these principal indicators for global comparison.

Per capita based metrics, such as GDP per capita, are widely used in SDG performance comparison, which assumes stock measures (e.g., GDP) scale linearly with population. However, this assumption does not always hold since it ignores the effect of agglomeration resulting from non-linear interactions in social dynamics. I find extensive empirical evidence that many

important national development indicators scale non-linearly with population size, which provides a quantitative argument against the mainstream practice to compare national development using per-capita measures. I further propose a quantitative framework to explain the scaling in nations originating from the scaling in cities.

The global progress to achieve the SDGs by 2030 has been stalled by the coronavirus disease 2019 (COVID-19) pandemic. Several studies have qualitatively assessed the impacts of COVID-19 on SDGs. Quantitative assessments, however, are rare, largely due to the complex non-linear relationship among SDG indicators making prediction difficult. I use machine learning approaches to capture the complex non-linear relationship between SDG indicators and evaluate the impacts of COVID-19 on SDGs. I find that the overall SDG performance declined by 7.7% in 2020 at the global scale, with the performance of 12 socioeconomic SDGs decreasing by 3.0-22.3% and that of 4 environmental SDGs increasing by 1.6-9.2%. By 2024, the progress of 12 SDGs will lag behind for one to eight years compared to their pre-COVID-19 trajectories, while extra time will be gained for 4 environment-related SDGs. Furthermore, the pandemic will cause more impact on emerging market and developing economy than on advanced economy, and the latter will recover more quickly to be close to their pre-COVID-19 trajectories by 2024.

Chapter 1. Introduction

1.1. Background

The concept of development has been variously defined in literature. The modernization theory can be traced back to the late 1940s, when capitalism and communism competed most furiously in the cold war political background[1]. Modernization refers to a progressive transition from a "tradition" to "modern" society, which can be accomplished by the adoption of western cultural and institutional practices[2]. In this general context, the core aim of developing countries was to catch up with the advanced industrialized countries by economic growth[1]. Dependency theory arose from a growing association of southern hemisphere nationalists (mainly Latin America and Africa) and Marxists[3]. It divided the world into "core" (i.e., developed countries) and "periphery" (i.e., developing countries). The periphery's disadvantageous position has caused it to be impoverished by the core, which makes development difficult[4]. Different from modernization theory, dependency theory defines development as a social process, which integrates social equality or basic human needs into economic growth[1]. In Sen's entitlement theory, development was a process of expanding the real freedoms that people enjoy[5]. Sen defined five types of freedoms including political freedoms, economic facilities, social opportunities, transparency guarantees, and protective security[5]. The theory has been widely recognized by the social study institutions. For example, the well-known Human Development Index developed by the United Nations is anchored in Sen's theory[6].

With the increasing conflict between human society and the environment, the dimensions of development have been extended to a sustainable thinking. The concept of sustainable development, which was derived mostly from the 1987 Brundtland report, is defined as meeting the needs of the present without compromising the ability of future generations to meet their own needs[7]. Sustainable development has shifted the focus of development towards social development and environmental protection for future generations in addition to economic development. In 2000, the United Nations set Millennium Development Goals (MDGs) to guide and help the development of the developing and least developed countries. Specifically, MDGs contain eight goals including eradicating extreme poverty and hunger, achieving universal primary education, promoting gender equality and empower women, reducing child mortality, improving maternal health, combating HIV/AIDS, malaria, and other diseases, ensuring environmental sustainability, developing a global partnership for development[8].

As an extension to MDGs, Sustainable Development Goals (SDGs) set by the United Nations in 2015 are a collection of 17 interlinked global goals designed to be a “blueprint to achieve a better and more sustainable future for all by 2030”, which has been adopted by 193 countries and regions[9]. To facilitate monitoring the progress towards achieving SDGs, the 17 goals are underpinned by 169 targets which are further measured by an even larger number of SDG indicators[9]. These SDG indicators are tracked at the national level by several global organizations. For example, the World Bank maintains a database of 351 indicators for 217 countries and regions with data available since 1990 to monitor the progress of each nation or region towards the 17 SDGs[10].

Three key challenges arise in SDG implementation for nations: 1) costly data collection, 2) biased performance comparison, and 3) less accurate performance prediction (Figure 1-1).

First, the large number of SDG indicators makes data collection expensive and time-consuming. For example, the total estimated cost is at nearly $45 billion for collecting data to quantify all SDG indicators for all countries over the SDG period[11], more than the UN's annual expenditure in 2016[12]. This calls for an urgent need to find a smaller number of principal indicators that are cheap to collect but can still provide sufficient information for monitoring the SDGs. Identifying such principal indicators from a whole indicator set is generally known as dimensionality reduction in which the number of variables in a dataset is reduced by removing some variables without losing valuable information (i.e., variance)[13]. Two primary methods for dimensionality reduction are principal component analysis (PCA) and factor analysis (FA). PCA conducts dimension reduction by projecting each data point into a few principal components to obtain lower-dimensional data while preserving as much of the data variation as possible[14, 15]. For example, Jiang *et al*. found a first component that can explain up to 85% variation of a set of 28 sustainable development indicators[16]. On the other hand, FA is a statistical method used to reduce the observed and correlated variables into a lower number of unobserved variables called factors plus error terms[17]. For instance, Laurett *et al*. concentrated 25 sustainable development-related variables into three factors including natural agriculture, innovation and technology, and environmental aspects using FA[18]. Both PCA and FA identify a smaller number of new variables respectively called principal components or factors, which are linear combinations of the original variables, to explain most of the variance of the dataset[19]. However, my goal here is to find a subset of the original variables rather than a set of new variables. Therefore, I use a hybrid approach by combining PCA and multiple regression to identify the principal indicators with the least collection difficulty.

Second, the use of per capita indicators makes the comparison of sustainable development performance biased. For example, a wide range of per capita indicators (e.g., GDP per capita and greenhouse gas emissions per capita) are frequently used to compare progress towards sustainable development among countries[10]. The per capita based comparison relies on a strong assumption that, on average, indicators measuring the size of stocks (e.g., GDP) scale linearly with the population[20]. However, this assumption does not always hold as it ignores the effect of agglomeration resulting from non-linear interactions in social dynamics[20]. The urban science literature has found that many socioeconomic outputs (e.g., GDP, wages, crimes and innovation) in cities can be characterized by the ubiquitous scaling law—$Y \sim Y_0 N^{\beta}$—where Y is an indicator of output, Y_0 is the baseline common to all cities, N is the city population size, and β is the scale-invariant elasticity indicating the percentage change in Y following a 1% increase in N[21, 22]. Non-linear scaling ($\beta \neq 1$) has been widely found in urban systems for distinct indicators[23-27]. However, little is known whether the per capita measures are suitable for national systems. Given that each country is essentially an ensemble of urban and rural areas, I hypothesize that similar scaling found for cities also exists for countries. At the same time, if such scaling does exist for countries, they are likely different from those found for cities, because an ensemble of cities does not equal a bigger city. Therefore, I examine sustainable development indicators for countries to test the scaling of these indicators with the population at the national scale and explore the origins of such scaling.

Third, the complex non-linear relationship among SDG indicators makes the traditional linear statistical prediction models less effective. As a result, quantitative evaluations of the coronavirus disease 2019 (COVID-19) impacts on SDGs are rare, although qualitative assessments exist[28-35]. For example, the UN's 2020 annual report on SDGs qualitatively showed worrisome initial impacts of COVID-19 on some specific goals and targets[28]. Similarly, Nundy *et al*. qualitatively evaluated the impact of COVID-19 on SDGs related to socioeconomic, energy-environment, and transport sectors in 2020[36]. One of the reasons why quantitative assessments are lacking is that non-linearities, which are common in complex systems, cannot be sufficiently captured by traditional linear models[37]. Without a quantitative evaluation, however, it is difficult to understand the impacts on specific SDGs, SDG targets, and SDG indicators for countries. On the other hand, machine learning approaches can estimate complex non-linear relationships between the response and predictors with better prediction accuracy[38-42]. In this dissertation, I develop machine learning models to quantitatively assess the impacts of COVID-19 on SDGs.

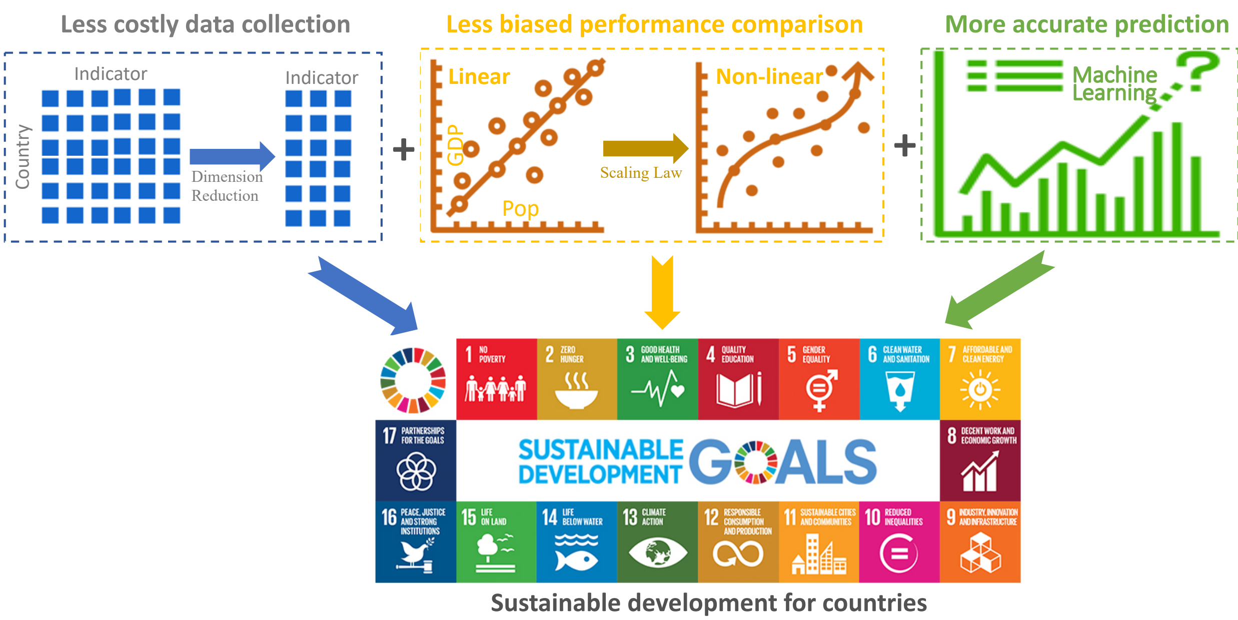

Figure 1-1. Three challenges in SDG implementation.

1.2. Research Questions

The overall research question of my dissertation is: how to use data-driven approaches to help address data collection, performance comparison, and prediction challenges in sustainable development for nations?

Regarding the data collection challenge, I focus on identifying the principal indicators to reduce the data collection cost (Chapter 2). Specifically, I address the following question:

- What are the principal indicators that can represent at least 90% of the variation of all the 351 SDG indicators from the World Bank database with the lowest difficulty of data collection?

Regarding the performance assessment challenge, I focus on examining the scaling of various metrics with population for countries based on the scaling law found in cities (Chapter 3). Specifically, I address the following questions:

(1) Do countries have similar scaling laws as found in cities?

(2) If so, how to explain the origin of such scaling in countries?

Regarding the performance prediction challenge, I focus on developing machine learning models to predict the impact of COVID-19 on SDGs (Chapter 4). Specifically, I address the following questions:

(1) What are the impacts of COVID-19 on SDGs?

(2) How do the impacts differ between emerging market and developing economy (EMDE) and advanced economy (AD)?

1.3. Structure of the Dissertation and Contributions

The remainder of the dissertation is organized as follows. Chapter 2 identifies the principal indicators of all SDG indicators. Chapter 3 examines the scaling of various national metrics with population. Chapter 4 develops machine learning models to predict the COVID-19 impact on SDGs.

In Chapter 2, I develop a hybrid model which integrates principal component analysis and multiple regression to identify the principal indicators of all SDG indicators considering the collection cost for each SDG indicator. The results can guide future investment in building the data infrastructure for SDG monitoring to give priorities to these principal indicators for global comparison. A manuscript based on this work has been published in the journal *Environmental Research Letters*[43].

In Chapter 3, I examine the scaling of various SDG indicators with population for countries and further develop a quantitative framework to explain the origins of such scaling. The results provide a quantitative argument against the mainstream practice of comparing national development using per capita measures. A manuscript based on this work is in preparation.

In Chapter 4, I develop machine learning models to predict the impact of COVID-19 on SDGs from 2020 to 2024 using projected GDP growth and population. The results help government and non-state stakeholders identify critical areas for targeted policy to resume and speed up the progress to achieve SDGs by 2030. A manuscript based on this work is in preparation.

Finally, in Chapter 5, I identify knowledge gaps, draw conclusions on the findings, and offer recommendations for future research.

Chapter 2. Principal Indicators for Monitoring Sustainable Development

2.1. Introduction

Collecting data to regularly monitor the SDGs is not an easy task[44-46]. Such efforts need significant investment in institutional infrastructure and financial resources and engagement with a vast number of stakeholders. For example, over 1,200 stakeholders worldwide have contributed to data collection for SDG indicators, including governments, NGOs, research institutions, multilateral organizations, and private sectors[47]. The total estimated cost is at nearly $45 billion for collecting data to measure all SDG indicators for all countries and regions until 2030 when the SDGs are supposed to be achieved[11], more than the UN's annual expenditure in 2016[12]. Despite many achievements, it is still challenging to annually update the sheer number of SDG indicators for all countries and regions[48, 49]. This challenge calls for alternative approaches to monitor the SDGs at a lower cost.

One way to reduce the data collection cost for SDG monitoring is to identify a subset of the SDG indicators as "principal indicators", so that the changes of these principal indicators can represent the changes of all indicators. Therefore, progress towards achieving the SDGs can be monitored by only using these principal indicators with much less cost and efforts, rather than relying on all indicators. Identifying such a subset of principal indicators from a whole set is generally known as dimensionality reduction in which the number of variables in a dataset is reduced by removing some variables without losing valuable information (i.e., variance)[13]. Dimensionality reduction requires strong correlations between variables. Indeed, many studies as well as my analysis (Figure A-1) have shown that the SDG goals, targets, and indicators are highly correlated with each other[50-54]. Such correlation indicates that, with appropriate methods, it is possible to extract a small number of principal indicators so that their variations can represent the variations of the entire set of SDG indicators.

The central question my dissertation aims to answer is, given the difficulty of data collection for individual SDG indicators, what are the principal indicators that can adequately monitor both the historical and future SDG progress with minimal effort of data collection. Specifically, I apply a hybrid approach by combining PCA and multiple regression to identify the principal indicators (see Data and Method section)[55, 56].

Using the hybrid approach of dimensionality reduction, I examine a World Bank dataset of 351 SDG indicators for 217 countries and regions from 2000 to 2017 (Method) to find principal indicators that are able to explain at least 90% of the variance—a benchmark criterion I choose—for all SDG indicators. Specifically, this dataset is approximately 42% complete with the amount of missing data ranging from 1% to 98% for individual SDG indicators and 38%-98% for countries and regions (Figures A-2&A-3). I use the ratio of missing data (i.e., missing rate) for each SDG indicator in the latest year as a proxy to measure the difficulty of data collection. I also use the missing rate of indicators as a constraint to select principal indicators. Higher missing rate means higher collection cost of the indicator. I firstly identify the best set of historical data that can be trained for selecting principal indicators for future SDG progress under different missing rate constraints. I secondly determine the number of principal indicators for future SDG progress under different missing rate constraint. Using the best training set and the number of principal indicators, I then select the final set of principal indicators that can represent at least 90% of both the past (2000-2017) and future (2018-2030) variances of the SDG indicators with the least effort of data collection in the future.

2.2. Data and Methods

2.2.1. Data

I use the World Bank dataset of SDG indicators obtained in July 2020 which originally includes 358 indicators for the 17 SDGs over the past 29 years for 217 countries and regions and 46 country groups (e.g., the Euro area, OECD members, and Least Developed Countries). In this research, I only use data from 2000 to 2017 because data in other years are substantially incomplete (Figure A-3). I also exclude seven indicators due to lack of data for 2000-2017. Lastly I only consider data for countries or regions excluding data for country groups. As a result, I have a dataset of 351 SDG indicators each of which is associated with one of the 17 SDGs for 217 countries and regions for each year from 2000 to 2017 (Table A-1). I use the portion of missing data of an indicator (i.e., missing rate) in the latest year with available data as proxy of the difficulty of data collection. Two assumptions are made here. First, low missing rate means it is relatively easy and cheap to collect data for these indicators for most countries and regions. Second, if a country or region collects data for an indicator in one year, it will likely continue to do so in the future. For most indicators, the latest available year is 2017, the last year in the dataset. However, there are some exceptions. For example, the latest data for indicator "CO_2 emissions (metric tons per capita)" in the World Bank dataset is for the year 2014, possibly because of delay in data compilation. For these exceptions, data in the actual latest year are used to measure missing rate to approximate the difficulty of data collection.

2.2.2. Explained Variance Calculation

I first calculate pairwise Pearson correlation coefficients for the 351 SDG indicators and generate a 351-by-351 correlation matrix. This is a non-positive-semidefinite (PSD) correlation matrix due to missing data of several indicators during several years. To prepare for the next step of calculating the explained variance of the subset indicators on the entire dataset, which requires a PSD correlation matrix[57], I calculate the nearest positive-semidefinite (PSD) correlation matrix using "nearPD" function in R[58]. The explained variance can be considered as the goodness of fit (R^2) of the multivariate multiple regression model in which the subset indicators are predictors and all the 351 indicators are responses.

Next, I calculate the explained variance of the subset of *k* indicators on the entire dataset (*X*) using the following equation[55]:

$$EP_{(k,X)} = [corr(X, P_k X)]^2 = \frac{tr([S^2]_{(k)} S_k^{-1})}{tr(S)} \tag{2-1}$$

where *corr* denotes the matrix correlation, *tr* is the trace of matrix, P_k is the matrix of orthogonal projections on the subspace spanned by given *k* indicators, S is the PSD correlation matrix from the above step, and S_k is the submatrix of matrix *S* with indices of *k* indicators. The algorithm for searching the highest explained variance of *k* indicators is shown in the reference[59]. These *k* indicators are defined as the principal indicators with size *k*. I then can identify the smallest number (*m*) of indicators for any threshold of explained variance (90% in this study). In practice, I use the "improve" function from the R package "subselect"[60] to achieve the largest explained variance. I then select the principal indicators for different missing rate thresholds.

I compare the explained variances on the entire dataset between using the identified principal indicators and using randomly selected subsets of indicators with the same size to demonstrate the uniqueness of the principal indicators (Figures A-4&A-5). I also provide the marginal explained variance to validate the selection of the principal indicators.

2.2.3. Marginal Explained Variance Calculation

To validate the selected principal indicators are good proxy for the entire dataset, I examine the marginal explained variance of the principal indicators and non-principal indicator. I calculate the marginal explained variance of each individual principal indicator *i* on the entire dataset ($MEP_{(i,X)}$), which is the difference between the explained variance of all principal indicators ($EP_{(k,X)}$) and the explained variance of the principal indicators except the target one ($EP_{(k-1,X)}$):

$$MEP_{(i,X)} = EP_{(k,X)} - EP_{(k-1,X)} \tag{2-2}$$

Similarly, the marginal explained variance of each non-principal indicator *j* can also be calculated. I first rank the *k* principal indicators based on their marginal explained variance, and then calculate the explained variance of the set of principal indicators except the one with the smallest marginal explained variance (set u). Next I calculate the explained variance of the set of indicators including the set u and one additional non-principal indicator (set *v*). The difference between the two explained variance is the marginal explained variance of the non-principal indicator ($MEP_{(j,X)}$):

$$MEP_{(j,X)} = EP_{(v,X)} - EP_{(u,X)} \tag{2-3}$$

An example of validation for 77 principal indicators that can explain 90% of the variance for the entire dataset when I don't consider difficulty of data collection (missing rate threshold = 100%) is shown in Figures A-6&A-7. Note that these 77 principal indicators can only represent 90% of the variance of the entire dataset rather than data for each year, which will need 94 principal indicators, when I don't consider the difficulty of data collection (missing rate threshold = 100%)

2.3. Results

2.3.1. Best Training Set

I first examine how much future variance of the SDG indicators can be explained by principal indicators identified from various training sets. Specifically, I split the entire dataset by years into a training set and a test set. In each split, the training set includes the data for all SDG indicators in all countries and regions in a given number of consecutive years, while the test set is the data for each single year after the last year of the training set representing the future. For example, if the training set is the data from 2000 to 2014, there are three test sets which are for 2015, 2016, and 2017, respectively. For each training set, I measure how much variance 100 principal indicators can explain for each corresponding test set as a benchmark. Then I vary the number of principal indicators to examine the impact on the explained variance.

Figure 2-1 shows the explained variance of selected principal indicators in each data split. Each panel (Panels A-F) selects principal indicators only from indicators with data missing rate lower than a threshold. Therefore the threshold of 100% (Panel A) means all indicators will be considered as candidates for principal indicators, implying that I do not consider the difficulty of data collection. In this case, principal indicators identified using the latest single-year data as the training set can explain the largest variance for test sets which represent future SDG progress. On the other hand, as shown in Panels B-F, the entire historical dataset is the best training set if I consider the difficulty of data collection (missing rate threshold ≠ 100%). For example, Panel F shows that, when I only select principal indicators from indicators with less than 50% missing rate, the longer the training set period is, the more variance can be explained for the test sets. I can find similar results when varying the number of principal indicators (Figure A-8). Therefore, I will use the entire dataset (2000-2017) as the training set to identify principal indicators that are expected to be able to explain the most variance of the 351 SDG indicators in the future.

Figure 2-1. Explained variance of the 100 principal indicators identified from training sets on a test set with fixed period between the test set year and the last year of the training set under different missing rate constraints. ΔT indicates the period between the test set year and the last of the training set years. (A) - (F) principal indicator with less than 50%, 60%, 70%, 80%, 90% and 100% missing rate respectively.

2.3.2. Principal Indicators for Past and Future SDG Progress

Using the entire historical dataset, I select principal indicators that can represent at least 90% of the variance of all SDG indicators in each year between 2000 and 2017 under various missing rate thresholds. I then use the total number of missing data points for the principal indicators in the most recent year to represent the difficulty of data collection. This criterion simultaneously considers both the number of principal indicators and the portion of missing data in each indicator. The set of principal indicators that has the least number of missing data points is considered as the best to represent the variances of the SDG indicators in the past. Since I select these principal indicators using the best training set identified before, the selected principal indicators are also expected to be able to represent the most variance of SDG indicators in the future.

As shown in Figure 2-2, Panel A, when the missing rate threshold is low, I have less candidate indicators to select from and thus more principal indicators are needed to explain at least 90% of the annual variances of the SDG indicator data in the past. I need 94 principal indicators to explain at least 90% of the variances when I don’t consider the difficulty of data collection (missing rate threshold = 100%). But the number of principal indicators increases to 99, 106, 118, 129, 147, and 159 when the missing rate threshold is 90%, 80%, 70%, 60%, 50%, and 48%, respectively (Figure A-9). Note that it is not possible to explain at least 90% of the variances anymore when the missing rate threshold is less than 48% (not enough candidate indicators).

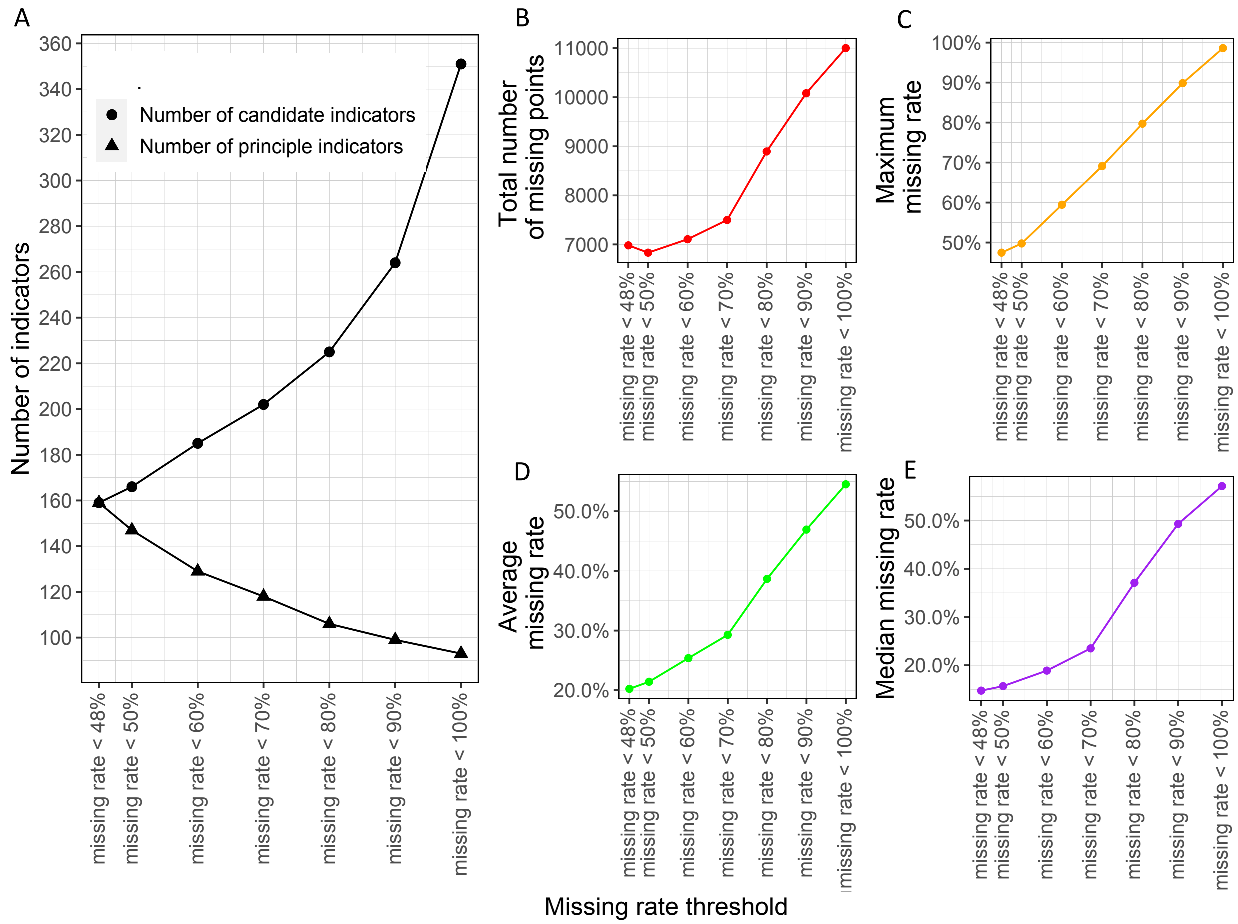

Figure 2-2. Results of principal indicators selection.

(A) Number of principal indicators to explain at least 90% of the annual variances from 2000 to 2017 under different missing rate thresholds. The line with dots represents the number of candidate indicators that meet the missing rate threshold requirement, and the line with triangles represents the number of principal indicators. (B) Total number of missing data points in each set of principal indicators. (C)-(E) Maximum, average, and median missing rate of principal indicators, respectively.

Panels B-E shows that 147 principal indicators identified under the 50% missing rate threshold (each principal indicator with no more than 50% data missing) have the lowest total number of missing data points (6,832). In addition, these 147 principal indicators also have low maximum, average, and median missing rates compared to other sets of principal indicators identified under other missing rate thresholds (Table 2-1). As a result, I consider these 147 indicators as the best set of principal indicators that are able to explain at least 90% of the annual variances of the SDG indicators in the past (2000-2017), are expected to explain the most annual variances in the future (2018-2030) (Figure 2-3), and has the lowest difficulty of data collection.

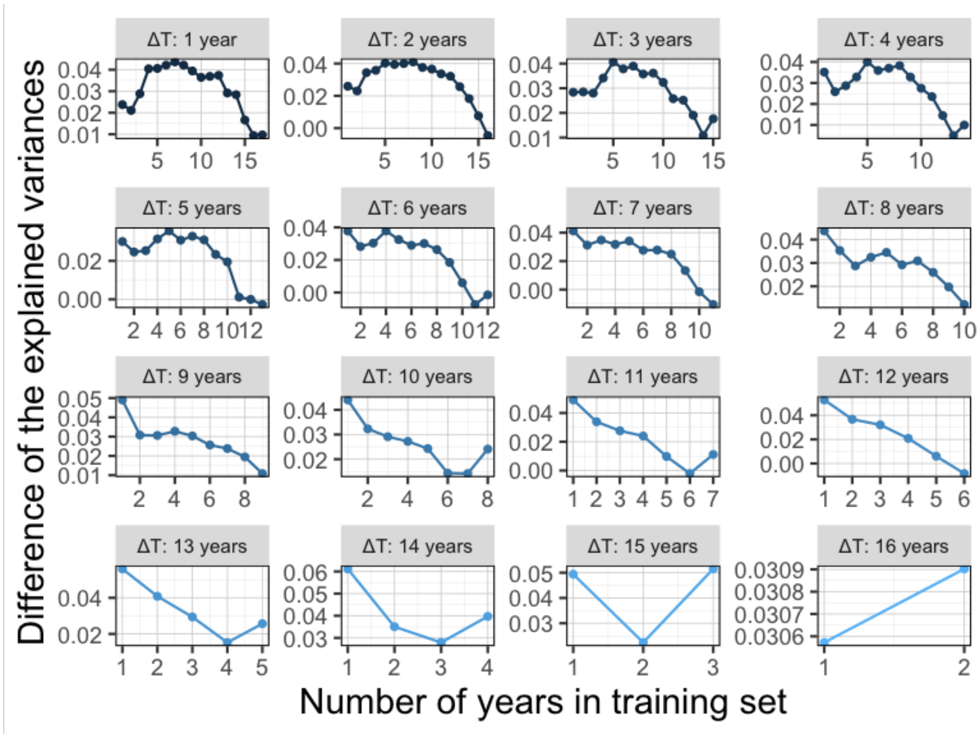

Figure 2-3. Difference of the explained variance of the 147 selected principal indicators with less than 50% missing rate on the training set and that on the test sets.
ΔT indicates the period between the test set year and the last year of the training set. Note that the average difference is only 2.5%. We further test the explained variance of the 147 principal indicators on the entire dataset (best training set). Result shows that they can explain over 92% of the variance on the training set, indicating these 147 principal indicators are expected to explain nearly 90% (92% - 2.5% ≈ 90%) variance on the future dataset.

Table 2-1. Principal indicators that can represent at least 90% of the variance of all SDG indicators. Note that they represent the past (2000-2017) under different missing rate thresholds**.**

Missing rate constraint	< 57%	< 55%	< 52%	< 50%
Number of principal indicators	133	138	143	147
Total number of missing points	6989	6956	6836	6832
Maximum missing rate	56.68%	54.38%	51.61%	49.77%
Average missing rate	24.22%	22.89%	22.03%	21.42%
Median missing rate	17.97%	15.67%	15.67%	15.67%

Figure 2-4 highlights the 147 principal indicators among all SDG indicators (Table A-2). These principal indicators belong to 14 of the 17 SDGs. No indicators in three SDGs—Goal 1

"No Poverty", Goal 13 "Climate Action", and Goal 16 "Peace, Justice and institutions"—are selected as principal indicators.

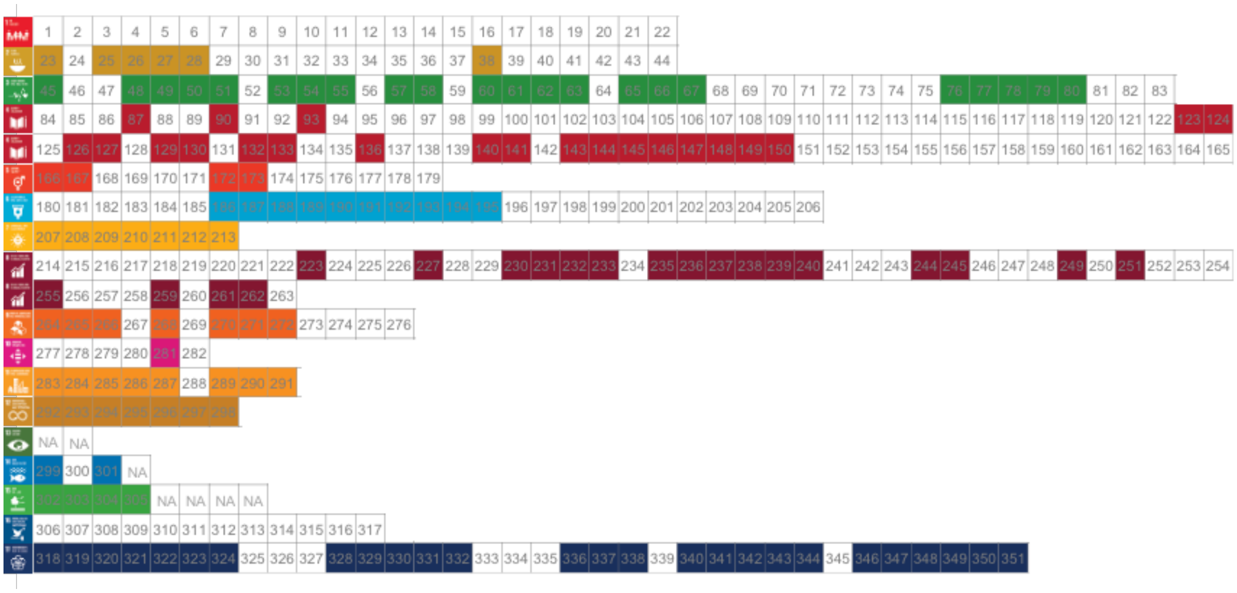

Figure 2-4. The 147 principal indicators highlighted among the full set of 351 SDG indicators.

For SDG 1 (No Poverty), data for its indicators are largely missing (missing rate > 85%). As a result, SDG 1 indicators are excluded as candidates when the missing rate threshold is lower than 85%. More importantly, SDG 1 indicators can be represented by many principal indicators which are highly correlated with national poverty measures. For example, it is widely recognized that access to sanitation infrastructure can help alleviate poverty[61, 62]. This is also supported by the strong correlation (Pearson correlation coefficient: -0.76) between the principal indicator "People using safely managed sanitation services (% of population)" (Goal 6 "Clean Water and Sanitation") and SDG 1 indicator "Rural poverty headcount ratio at national poverty lines (% of rural population)".

For SDG 13 (Climate Action), there are only two indicators and both do not have any data. Even if data were available for these SDG 13 indicators, they may still not be selected as principal indicators because many existing principal indicators are closely related to SDG 13 and could well represent the variances of SDG 13 indicators. Note that the UN uses seven different indicators for SDG 13, most of which are global-scale indicators, such as “Number of countries with national and local disaster risk reduction strategies”. Therefore they are not included in the World Bank dataset used in this study.

For SDG 16 (Peace, Justice and Institutions), its indicators have more than 66% of missing rate, and thus are excluded as candidates for principal indicators when the missing rate threshold is lower than 66%. Similarly, some principal indicators can already represent SDG 16. For example, SDG 3 indicator “Mortality rate attributed to unsafe water, unsafe sanitation and lack of hygiene (per 100,000 population)” is highly correlated (Pearson correlation coefficient: -0.91) with SDG 16 indicator “Completeness of birth registration (%)”.

2.4. Discussion

I identify 147 principal indicators that can represent at least 90% of the yearly variance of a full set of 351 SDG indicators in the past (2000-2017) and are expected to do so for the future (2018-2030) with the lowest difficulty of data collection. Without tracking the full set of 351 SDG indicators many of which have highly incomplete data, these 147 principal indicators are sufficient to evaluate and monitor the progress of countries and regions towards SDGs.

The UN identifies invisibility and inequality as the two big global challenges for the current state of SDG data[63], and the large amount of data (unaffordable cost) and declining finance are two major causes[11, 64]. The principal indicators I identified can help address these challenges. These principal indicators have relatively better data availability and can sufficiently monitor SDG progress. They can thus reduce the amount of data needed for SDG monitoring. Moreover, with limited and even declining financial resources, investment in SDG data infrastructure needs to be strategic and considers the principal indicators as priorities, especially for developing countries or regions with substantial data challenges (Table A-3).

The results do not necessarily recommend to stop tracking non-principal indicators, as established systems might already exist to collect data for those indicators for other purposes. However, my method is based on minimizing the difficulty of data collection; therefore indicators with established systems across countries and regions (thus likely low missing rate) are highly likely to be selected as principal indicators. Indeed, the 147 principal indicators generally have better data availability than non-principal indicators, with the average and median missing rates of 21.4% and 15.7%, respectively. In contrast, the average and median missing rates of the non-principal indicators are 79.6% and 84.3%, respectively. The situation that an indicator is well tracked in some countries or regions but not in others is rare. About 90% of the countries and regions (194 out of 217) in the dataset have very similar structure of missing rates across indicators (correlation coefficients > 0.5) during the study period (Figure A-10). This means, if an indicator does not have data in some countries or regions, it will likely be the same in others. Regardless, investment in SDG data infrastructure should give priorities to these principal indicators for better cross-country (region) comparison, as they have low missing data rates in the past and the difficulty of future data collection is low.

To ensure the representativeness of the principal indicators for all SDGs, I can force to select at least one indicator from each SDG as principal indicators, except for SDG 13 indicators of which do not have any data. By adding each indicator in SDG 1 and 16 as a principal indicator respectively, the additional explained variances are similar and small (between 0.003 and 0.006) (Figures A-11&A-12). Given this, I recommend to select the indicators "Poverty headcount ratio at national poverty lines (% of population)" and "Intentional homicides (per 100,000 people)", because they have the lowest missing rates among all indicators in SDG 1 (85%) and SDG 16 (55%), respectively.

Building on the principal indicators, I may consider developing an integrated index or a composite indicator to represent the SDG indicators for an overall evaluation of SDG progress for countries and regions[49]. Given that the data availability of many non-principal indicators is low, it may be better to use the principal indicators rather than the entire set of SDG indicators to develop the index or composite indicator.

2.5. Limitations and Future Research

The principal indicators are identified based on the historical correlations between individual indicators. However, some correlations may change over time. For example, poverty and food security are often correlated strongly with each other; but it is possible that poverty is alleviated by growing cash crops which may worsen food security. Therefore, a regular examination of the principal indicators is necessary to identify those changed correlation relationships and update the principal indicators. In addition, the correlations between SDG indicators do not exactly mean causality. Thus, the results are not intended to direct investment on SDGs themselves, but to guide investment on data infrastructure to monitor SDGs.

The principal indicators can represent 90% of the variance of all the indicators on average. Our method considers all the indicators as a whole but is unable to identify how much variance of a specific indicator can be represented by principal indicators. This means some indicators are well represented but others are not. In addition, I set at least 90% of the variance explained as the benchmark criterion to select the principal indicators, which is relatively arbitrary. In future research, this criterion needs to be further refined to consider preferences from stakeholders.

Chapter 3. Scaling in Nations and Its Origins

3.1. Introduction

A wide range of indicators are frequently used to compare progress towards sustainable development among countries. Given that countries vary in population size, many of these indicators that measure the level of stocks are normalized by population for fair comparisons, such as gross domestic product (GDP) per capita, health expenditure per capita, and greenhouse gas (GHG) emissions per capita[10, 65-67]. Despite of wide use, population-normalized indicators rely on a strong assumption that, on average, they increase linearly with the population[20]. However, this assumption is not always hold since it ignores the effect of agglomeration resulting from non-linear interactions in social dynamics[20]. Therefore, a more appropriate approach is needed to compare countries for their progress toward sustainable development by taking into account the non-linear relationship between the population and sustainable development indicators.

A possible solution lies in the lessons learned from the emerging field of urban science. The urban science literature has found that many outputs (e.g., GDP, wages, crimes, innovation, and contagious disease) in cities can be determined by the ubiquitous scaling law—$Y \sim Y_0 N^{\beta}$—where Y is an indicator of output, Y_0 is the baseline common to all cities, N is city population size, and β is the scale-invariant elasticity indicating the percentage change in Y following a 1% increase in N[21, 22]. Non-linear scaling ($\beta \neq 1$) has been widely found in urban systems with distinct indicators[21-23, 68-71]. In 2007, Bettencourt *et al.* firstly found that socioeconomic indicators, such as GDP, wages, patents, serious crime and AIDS, scale super-linearly ($\beta \approx 1.2$) with population among cities belonging to the same urban system[23]. Since then, the non-linear scaling law has been validated by various following literature across different nations and times[24-27]. In 2013, Bettencourt further developed a quantitative theoretical framework to explore the origins of non-linear scaling in cities[68]. The basic idea of non-linear scaling is derived from the non-linear social interactions (e.g., friendship, employment and acquaintance)[68]. Little is known about such scaling in countries. Each country is essentially an ensemble of urban areas and rural areas[72]. It is thus reasonable to hypothesize that scaling law found for cities may also exist for countries to some extent. However, they are likely to be different from that found for cities, because an ensemble of cities does not equal to a larger city, and there will always be people living in rural areas. Therefore, uncovering the scaling for countries can greatly improve the way to evaluate the growth of countries and provide policy implications towards achieving sustainable development goals. Here I examine 58 sustainable development indicators for 213 countries and regions from 1995 to 2019 compiled to test the scaling of these indicators with population and explore the origins of such scaling.

3.2. Data and Methods

3.2.1. Data

I collected data of 58 development indicators of 213 countries and regions for 1995-2019 from various databases including World Bank database[73], Our World in Data[74], and United Nation Crime database[75]. These indicators are categorized into three groups including 36 indicators of socioeconomic activities and 8 indicators of public health, and 14 indicators of individual needs (Figure 3-1).

3.2.2. Methods

The scaling of an indicator is expressed as follows[21, 22]:

$$Y = Y_0 N^{\beta} \tag{3-1}$$

where Y indicates a certain indicator (e.g., GDP) of a country, N is the total population of a country, Y_0 is a normalization constant, and β is the scale-invariant elasticity indicating the percentage change in Y following a 1% increase in N. If I take the log for both sides, the equation can be rewritten as follows:

$$\log Y = \log Y_0 + \beta \log N \tag{3-2}$$

It becomes a linear line in log-log scale where β represents the slope of the linear line. There can be three categories of β, namely super-linear with population (β>1) which means countries with larger population have larger value per capita indicator; linear with population (β=1) which means countries with different volume of population have the same value of per capita indicator; and sublinear with population (β<1) which means countries with larger population have smaller value per capita indicator[76]. I fit the data by using ordinary least squares (OLS) to find β.

In addition, the residual (ε) of the above regression model (i.e., scale-independent indicator) is used to re-rank the countries[25, 26].

$$\varepsilon = \log \frac{Y}{Y_0 N^{\beta}} \tag{3-3}$$

For the positive directional indicator like GDP, larger ε means better performance. For the negative directional indicator like CO_2 emissions, smaller ε means better performance.

3.3. Results

3.3.1. Empirical Scaling at Country Level

By analyzing data of 213 countries and regions (Table B-1) from 1995 to 2019[73-75], I found important development indicators scale with population universally (Figure 3-1 and Table B-2).

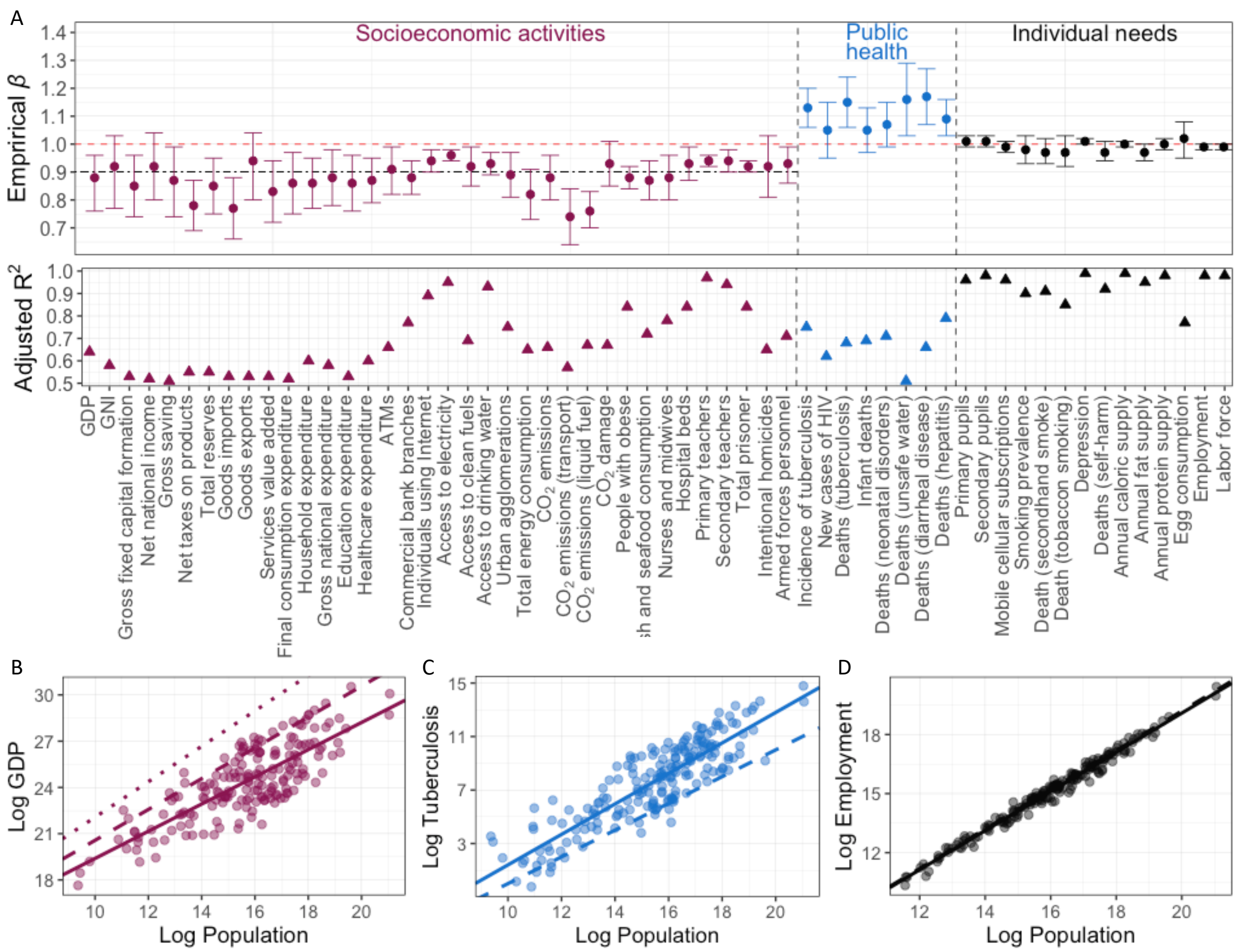

Figure 3-1. Empirical scaling results for indicators of socioeconomic activities, public health and individual needs. (A) Empirical scaling exponents with 90% confidence interval for indicators of socioeconomic activities, public health, and individual needs. Dot-dash line shows the approximate scaling exponent (β=0.9) for most socioeconomic activity indicators, and dash line shows the linear scaling (β=1). Examples of scaling relationships in countries for GDP (B), incidence from tuberculosis (C), and employment (D) in 2019; solid line shows the best-fit relation, dash line shows the linear scaling, and dotted line shows the scaling of the same indicator in cities.

Specifically, I find indicators of socioeconomic activities scale sub-linearly (β<1) with population, implying the growth rate of these indicators declines as population increases. On the one hand, some indicators represent socioeconomic welfare such as gross domestic product (GDP), net national income, access to healthcare, and access to safe drinking water and electricity; thus sub-linear scaling indicates compromised welfare for each individual with increased population. On the other hand, higher value of indicators such as CO_2 emissions, energy consumption, and number of prisoners are undesired; thus sub-linear scaling reflects higher per-capita efficiency with larger population. Such result is contrary to the scaling of similar indicators in cities which scale super-linearly with urban population (β≈1.15)[22-26, 68, 69, 76-78].

The results also show public health indicators scale super-linearly with population (β_p≈1.1), such as "infant death", "death from hepatitis", and "incidence of tuberculosis". This suggests that the performance of public health tends to decline for individuals with increased population in a country. The super-linear scaling of public health indicators can be explained by the sub-linear scaling of the socioeconomic activity indicators related to healthcare such as "healthcare expenditure", "number of nurses and midwives", and "number of hospital beds" (β_h≈0.9). Specifically, the number of death or disease Y_p=Nf where f is the per-capita death or disease in a country which is correlated to the inverse of per-capita access to healthcare ($Y_h/N = Y_{h0}N^{\beta_h}/N = Y_{h0}N^{\beta_h-1}$); thus $Y_p = Nf \sim N/(Y_h/N) \sim N^{2-\beta_h}$ where the exponent β_p=2–β_h≈1.1.

I also find indicators of individual needs scale linearly with population in countries, which has also been observed in cities[23, 24]. This indicates that, on average, individuals in different countries tend to have the same level of demand related to these indicators, regardless of the size of population. The fact that individual need indicators scale linearly in both countries and cities can be explained by that the terminal units of socioeconomic networks in both countries and cities are the same—individuals—and their size is invariant[76].

The exponents of most indicators are consistent across different years (Figure 3-2), indicating the scaling of these indicators could be the result of some fundamental mechanisms governing the socioeconomic dynamics of countries. However, exponents of a few indicators like “Mobile subscriptions” and “Internet users” continuously grow or decline. This is likely because the popularizing rate of these basic services would increase until 100% (i.e., same per capita across countries) as the economy grows, which also means exponents of these basic service-related indicators will become linear in the future.

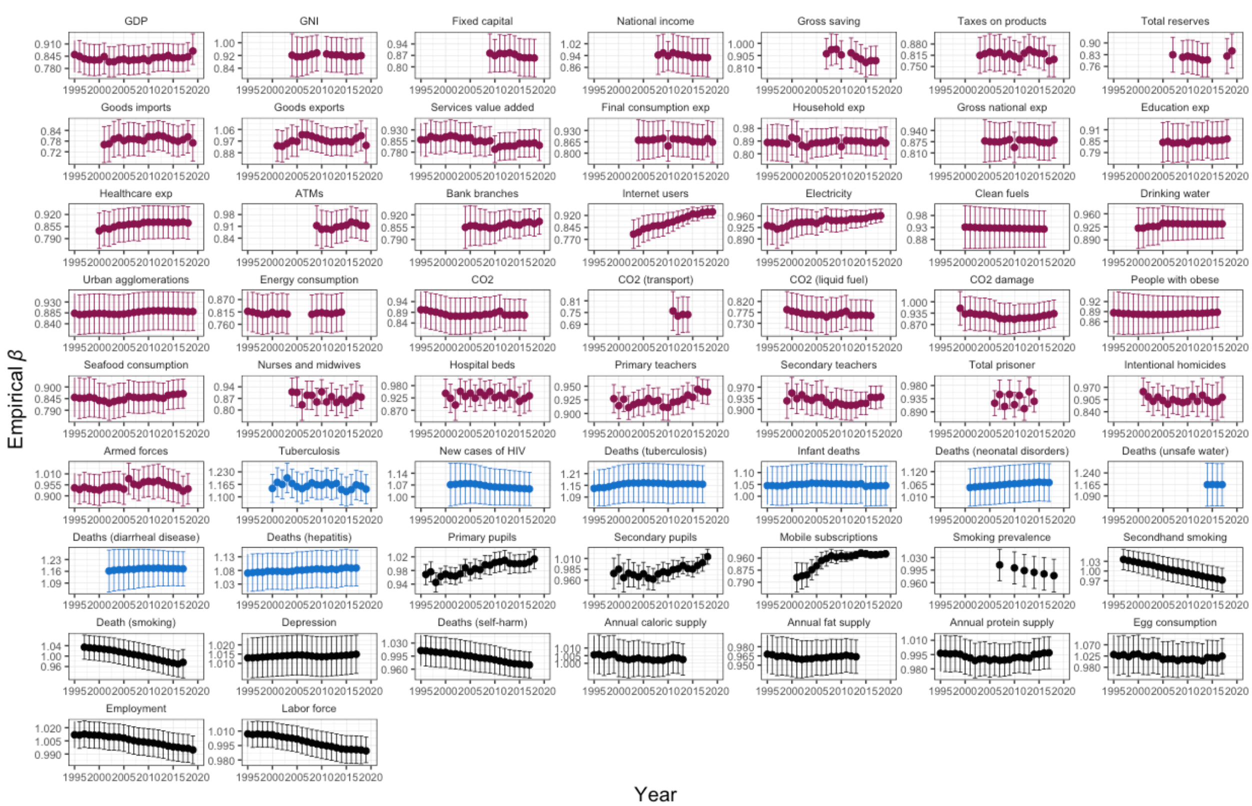

Figure 3-2. Annual scaling exponents for indicators of socioeconomic activities, public health and individual needs.

3.3.2. Developing a Theory of Scaling at the Country Level

Socioeconomic activity indicators scale differently in countries and cities: sub-linearly in countries and super-linearly in cities. This implies aggregation effects exist in cities from the concentration of population[23, 26], but do not exist in countries with population increase. It also indicates there are some intrinsic relations between the scaling in countries and that in cities. Given that a country is an ensemble of urban and rural areas, I propose a theory to explore the origins of scaling of indicators of socioeconomic activities and individual needs in countries based on the scaling laws observed in cities.

To understand the scaling exponent of development indicators among countries, I need to estimate the value of the development indicator for each country given its population size. A country is an ensemble of urban and rural areas; thus the total level of a development indicator of a given country is the sum of its total urban and rural parts which in turn are the sum of the levels of each urban and rural area, respectively. This can be written as $Y_{total} = Y_{urban} + Y_{rural} = Y_0 \sum_i N_{u,i}^{\beta_u} + \alpha Y_0 \sum_j N_{r,j}^{\beta_r}$, where Y_0 is the common economic base of all urban areas in the same country which differs across countries, $N_{u,i}$ is the population size of city i, and the scaling exponent β_u is from urban scaling literature ($\beta_u = 1.15$ for socioeconomic activity indicators and $\beta_u = 1$ for individual need indicators), αY_0 is the common economic base of all rural areas for the country assumed to be proportional to the urban economic base Y_0, α is the ratio between rural and urban economic base of this country and differs across countries, $N_{r,j}$ is the population size of village j. In addition, I assume no aggregation effect between the rural portion of Y (Y_{rural}) and rural population. This means $\beta_u = 1$ for any development indicator. The total value of a development indicator of a given country can be written as $Y_{total} = Y_{urban} + Y_{rural} = Y_0 \sum_i N_{u,i}^{\beta_u} + \alpha Y_0 \sum_j N_{r,j}^{\beta_r} = Y_0 \sum_i N_{u,i}^{\beta_u} + \alpha Y_0 N_r^1$, where N_r is the total rural population of the country[73]. Y_{urban} and Y_{rural} of a country can then be estimated once Y_0, the distribution of urban population (i.e., population for each urban area), and α are given.

To quantify my theory, the parameters need to be estimated. To estimate Y_0 of a given country, I need to have empirical data on the urban portion of Y_0 and urban population for all cities, and then conduct the regression to estimate the Y_0. However, such data are not available for all indicators in all countries. In addition, given the high explanation power (high R^2) for the urban population on urban development indicator[24, 68, 76], the true Y_0 of a given country should be close to Y_0 of each city in this country[24, 68, 76], especially for high-urbanized countries. Therefore, I can use pairwise empirical data of urban indicators and population of cities from as many countries as I can to estimate the range of Y_0.

I take GDP as an example of socioeconomic activity indicators given its relatively abundant data (Figure 3-3). I collect the pairwise empirical data on urban GDP and population of almost 900 cities from 150 countries and regions[79]. I calculate Y_0 for each city using $GDP_{urban}/N_{urban}^{1.15\pm\delta}$, where GDP_{urban} is the GDP of the city and N_{urban} is the population of the city. I add a parameter δ to consider the uncertainty of the estimated Y_0 of GDP. Previous studies show that the scaling exponent of the GDP in the urban system ranges around 1.1 and 1.2[24, 68], which helps us to set δ equals 0.05. I then estimate the range of Y_0 for each country using the minimal and maximum values of the Y_0 of its cities. This means the minimal value of Y_0 for each country is the minimal of Y_0 of its cities, so as to the maximum value.

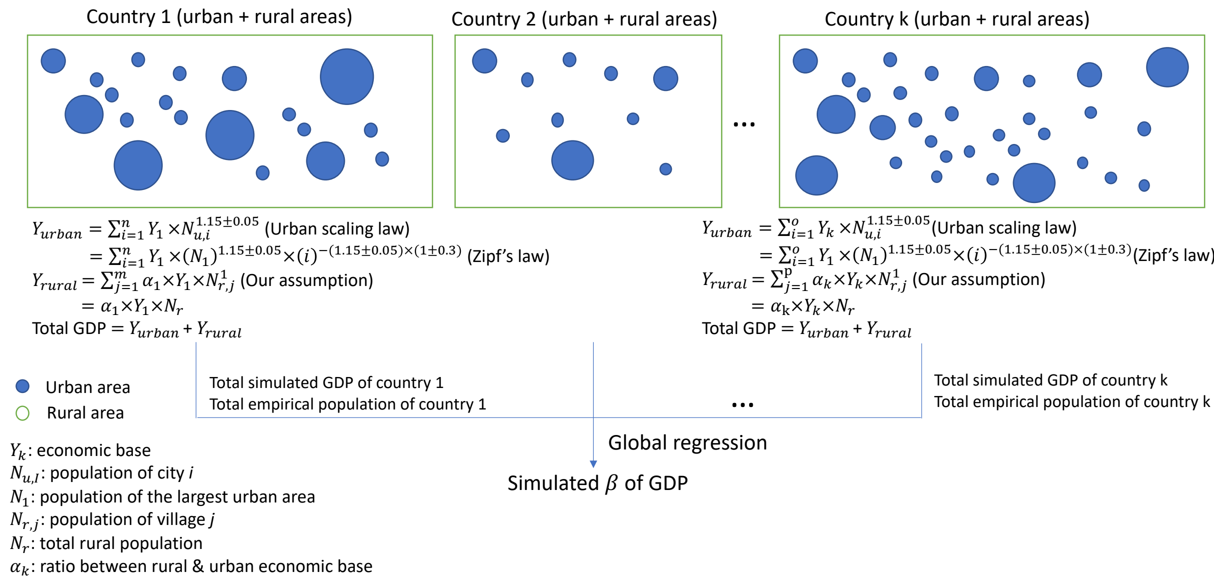

Figure 3-3. Basic framework for understanding scaling exponent of indicators of socioeconomic activity. Note that this is an example of GDP.

I use the Zipf's law to approximate the distribution of urban population (population size for each urban area) among cities in a country. Zipf's law implies that the city in any country with the largest population is generally twice as large as the next largest, and so on[79, 80]. This could be formularized as $N_{u,i} = N_1 i^{-1}$, where $N_{u,i}$ is the population of a city i, i is the rank of the population size of the city, and N_1 is the population size of the largest city. However, many empirical studies found that Zipf's exponent can vary around 1 depending on the country, the time period, the definition of cities used or the fitting method[79, 80]. Therefore, I extend the Zipf's law function as $N_{u,i} = N_1 i^{-(1\pm\zeta)}$ to consider the uncertainty of the approximated distribution of urban population. Previous empirical studies show that most of the Zipf's exponents vary around -0.7 to -1.3[81, 82], which help us to set ζ equals 0.3. I collect urban population data for almost 1,900 cities with more than 300,000 people of 150 countries and regions[83]. Given the total urban population and maximum urban population for each country are known, I only need to estimate the urban population less than 300,000 using the extended Zipf's law.

After quantifying the theory, I run simulations to test it. I vary α within the range [0.7, 1.3] to consider the uncertainty of the simulated rural GDP. For each simulation, I randomly and independently select a value for each parameter within the parameter interval for each country. After randomly simulating the GDP for each country, I conduct the regression to find the simulated β. I repeat the simulation process for 10,000 times. Note that the random selection of parameter means the selected value of a given parameter follows the uniform distribution including $Y_0 \sim U\big(min(Y_0), max(Y_0)\big)$, $\delta \sim U(-0.05, 0.05)$, $\zeta \sim U(-0.3,0.3)$, and $\alpha \sim U(0.7,1.3)$. In other words, each value within the parameter interval has the same probability to be selected. To consider the uncertainty of the simulation due to the distribution of the parameters, I also consider the parameters follow another widely observed distribution, normal distribution. This means $Y_0 \sim N\Big(\big(max(Y_0) - min(Y_0)\big)/2, \big(\big(max(Y_0) - min(Y_0)\big)/2 - min(Y_0)\ \big)/3\ \Big)$, $\delta \sim N(0,0.05/3)$, $\zeta \sim N(0,0.3/3)$, and $\alpha \sim N(1,0.3/3)$.

I also take employment as an example of individual need indicators. I collect pairwise empirical data on urban employment and population of almost 900 cities from 150 countries and regions[84]. I calculate Y_0 for each city using $Employment_{urban}/N_{urban}^{1\pm\delta}$, where $Employment_{urban}$ is the employment of the city and N_{urban} is the population of the city. Previous empirical studies show that the scaling exponent of the employment in the urban system ranges around 0.99 and 1.02[23]. Therefore, I add a parameter δ (δ = 0.02) to consider the uncertainty of the estimated Y_0 for employment. Similarly, I simulate the β of employment using the parameters under uniform and normal distributions, respectively.

The results show that the simulated samples derive similar scaling exponents from the empirical observations (Figure 3-4). Specifically, the median of the simulated β for socioeconomic activity indicators and individual need indicators are close to those from empirical data (0.90 vs. 0.88 and 1.00 vs. 0.99, respectively).

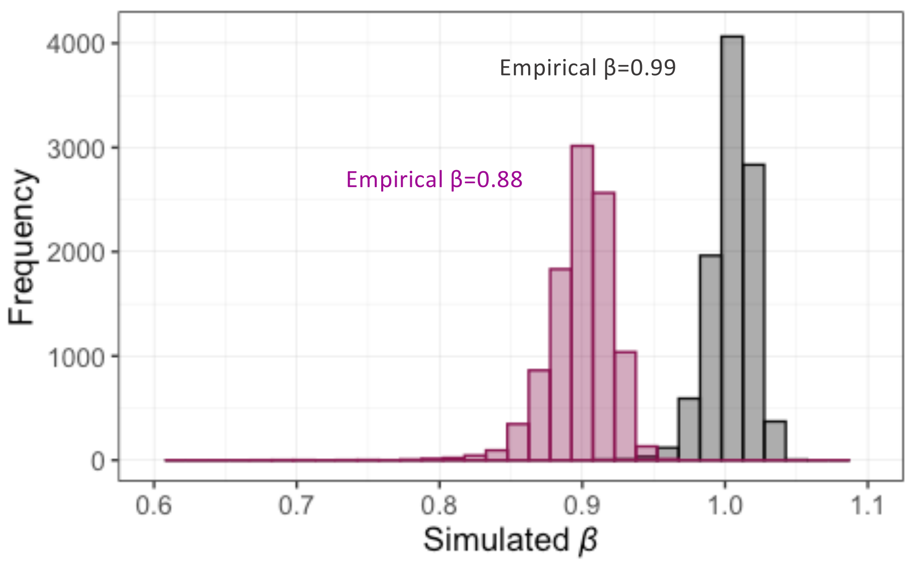

Figure 3-4. Histogram of β for indicators of socioeconomic activity and individual need indicators in countries from 10,000 simulations.
Median value of the simulated β for socioeconomic activity (red) and individual need indicators (grey) are 0.90 and 1.00, respectively. Distribution of parameters is normal. Uniform distribution generates similar results (Figure B-1).

3.3.3. Scaling Transition in Nations

Previous studies found that there are three kinds of population growth patterns based the scaling exponent[23, 76]. As shown in Figure 3 in ref[23], β<1 leads to a sigmoid population growth pattern, and population growth ceases in long term as it reaches a finite carry capacity. This is shown in the biological systems and companies where the organism ultimately dies[76, 85] and the company demises[86]. β=1 leads to an exponential population growth pattern. β>1 leads to a growth which is faster than exponential population growth and scaling diverges within a finite time and collapse due to the limited resource. This means cities are destined the eventually stop growing[23]. However, this collapse could be avoided by innovation and technology to reset the initial conditions (Figure 4 in ref[23]). In that case, a new cycle is initiated, and cities continue to grow. The reset process could be continually repeated and lead to multiple cycles, which therefore pushes the potential collapse into future. The side-effect of this reset is the time to collapse in the following cycle becomes shorter, which means major innovations must arise at an accelerated rate[23, 76].

The results show that countries are more like biological systems and companies rather than cities, in which development outputs grow sub-linearly with population. This indicates countries will eventually stop growing or even collapse. How can countries grow continuously, or is it even possible? Urbanization might be the answer, because, theoretically, cities grow super-linearly and their growth never stops.

To test this hypothesis, I examine 58 highly urbanized countries (urbanization rate in 2019>80%). Results show that the scaling exponents of most socioeconomic activity indicators increase from around 0.9 (sub-linear scaling, Figure B-2, Figure 3-5, Panel A) to close to 1 (linear scaling, Figure B-2, Figure 3-5, Panel A). The scaling of individual need indicators is still linear for these highly urbanized countries. In addition, the values of Adj-R^2 for most indicators are improved, indicating population can better explain the variations of these indicators when countries become more urbanized (Figure B-2, Figure 3-5, Panel B). I also simulate β for socioeconomic activity indicators and individual need indicators for these highly urbanized countries. Results show that the simulated β are very close to the empirical observations (Figure 3-5, Panel C), 1.02 vs. 0.99 for socioeconomic activity indicators and 0.97 vs. 0.99 for individual need indicators. These results indicate urbanization can potentially help countries grow with increased scaling exponents from sub-linear to linear. However, is it possible for countries to grow super-linearly?

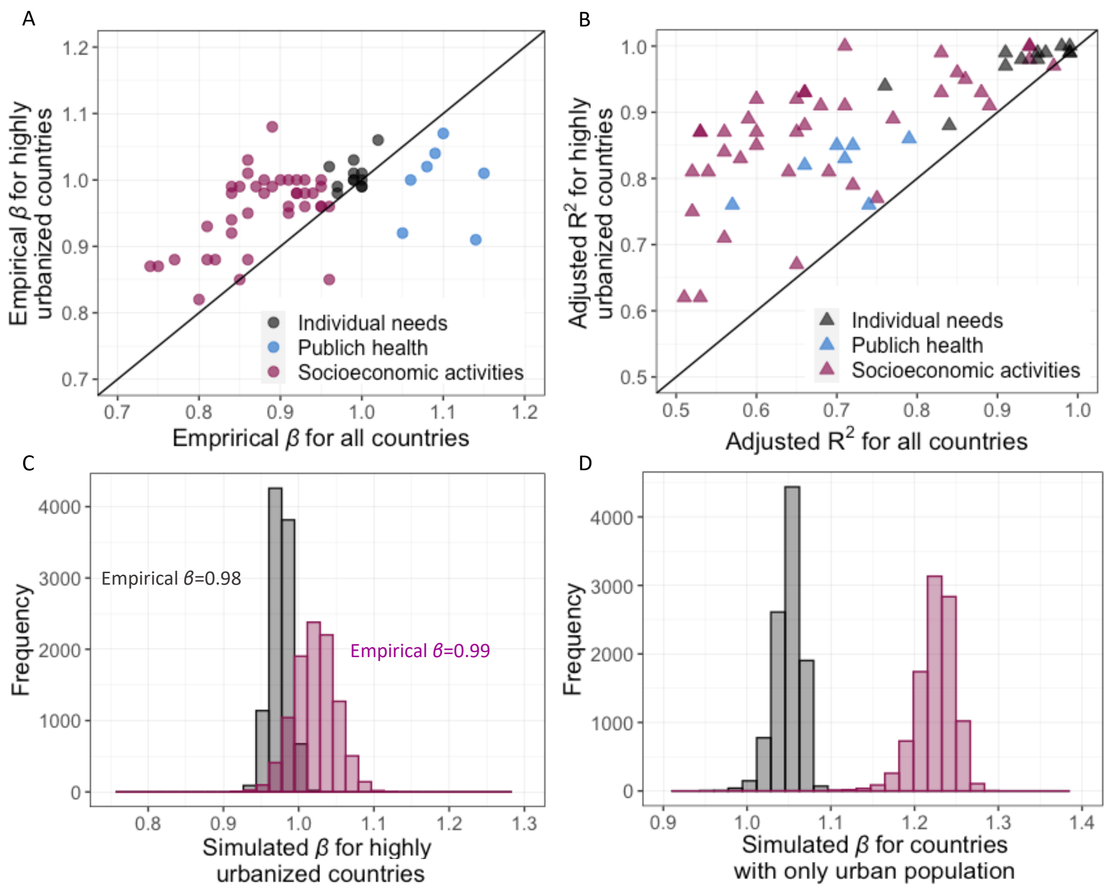

Figure 3-5. Histogram of β for indicators of socioeconomic activity and individual need indicators for simulation. Comparison of empirical scaling exponents (A) and adjusted R^2 (B) of socioeconomic activity indicators between highly urbanized countries and all countries. Histogram of simulated β for socioeconomic activity indicators (red) and individual need indicators (grey) for highly urbanized countries (C) and for all countries with only urban population (D) from 10,000 simulations. Distributions of parameters in (C) and (D) is normal. Uniform distribution generates similar results (Figure B-3, Panels A&B).

If each country is a city (e.g., Singapore), the scaling of countries will be super-linear, leading to open-ended growth. But what if each country is fully urbanized but with multiple cities? I simulate the scaling of the development indicators only considering each country's existing cities[84]. I find the scaling of socioeconomic activity indicators would become super-linear (β≈1.06) (Figure B-4, Panel A), while that of individual need indicators would remain linear (β≈1.00) (Figure B-4, Panel B). Alternatively, assuming each country only has its current urban population, I find that the simulated β for socioeconomic activity indicators would be around 1.2, indicating super-linear, while that of individual need indicators would still be around

1.00 (Figure 3-5, Panel D). This means urbanization can indeed lead to continuous growth for countries. It is generally accepted that urbanization promotes economic growth to some extent as it released the agricultural labor into industrial service based economy[87], and the aggregated population in cities increase the social interactions and balance benefits and costs in a way that leads to super-linear growth for socioeconomic properties[68]. These theories could also be supported by modern statistics[88]. However, super-linear growth comes with super-linear increases of undesired socioeconomic outcomes (e.g., crime and resource consumption)[23]. This calls for policy attentions to these accompanying, unavoidable undesired consequences of urbanization.

3.4. Discussion

Despite the great diversity and complexity of countries, the findings suggest national metrics follow common scaling relationships with population size. I have also shown that the scaling in countries is largely driven by the scaling in cities and super-linear growth in countries is largely due to urbanization. By viewing countries as a structure that include an ensemble of self-similar cities and rural areas, I found these systems are governed by universal mechanisms regardless of social, economic, political, cultural, and geographical variabilities. Such findings provide a quantitative and mechanistic understanding of national development. A critical implication for development immediately follows. Keeping other factors constant, if a country could concentrate people and resources in megacities while ensuring social cohesion and environmental sustainability, its development indicators have potential to significantly improve.

Here is an example of GDP. The total GDP of a given country is the sum of its total urban and rural GDP, which can be expressed as $Y = \sum_{i=1}^{n} Y_0 \times N_i^{1.15 \pm \delta} + \sum_{l=1}^{m} \alpha_0 \times Y_0 \times N_{rural}^{1}$. The first strategy is to continue the urbanization process to let the super-linear scaling effect existing in the urban system dominates national development. This is because $Y_{urban} > Y$ given $Y_{urban} = \sum_{i=1}^{n-1} Y_0 \times N_i^{1.15\pm\delta} + Y_0 \times (N_n + \Delta N)^{1.15\pm\delta} + \sum_{l=1}^{m} \alpha_0 \times Y_0 \times (N_{rural} - \Delta N)^1$, where ΔN is the size of rural population moving to the city with the least population (N_n). This also holds if the rural population (ΔN) moves to any city. The second strategy is to concentrate urban population given the constant urban population and rural population. I propose two specific ways to achieve it. First, the country can have fewer but larger cities. This is because $Y_{less} > Y$ given $Y_{less} = \sum_{i=1}^{n-2} Y_0 \times N_i^{1.15\pm\delta} + Y_0 \times (N_{n-1} + N_n)^{1.15\pm\delta} + \sum_{l=1}^{m} \alpha_0 \times Y_0 \times (N_{rural}^1)$, where N_n is the size of population of the smallest city. This also holds if any two cities merge as one. The extreme case is the country only has one city. Having fewer but larger cities might not be feasible for all countries. An equivalently effective approach is to better connect cities with better infrastructure such as high-speed rail and the Internet. Second, the country can encourage mega cities to concentrate its urban population. This because $Y_{mega} > Y$ given $Y_{mega} = \sum_{i=2}^{n-1} Y_0 \times N_i^{1.15\pm\delta} + Y_0 \times (N_1 + \Delta N)^{1.15\pm\delta} + Y_0 \times (N_n - \Delta N)^{1.15\pm\delta} + \sum_{l=1}^{m} \alpha_0 \times Y_0 \times (N_{rural}^1)$, where N_1 is the size of population of the largest city and ΔN is the size of population moving from the smallest city to the largest city. This also holds if ΔN is from a smaller city to a larger city. The extreme case is that the country has one mega city and the rest of the urban population is allocated in extremely small cities. A more practical scenario is to have multiple megacities to host the majority of urban population.

The practical implications of the findings highlight the importance of understanding the limitation and possibility of country growth. These scaling relationships predict many dimensions of development a country can expect with respect to population change and urbanization. Such predictions help policymakers set realistic targets for development policy and develop strategies to address unintended consequences. These findings also provide a quantitative argument against mainstream practice of comparing national development using per-capita measures[89-95], which assumes development indicators scale linearly with population[25, 96]. However, this assumption does not always hold, since it ignores the effect of agglomeration resulting from non-linear interactions in social dynamics. New rankings of nations based on deviations from the scaling laws provide new and more accurate comparison of the performance of national development (Methods, Supplementary Note B-1 and Table B-3).

3.5. Limitations and Future Research

This study assumes the linear scaling between the development indicators and population in rural areas due to lack of empirical data. For future research, I will collect empirical data to improve the quantitative framework explaining the origin of scaling in countries by considering the non-linear relationship between development indicators and population in rural areas.

The super-linear and sub-linear scaling exponents only represent the general pattern on average at the global scale. The deviation of various development indicators from the scaling is particularly important to understand how local characteristics play a role in national development. Future research can further explore the scaling relationship at the country level to provide a unique perspective on how socioeconomic dynamics shape the development of a country and its impacts on energy, resources, and the environment. This insight will help identify pathways of sustainability transition towards open-ended growth and continuous improvement of human living standards within the planetary boundary.

Chapter 4. Impact of COVID-19 on Sustainable Development Goals

4.1. Introduction

The global progress to achieve the UN SDGs by 2030 has been stalled by the COVID-19 pandemic. To date, COVID-19 has already caused over 145 million confirmed cases and 3.1 million deaths[97]. As a result of mitigation measures such as lockdown, COVID-19 has also greatly affected the global economy. The world's GDP is projected to decline by 4.4% in 2020, almost three times worse than that in the Great Recession (-1.6% in 2008)[98]. Consequentially, financial and institutional resources that would be available to enhance SDGs will likely go away by a large extent. Achieving SDGs by 2030 post COVID-19 becomes more challenging if not impossible.

A few studies have assessed the impacts of COVID-19 on SDGs[28-34]. However, all these studies focus on qualitative assessment of COVID-19 impacts on SDGs or SDG targets. Without a quantitative evaluation, it is difficult to understand the different impacts on specific SDGs, SDG targets, and SDG indicators for developed and developing economies. Such an understanding is urgently important for government and non-state stakeholders to identify critical areas for targeted policy to resume and speed up the progress to achieve SDGs by 2030.

To fill this knowledge gap, I predicted the quantitative impacts of COVID-19 on SDGs at the indicator level using machine learning. The prediction is based on the expected changes in GDP and population, because both historical data and future projections related to GDP and population are widely available for developing models and the success of SDGs highly depends on economic growth. The model can predict 42 SDG indicators in 31 targets and 16 SDGs with reasonable accuracy (Methods and Table C-1). Other indicators are thus excluded due to either lack of data or low prediction accuracy (testing $R^2 < 0.6$) including all indicators in SDG 5 (Gender Equality). As a result, my analysis focuses on these 43 SDG indicators which are most relevant to GDP and population. Specifically, I addressed two research questions. First, what are the global impacts of COVID-19 on each SDG? Second, how do the impacts differ between emerging market and developing economy (EMDE) and advanced economy (AD)?

To answer these questions, I first used historical data to develop and test a variety of supervised machine learning models with cross-validation to predict each SDG indicator (response) based on four predictors (population, GDP, annual GDP growth rate, and time). I then predicted each SDG indicator between 2020 and 2024 using the best model and projected GDP and population. To reflect the impact of COVID-19, I used four sets of GDP projection data to represent one no-COVID-19 scenario and three post-COVID-19 scenarios. Specifically, the International Monetary Fund (IMF) released two GDP projections in October 2019 and October 2020[98, 99] which I used to represent the no-COVID-19 scenario and a COVID-19 (S1) scenario, respectively. Specifically, the COVID-19 (S1) scenario is very optimistic that the GDP will quickly recover to pre-COVID-19 trajectory in 2021 with the global GDP growth rate of 5.2%. Given the continuation of the COVID-19 pandemic worldwide, mitigation measures affecting the economy are likely to be continued at least until 2022[100]. Therefore, I also examined two less optimistic COVID-19 scenarios in which the GDP recovers to the pre-COVID-19 trajectory in 2022 (COVID-19 (S2)) and 2023 (COVID-19 (S3)), respectively. Note that the GDP projections of the three COVID-19 scenarios in 2020 are the same. As the uncertainty of longer GDP projection becomes increasingly higher, I did not predict the SDG indicators beyond 2024. Next, I normalized and aggregated the predicted SDG indicators into SDG performance. Specifically, the SDG performance is a metric based on multiple SDG indicators to represent the overall performance towards achieving each SDG. A higher value is more desired indicating closer to achieving SDG (see details in the Methods). I quantified the impact of COVID-19 using the predicted SDG performances and indicators in the no-COVID-19 and the COVID-19 scenarios in the same year. In other words, I exclusively focused on how the SDGs would be with COVID-19 as compared to how they would be with COVID-19 during 2020-2024, rather than how the SDGs will change from 2019.

4.2. Data and Method

4.2.1. Indicator Selection and Data Sources

I proposed three criteria to select predictors including 1) the availability of both prediction and historical data; 2) the association with global sustainable development; 3) low correlation among predictors. The population- and economy-related indicators meet both the first two criteria[23, 101-103]. For the population-related indicators, I selected the "Total population", "Urban population", "Female population", "Male population", "Population ages 0-14", "Population ages 15-64, "Population above 65" and "Annual population growth rate (%)"as candidates. For the economy-related indicators, I selected "GDP (current US$)", "GDP (constant 2010 US$)", "Annual GDP growth rate (%)", "GDP per capita (constant 2010 US$/capita)", and "GDP per capita (current 2020 US$/capita)" as candidates. Figure 4-1 shows the Pearson correlation matrix among these candidate predictors. Note that population-related indicators are highly correlated with each other with an average 0.78 Pearson correlation coefficient. This means I can only use one indicator to represent their total information (variation). I selected the indicator "Total population" as it is the most comprehensive one. The economy-related indicator

“GDP (constant 2010 US$)” is highly correlated with “GDP (current US$)”. I selected “GDP (constant 2010)” as it is inflation-adjusted and measures the real change in GDP [104]. I also selected “GDP growth rate” as a predictor. This indicator is also highly correlated (0.61 Pearson correlation coefficient) with “Annual population growth rate (%)”. I discarded “GDP per capita (constant 2020 US$/capita)” and “GDP per capita (current 2020 US$/capita)” as they are the linear combination of the “GDP” and “Total population”. In addition, I also incorporated “Time (measured by year)” to capture the potential variation associated with time.

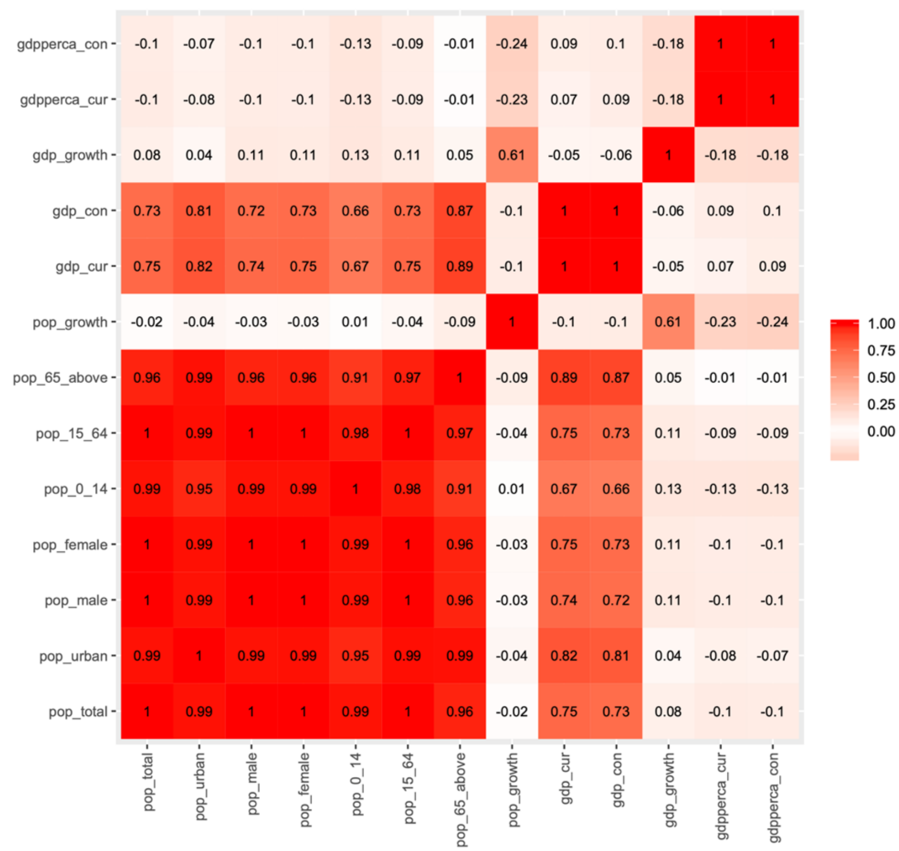

Figure 4-1. Pearson correlation matrix of candidate predictors.

I selected candidate indicators from datasets provided by the UN[105], World Bank[106], and the 2020 SDG Index and Dashboards Report[30]. The 2020 SDG Index and Dashboards Report were published by the Sustainable Development Solutions Network which operates under the

UN auspices to promote the implementation of the SDGs and the Paris Climate Agreement. There are in total 42 SDG indicators in my dataset covering 16 SDGs and 31 SDG targets for 213 countries and regions. The temporal coverage of individual SDG indicators varies in the dataset, with the longest from 1990 to 2019. The historical data of all the predictors are from the World Bank[107]. The projected data of the predictor "GDP growth (%)" and "GDP (constant price)" under the COVID-19 (S1) scenario are from the newest IMF World Economic Outlook database (released in October 2020)[98]. I also considered two less optimistic scenarios in which GDP recovers to the 2019 level in 2022 and 2023, respectively (Figure 4-2). The hypothetical projected GDP data under the no-COVID-19 scenario are from the same source released in October 2019 before COVID-19[99]. The projected data of the predictor "Total population" are from the UN's World Population Prospects database in 2019[108]. I collected the projected data for 187 countries and regions (Table C-2). The classifications of EMDE (149 countries and regions) and AE (38 countries and regions)are from IMF[98]. I predicted the annual value of each SDG indicator from 2020 to 2024 based on the available data for these predictors.

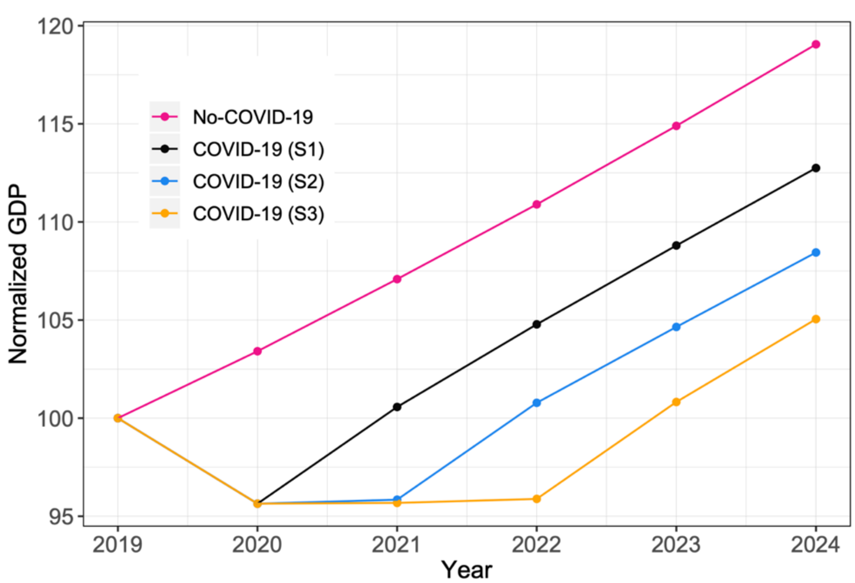

Figure 4-2. Normalized GDP under the no-COVID-19 and three COVID-19 scenarios.

Note that the no-COVID-19 scenario and COVID-19 (S1) scenario is projected by the IMF[98, 99]. I further considered two less optimistic COVID-19 scenarios in which the GDP recovers to pre-COVID-19 trajectories in 2022 (COVID-19 (S2)) and 2023 (COVID-19 (S3)), respectively.

4.2.2. Machine Learning Models for Prediction

Compared with the traditional methods, machine learning approaches can generally estimate complex non-linear relationship between response and predictors and show better prediction accuracy[38-42]. I developed and tested three types of widely used machine learning models, including support vector machine (e1071 package[109] in R), random forests (randomForest package[110] in R) and extreme gradient boosting (xgboost package[111] in R), to model the historical relationship between the four predictors (GDP, GDP growth rate, Total population, and Time) and the response (each SDG indicator). For each response, I selected the best model (with the highest R^2 on test sets) to predict the future. Specifically, I split the entire dataset by years into a training set and several test sets. The number of test sets is based on the last available year of the SDG indicator. For example, if the last year of an indicator is 2018, the last six years are the period of the test set with data in each year as a separate test set. The rest of the data as a whole are the training set. For the model training, 3-fold cross-validation is used to optimize the hypermeters and avoid overfitting. Importance of the predictors can be found in the Figure C-2. I used the coefficient of determination (R^2) to evaluate the prediction accuracy. I use 60% explained variance as the criterion for model selection (i.e., $R^2 >= 0.6$ on each test set) for each SDG indicator (Figure C-1). This means the major variation (>= 60%) of a specific indicator can be captured in the model, but the predicted value maybe not as reliable for individual countries (see an example in Figure C-3). Therefore, I only focus on country groups (AE, EMDE, and global) for the analysis, rather than focusing on individual countries. For the prediction, I re-trained the best model with the entire data set for each SDG indicator. I also used bootstrap sampling to reduce uncertainty which is a robust method to calculate confidence intervals for machine learning algorithms[112]. I calculated the confidence intervals of the prediction results by bootstrap resampling the training set for 100 times and filtered out the 5% quantile, 50% quantile (median value), and 95% quantile prediction values. I focused on the median value in the discussion as it will happen with the highest probability.

4.2.3. Normalization and Aggregation

To ensure comparability across different SDGs, the predicted indicator values for each SDG were normalized. I proposed a simpler normalization method rather than using the min-max normalization method[49, 113] for two reasons. First, the purpose of the min-max method is to compare the progress of SDGs among many countries across years with a maximum value of 100. However, the main goal of the research is to analyze the effect of COVID-19 at the global level and country groups level, which means the performance of an SDG indicator in 2019 should be the base (i.e., SDG performance = 100). Second, for the min-max method, I need to first select the lower and upper bound, which are usually set by the 2.5th quantile or top five performers[49, 113, 114]. This is impractical for us because I only focused on five years for the prediction (2020-2024). The simpler normalization method is represented using the following formulas: $SDG\ indicator\ performance =$

$$\begin{cases} \frac{x}{x_{2019}} \times 100\ for\ postive\ directional\ indicator\ \ (e.g., GDP\ per\ capita) \\ \frac{x_{2019}}{x} \times 100\ for\ negative\ directionalindicator\ (e.g., GHG\ emissions) \end{cases} \tag{4-1}$$

where $SDG\ indicator\ performance$ represents the normalized performance for a given SDG indicator, x is the value of a given SDG indicator before normalization, x_{2019} stands for the value of the indicator in 2019. "Positive directional indicator" means larger value corresponds to desired performance (e.g., GDP per capita), while "negative directional indicator" means the opposite (e.g., GHG emissions). The direction of the indicator is shown in Table C-3. Note that this normalization method cannot be directly applied to indicators with negative value such as "GDP growth (%)"as it will mislead the performance for the following two reasons. The indicator "GDP growth (%)" is not feasible for the proposed simple normalization method for the two reasons. First, there will be negative values which mislead the direction of the SDG indicator performance in two cases. For example, the value of "GDP growth (%)" is 2.4% in 2019 and -4.4% in 2020, which would lead to the normalized performance in 2020 of -183 using the normalization method. Another case is that the value of "GDP growth (%)" is 0.4% in 2019 and -4.4% in 2020, which means the normalized performance in 2020 would be -1,100. The latter case is obviously better than the former, but the normalized SDG indicator performance shows the opposite (-1,100 worse than -183). Second, the high variation of "GDP growth (%)" will mislead the performance of SDG 8. The value of "GDP growth (%)" decreases from 2.4% in 2019 to -4.4% in 2020 and back to 5.6% in 2021 under the COVID-19 scenario. This means the normalized SDG indicator performance would be -183 in 2020 and then back to 233 in 2021 (Figure C-4). The high variation will dominate the performance of SDG 8 and dilute the impact of other indicators, as shown in Figure C-4 that the performance of SDG 8 will decline by 61% in 2020 under the COVID-19 scenario and then become even higher than that under the no-COVID-19 scenario in 2021. Therefore, I proposed a piecewise function to re-normalize the indicator "GDP growth (%)". I assigned 0 value for the negative growth rate, and cut the change of GDP performance by 2/3 for the positive growth rate (Figure C-4). For example, if "GDP growth (%)" decreases from 2.4% in 2019 to -4.4% in 2020 and increases back to 5.6% in 2021, the re-normalized value will be 0 in 2020 and 144 (100 + ((5.6% / 2.4%) - 100) / 3 = 144) in 2021 (Figure C-4). The re-normalization will not change the trend of "GDP growth (%)", but helps show the effect of other indicators in SDG 8 (Figure C-4). I also tried other ratios like 3/4 which yielded similar results. For these cases, I used a piecewise function for normalization (Figure C-4). After normalizing all SDG indicators, I aggregated all the performances of related indicators using the arithmetic mean to yield the performances for specific SDGs[49, 113]. Then I aggregated all SDG performances using the arithmetic mean to yield an overall performance[49, 113].

4.3. Results

4.3.1. Global Impact in 2020

I found that a 7.7% decline of the overall SDG performance is expected in 2020 compared to no-COVID-19 scenario in the same year, (i.e., the difference of the SDG performance in 2020 in two scenarios compared to the SDG performance in 2020 in the no-COVID-19 scenario) (Figure 4-3). At the SDG level, the performances of 12 socioeconomic-related SDGs are expected to decline by 3.0-22.3 % in 2020, while those of 4 environment-related SDGs will increase by 1.6-9.2%.

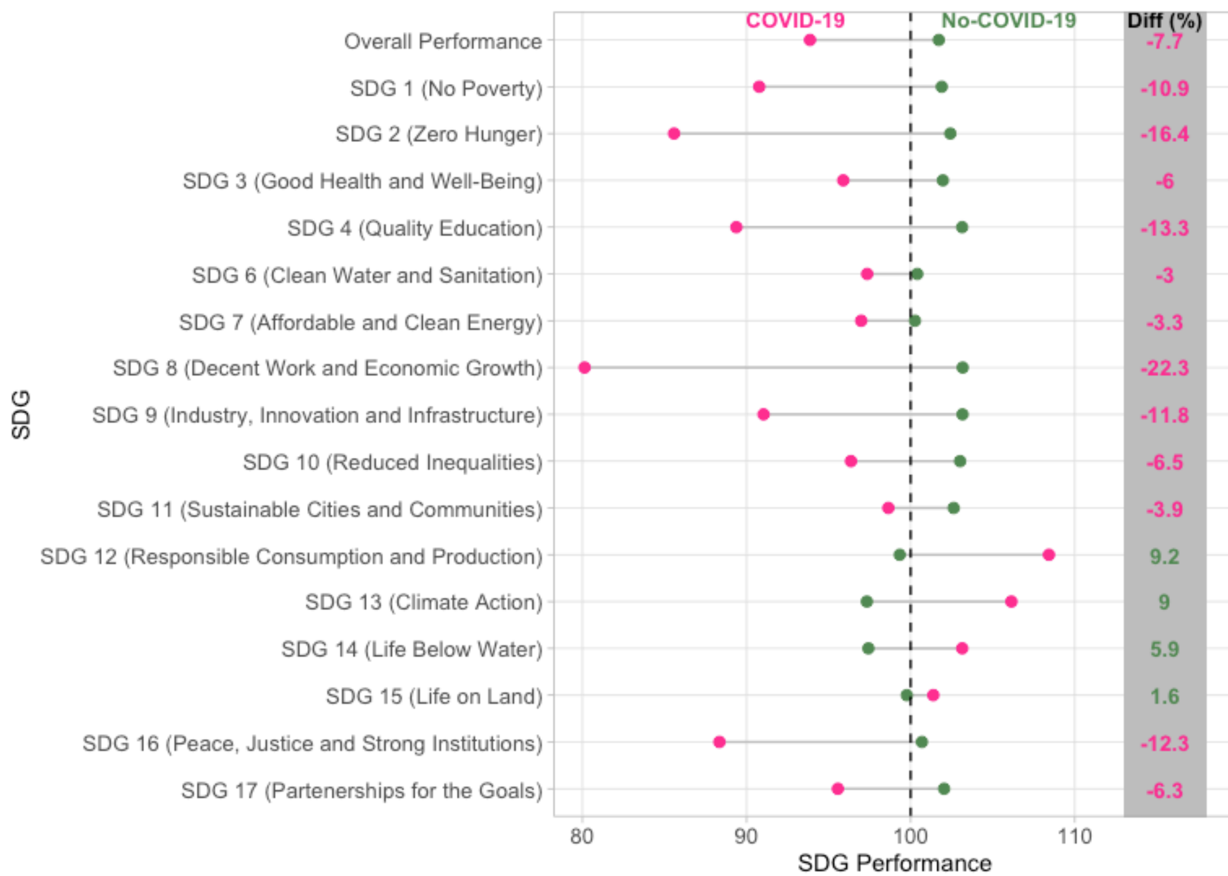

Figure 4-3. Comparison of SDG performances in 2020 under the COVID-19 and no-COVID-19 scenarios. "Diff (%)" denotes the percentage change of the SDG performance in 2020 in the COVID-19 scenario as compared to that in the no-COVID-19 scenario, representing the impact of COVID-19 on the SDG in 2020. SDG performances are normalized based on those in 2019 (SDG performance = 100 in 2019). Note that SDG 5 (Gender Equality) is excluded as none of its indicators can be predicted with reasonable accuracy ($R^2 < 0.6$). Note that the projections of the predictors in 2020 are the same under three COVID-19 scenarios.

The SDGs with declining performances in 2020 due to COVID-19 all highly depend on economic development. Among them, SDG 8 (Decent Work and Economic Growth) will suffer the greatest decline (-22.3%) in 2020. All its six indicators would decline (Figure C-5) with the largest for, not surprisingly, the indicator "GDP growth (%)" (-100%) (Figure C-6). The second largest predicted decline is for SDG 2 (Zero Hunger) with 16.4% decrease in its performance. Specifically, the indicator "Number of people with undernourishment" in 2020 is predicted to increase from 0.79 billion to 0.95 billion due to COVID-19 (Figure C-6). The latest UN Sustainable Development Goals Report predicts that small-scale producers are hit hard by the pandemic[28]. The performance of SDG 4 (Quality Education) will decrease by 13.3% as the third largest decline. More than 8 million primary children are predicted to be out of school due to COVID-19 in 2020, making its indicator “Number of primary children out of school” up to around 60 million in 2020. This is largely due to remote learning remains out of reach for many students especially those in developing countries[28]. For SDG 16 (Peace, Justice and Strong Institutions), the next largest declining SDG (-12.3%), “Corruption perception index (worst 0-100 best)” will decrease from 45.4 in the no-COVID-19 scenario to 39.8 in the COVID-19 scenario. This is reflected by studies such as Gallego et al. which found increased corruption due to relaxed public procurement rules and procedures in many places to expedite transactions for pandemic mitigation [115]. SDG 9 performance will decline by 11.8% (Industry, Innovation and Infrastructure). Notably, the indicator “Air transport, passengers carried” will decrease from 4.8 billion without COVID-19 to 3.0 billion with COVID-19 (Figure C-6), which is widely expected and observed due to travel restrictions during the pandemic[116]. For SDG 1 (No Poverty, -10.9%), the prediction shows about 200 million additional people will be “living less than $3.20 a day” due to COVID-19 in 2020. The UN also expects that COVID-19 will cause the first increase in extreme poverty in decades with 71 million people being dragged back into extreme poverty (less than $1.25 per day)[28].

While the SDGs depending on economic development are projected to suffer from COVID-19, other SDGs that are more relevant to the environment will actually be improved in 2020 during the pandemic. Specifically, the performances of SDG 12 (Responsible Consumption and Production), SDG 13 (Climate Action), SDG 14 (Life Below Water), and SDG 15 (Life on Land) will increase by 9.2%, 9.0%, 5.9%, and 1.6%, respectively, in 2020 in the COVID-19 scenario compared to the no-COVID-19 scenario. The prediction shows the per capita impacts on aquatic and terrestrial ecosystems will decrease by 9.2%, 5.9% and 1.6% due to COVID-19, respectively, approximated by the predicted changes of the SDG 12 indicator “Forest rents ($/capita)”, SDG 14 indicator “Fisheries production (kg/capita)”, and SDG 15 indicator “Forest area as a proportion of total land area (%)” (Figure C-6). This is also reflected in Sachs et al. which considered economic decline induced by COVID-19 will cause a short-term reduction in threats to the ecosystem and consumption of natural resources[30]. For SDG 13, the indicator “Energy-related carbon emissions (kg/capita)” will decline from 4.9 kg/capita in the no-COVID-19 scenario to 4.5 kg/capita in the COVID-19 scenario based on the projection. This is equivalent to an annual reduction of 5.9% in global carbon dioxide (CO_2) emissions in 2020 with COVID-19 from the 2019 level. Similarly, Liu et al. estimated the global CO_2 emissions declined by 8.8% in the first half of 2020[117], and their follow-up estimates indicate a 5.5% reduction in 2020 until October 31 compared to the same period in 2019[118]. The UN also predicted that COVID-19 will result a 6.0% drop in greenhouse gas (GHG) emissions for 2020[28].

4.3.2. Global Impact by 2024

Figure 4-4 (Panels A&B) show the impact of COVID-19 by 2024 on SDGs. In particular, the difference of the overall SDG performance in 2021 between the COVID-19 (S1) and no-COVID-19 scenarios is only 2.5, down from 7.8 in 2020, indicating in 2021 SDGs are closer to what they would be without COVID-19 than they are in 2020. This is due to the optimistic projection of over 5% annual GDP growth in 2021 by IMF[98]. However, in COVID-19 (S2) and (S3) scenarios in which global GDP stagnates in 2021 (pandemic continues in 2021), the difference of the 2021 overall SDG performance compared to that in the no-COVID-19 scenario are 6.4 and 6.5, respectively. Prolonged pandemic slows down the economic recovery and thus slows down the global SDG progress.

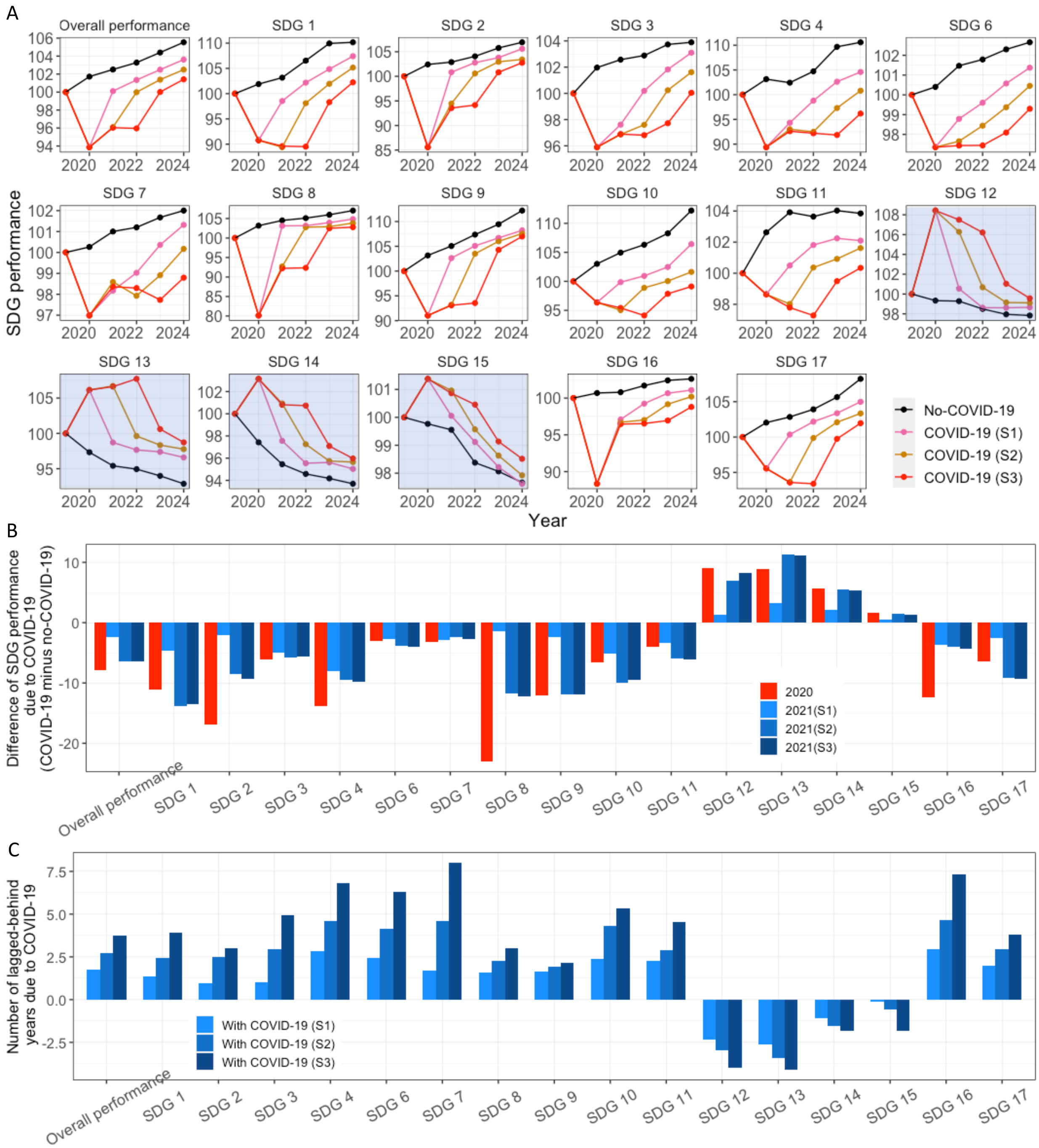

Figure 4-4. Global impact of COVID-19 on SDGs by 2024.
(A) Comparison of SDG performances between the no-COVID-19 and three COVID-19 scenarios from 2020 to 2024. Four environment-related SDGs with declining performances in the no-COVID-19 scenario are differentiated with different background colors. (B) Difference of SDG performances between the no-COVID-19 and each of the three COVID-19 scenarios in 2020 and 2021. Note that the projections of predictors are the same in 2020 under the three COVID-19 scenarios. (C) Number of years lagging behind the original trajectory for each SDG by 2024 due to COVID-19 under the three COVID-19 scenarios.

Among the 12 socioeconomic-related SDGs whose performances declined in 2020 due to COVID-19, in general, quicker GDP recovery will lead to quicker SDG performance recovery (Figure 4-4). For example, the differences of all the 12 SDGs in 2021 between the COVID-19 (S1) and no-COVID-19 scenario will be smaller than those in 2020. None of the 12 SDG performances will be able to reach the level they would be without COVID-19 in 2021 in all three COVID-19 scenarios. Among the four environment-related SDGs the performances of which increased in 2020 due to COVID-19, quicker GDP recovery will lead to quicker SDG performance decline. For example, the performances of the four SDGs in 2021 will be very close to their 2019 levels under the COVID-19 (S1) scenario, but will be still higher than their 2019 levels under the COVID-19 (S2) and COVID-19 (S3) scenarios.

Figure 4-4 (Panel C) shows how long COVID-19 will make each SDG lag behind its original trajectory without COVID-19 until 2024, defined as the difference of SDG performances in 2024 with and without COVID-19 divided by the average annual change of the SDG performance between 2019 and 2024 without COVID-19. This measure indicates the time (in years) it would take for each SDG to come back to its original progress without COVID-19. Overall, global SDG progress will lag behind the original trajectory by 1.9 to 4.1 years in the three COVID-19 scenarios, roughly equivalent to delay of achieving SDGs for 1.9 to 4.1 years due to COVID-19. For individual SDGs, although SDG 2 (Zero Hunger), SDG 8 (Decent Work and Economic Growth) and SDG 9 (Industry, Innovation and Infrastructure) will be greatly affected by COVID-19 in 2020 (16.4%, 22.3% and 11.8 declines), they will recover relatively quickly compared to their original trajectories without COVID-19, making them three of the least lagged SDGs due to COVID-19 by 2024 (about 1.0 to 3.0 year). In contrast, SDG 7 (Affordable and Clean Energy) will decline only by 3.3% in 2020 due to COVID-19, but it lags behind its original trajectory for approximately 1.8 to 8.0 years by 2024 as one of the most lagged SDGs. This is because the relative slow increment in performance of SDG 7 (with the annual increment of 0.4). This could also explain the relative long lags in SDG 6. Note that the pandemic will also slow down the process of environmental deterioration and gain us more time (0.9-4.1 years) to stabilize and reverse the originally declining trajectories of SDGs 12, 13, 14 and 15. Figure 4-4 (Panel C) also shows that the progresses of 12 socioeconomic-related SDGs will be further lagged-behind due to the slower GDP recovery, and the worsening of four environment-related SDGs (12, 13, 14 and 15) will be further slowed due to the slower GDP recovery.

4.3.3. Different Impacts for EMDE and AE countries

When the model is tested for EMDE and AE countries separately, fewer SDG indicators can be predicted with reasonable accuracy ($R^2 >= 0.6$): 27 indicators in 14 SDGs for EMDE and 18 indicators in only 8 SDGs for AE (Figure C-7&C-8). This is largely because of smaller sample size in split datasets for the two country groups. Therefore, I only compared the impacts of COVID-19 in EMDE and AE countries on the performance of individual SDG indicators (2019 = 100).

The results show COVID-19 will have severe negative impacts on SDG indicator performances for both EMDE and AE countries in 2020, with EMDE countries hit harder (Figure 4-5). Specifically, the median declines of individual SDG indicator performances in 2020 due to COVID-19 are -6.3% and -5.1% for EMDE and AE, respectively. This indicates EMDE countries are more vulnerable to economic downturn in sustainable development. The indicator "GDP growth (%)" in SDG 8 will decline the most for both EMDE (4.5% no-COVID-19 vs. -3.2% COVID-19) and AE (1.7% no-COVID-19 vs. -5.8% COVID-19) in 2020 among all the predicted SDG indicators. The other indicator that declines the most for both EMDE and AD is "Air transport, passengers carried (billion people)" in SDG 9, from 2.5 billion for EMDE and 2.2 billion for AE without COVID-19 to 1.5 billion with COVID-19 in 2020, respectively. The indicator "Undernourishment (%)"in SDG 2 will increase from 2.8% to 5.6% due to COVID-19 in 2020 for AE, making its performance declining by 50.0%, while the decline of the performance of the same indicator in EMDE is only 14.3%. However, the percentage of population undernourished in EMDE (14.1%) is still much higher than that in AE (5.6%) in the COVID-19 scenario in 2020. On the other hand, the performances of environment-related SDG indicators increase for both EMDE and AE in 2020. In particular, the performance of indicator "Forest rents ($/capita)" in SDG 12 has the largest increases for both EMDE and AE (18.9% and 22.7%, respectively), indicating lessened impact on terrestrial ecosystems in both country groups.

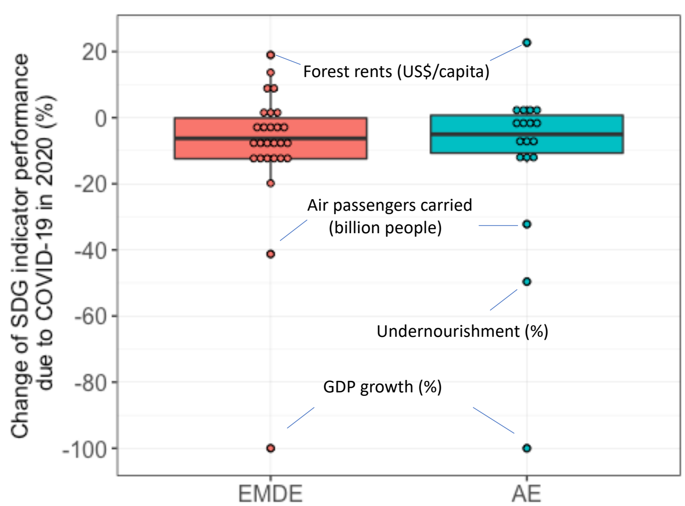

Figure 4-5. Impacts of COVID-19 on SDG indicator performance for EMDE and AE countries in 2020.

As shown in Figure 4-6 (Panels A-C), by 2024, the median changes of SDG indicator performances compared to the no-COVID-19 scenario are -2.3% to -5.5% and -1.5% to -2.8% for EMDE and AE, respectively. The largest decline for AE will be the performances of the indicators “Exports of goods and service ($/capita)” (-7.4% to -7.5%) and “Triadic patent (per thousand people)” (-5.9% to -11.8%) in 2024 due to COVID-19. For EMDE, the performances of the indicators “Manufacturing ($/capita)” and “Labour ($/capita)” will decline the most (-11.4% to -18.8% and -7.5% to -20.4%) in 2024 due to COVID-19. These results represent long-lasting impacts of COVID-19 on the global production and consumption system. Indicator “GDP growth (%)” will increase the most (7.1% to 28.1%) for AE in 2024. The largest increase for EMDE will be the performance of the indicator “Energy-related carbon emissions (kg/capita)” in 2024. While economic recovery is welcome, a strong “rebound” of GHG emissions is worrisome.

Figure 4-6 (Panels D-F) shows the number of years each SDG indicator lags behind its original trajectory without COVID-19 by 2024 for EMDE and AE. Because AE countries generally have smaller declines across all SDGs, they actually will be closer to their original trajectories by 2024 compared to EMDE countries. This is counterintuitive as the EMDE countries are predicted to own the faster post-COVID-19 economic recovery by IMF (S1). Specifically, IMF predicted that average GDP per capita of AE countries will recover to the 2019 level by 2023, but EMDE countries will be back to the same level two years earlier by 2021. The faster post-COVID-19 economic recovery for EMDE countries compared with AE countries will still remain under other two COVID-19 scenarios (S2 and S3). This may show the better resilience of the AE countries on the pandemic, which highlights the importance of sustainable development. The slower economic recovery for AE countries also explains additional time gained for SDG indicators such as “Energy-related carbon emissions (kg/capita)” with nearly 4 to 5 years. Note that the indicator “Suicide mortality rate (%)” will be lagged most for EMDE countries under COVID-19 scenarios (S2 and S3), which is due to the originally slow progress in the no-COVID-19 scenario (annual increment of 0.4).

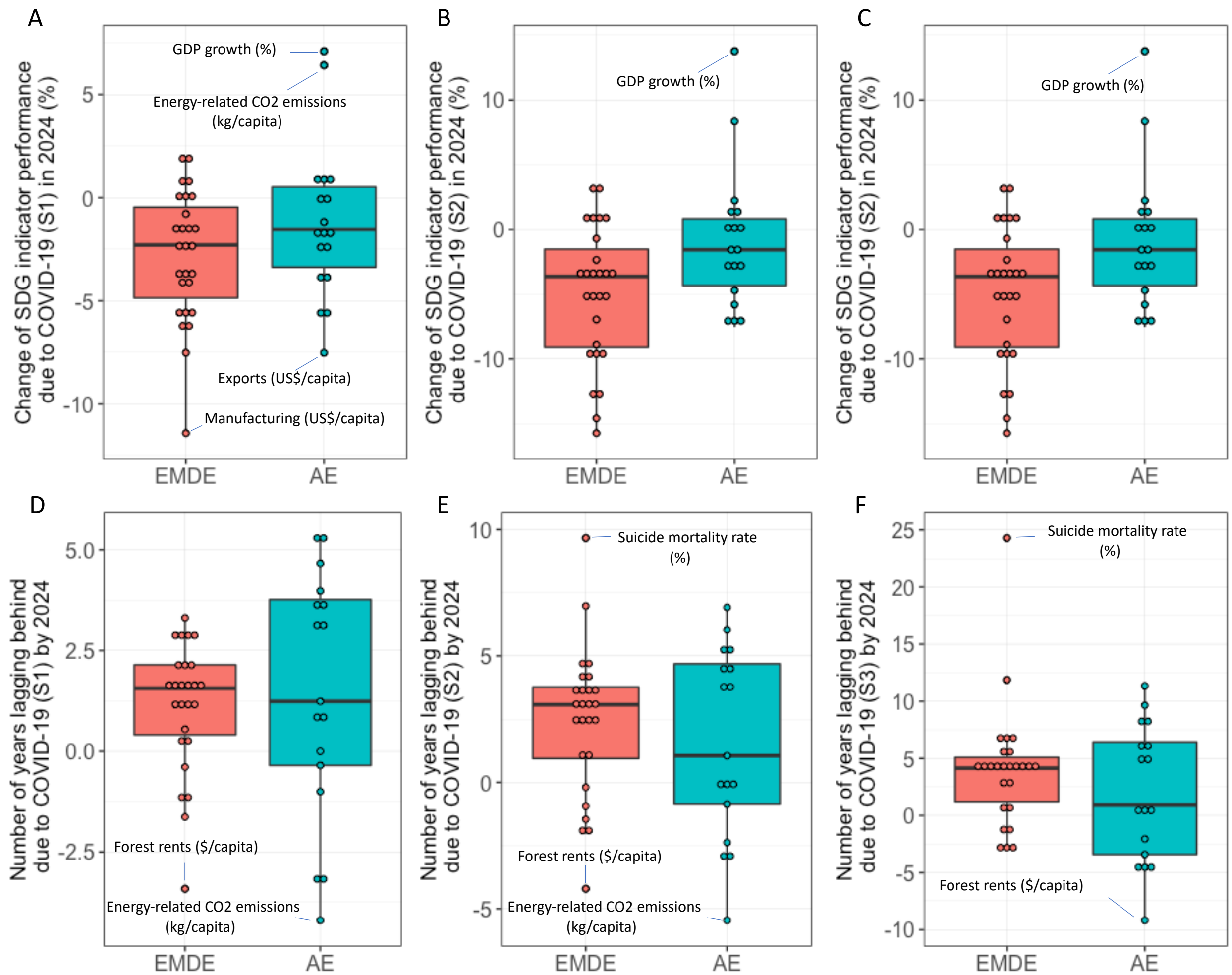

Figure 4-6. Impacts of COVID-19 on SDG indicator performances for EMDE and AD countries. (A-C) Impacts in 2024 under the three COVID-19 scenarios (S1, S2, and S3). (D-F) Number of years lagging behind the original trajectory without COVID-19 for each SDG indicator by 2024 under the three COVID-19 scenarios (S1, S2, and S3). In each boxplot, the central rectangle box spans the first to the third quartile. The central line segment inside the rectangle represents the median value. Only the indicators with testing R^2 >= 0.6 are shown.

4.4. Discussion

This study predicts SDG indicators from 2020 to 2024 in a no-COVID-19 scenario and the three COVID-19 scenarios based on projected GDP and population in each country or region. Prior to this work, most existing studies have only qualitatively evaluated the impact of COVID-

19 on SDGs, but a quantitative assessment was still lacking. The study shows COVID-19 will lead to declines of 12 socioeconomic-related SDG performances in 2020. SDGs and SDG indicators closely related to economic growth will be affected the most, such as SDG 8 (Decent Work and Economic Growth) and SDG 2 (Zero Hunger). On the other hand, four environment-related SDGs will actually be improved, likely due to reduced human activities during COVID-19, including SDG 12 (Responsible Consumption and Production), SDG 13 (Climate Action), SDG 14 (Life Below Water), and SDG 15 (Life on Land).

After 2020, the quicker GDP recovers, the quicker non-environment-related SDG performances will recover and the quicker the environment-related SDG performances will worsen. By 2024, there will still be one to eight years lagging behind for most SDGs compared to the situation without COVID-19. At the same time, the downward trajectories of the four environment-related SDGs will be slowed down for -0.1-4.1 years.

The impacts of COVID-19 on SDGs are different for different countries. EMDE countries will be affected almost twice more than AE countries in 2020. The recovery of EMDE countries are relatively slower than that of AE countries. By 2024, SDGs of the AE countries will be closer to their pre-COVID-19 trajectories than those of the EMDE countries.

The results are largely based on post-COVID-19 GDP projections. The results imply the pivotal role of rapid economic recovery on SDGs. Indeed, continuous economic growth is considered as one of the necessary condition for the success of SDGs[31, 119]. With a slower economic recovery, the recovery of SDGs will be slower and the gap caused by COVID-19 will be larger. Note that economic growth is also a barrier for improving certain environmental conditions, as indicated by the findings of improved SDG 12, 13, 14 and 15 due to COVID-19.

Post-COVID-19 economic recovery should emphasize in areas that can help decouple economic growth from negative environmental impacts.

Before COVID-19, the four environment-related SDGs—SDG 12, 13, 14 and 15—had already experienced worrisome declines moving away from the 2030 goals. COVID-19 will actually reverse the declines in these SDGs in 2020 by mitigating related environmental pressures. This is largely due to reduced human activities during the pandemic. However, as soon as the economy starts to recover after 2020, these SDGs start to come back to their original downward trajectories with declining environment quality. Nevertheless, I will still gain some extra time from COVID-19 for the four environment-related SDGs which provides a great opportunity to accelerate the global transition towards environmental sustainability. For example, previous studies estimated that the average annual low-carbon investment under a Paris-compatible pathway is about USD 1.4 trillion per year globally between 2020 and 2024[119, 120], which could be lower considering the extra time gained from COVID-19. This can be just about 10% of the total pledged COVID-19 stimulus to date[120].

This study lays a foundation for further exploring the impacts of COVID-19 on SDGs. The results reveal the different impacts of COVID-19 on individual SDGs and SDG indicators for different groups of economies. These impacts can be substantial and can greatly slow down the progress for most SDGs. Overall, COVID-19 will make global SDG progresses lag behind the original trajectory without COVID-19. Given that the SDG progress has already been difficult before COVID-19, the challenge of achieving SDGs by 2030 becomes even larger due to the pandemic. These results suggest stronger and targeted efforts are needed for SDGs post-COVID-19. The results rely on machine learning models driven by GDP and population projections. Other factors, such as technology development and new policy intervention, could also play critical roles in driving SDGs, but are excluded in the model due to the lack of reliable future projections. Future research should explore ways to incorporate other relevant variables in the prediction. The results are also based on the assumption that the tested relationship between the predictors (GDP, population, etc.) and each of the responses (SDG indicators) will also remain in the future. In addition, I also found pandemic-related indicators are scarce in existing SDG indicators, especially for SDG 3 (Good Health and Well-Being). Currently there is no indicator in SDG 3 directly on pandemics. Future efforts should consider including pandemic-related indicators in the suite of SDG indicators to better reflect the impact of pandemics on sustainable development.

4.5. Limitations and Future Research

The predicted SDG performance is aggregated from only 43 SDG indicators which are most relevant to GDP and population. This means the information of other SDG indicators, like gender, may not be reflected in the SDG performance, which increases the uncertainty of the predicted SDG performance. This is because the main predictors in this study are population and GDP. For future research, I will use additional predictors to increase the predictive ability of the model and reduce the uncertainty of the predicted SDG performance.

This research only focuses on the predicted SDG performance at the global level rather than at the individual country level. This is because I developed the prediction model at the global level and the model can only capture major variations (>= 60%) of a specific indicator, but the predicted value may not be as reliable for individual countries. For future study, I will develop models at the county level to increase the reliability of predicted values for individual countries.

Chapter 5. Conclusions and Future Research

Drawing from the rapid development of data science, my research applies data-driven methods to provide efficient and effective solutions to assist sustainable development for nations. Specifically, I focus on addressing the challenges at data collection, performance comparison, and prediction in the implementation of SDGs.

First, for data collection (Chapter 2), collecting a large number of SDG indicators with limited resources is extremely challenging. Reducing data demand by finding the principal indicators provides a practical solution for the challenge. Using principal component analysis and multiple regression considering collection cost for each indicator, I identify the principal indicators to represent almost full information of all the SDG indicators. Results show that 147 principal indicators can represent at least 90% of the annual variances of 351 SDG indicators in the past (2000-2017) and are expected to do so for the future (2018-2030) with the lowest difficulty of data collection. However, I do not necessarily recommend to only track principal indicators, as established systems may already exist to collect data for other indicators for other purposes. I would also recommend to regularly examine the principal indicators in the future to reflect the changes of data collection infrastructure. Principal indicators are identified based on the historical correlations between individual indicators. However, some correlations may change over time. For future study, I may consider developing an integrated index based on principal indicators to represent the SDG indicators for an overall evaluation of SDG progress for countries and regions.

Second, for performance comparison (Chapter 3), per capita based measures ignore the effect of agglomeration resulting from non-linear interactions in social dynamics. There needs a more appropriate approach to compare countries for their progress toward sustainable development by taking into account the non-linear relationship between population and sustainable development indicators. Building upon the scaling law in cities, I examine the scaling of sustainable development indicators with the population in countries and develop a quantitative framework to explain the origins of such scaling. Empirical results show that indicators of socioeconomic activities scale sub-linearly ($\beta \approx 0.9$), public health indicators scale super-linearly ($\beta \approx 1.1$), and indicators of individual needs scale linearly ($\beta \approx 1.0$) with the population in countries. I also show that, keeping other factors constant, if a country could concentrate people and resources in megacities while ensuring social cohesion and environmental sustainability, its development indicators would significantly improve. For future research, I will improve the quantitative framework explaining the origin of scaling in countries by considering the non-linear relationship between development indicators and population in rural areas.

Third, for performance prediction (Chapter 4), complex non-linear relationship among the SDG indicators makes prediction difficult. For example, several studies attempted to evaluate the impacts of COVID-19 on SDGs, but can only do so in a non-quantitative way. I develop machine learning models to quantitatively predict the impact of COVID-19 on SDGs. Results show that the overall SDG performance declined by 7.7% in 2020 at the global scale, with the performance of 12 socioeconomic SDGs decreasing by 3.0-22.3% and that of 4 environmental SDGs increasing by 1.6-9.2%. By 2024, the progress of 12 SDGs will lag behind for one to eight years compared to their pre-COVID-19 trajectories, while extra time will be gained for 4 environment-related SDGs. In addition, the pandemic will cause more negative impacts on SDGs for countries in emerging market and developing economy than for those in advanced economy. Future efforts should consider including pandemic-related indicators in the suite of SDG indicators to better reflect the impact of pandemics on sustainable development.

Appendices

Appendix A. Supporting Information for Chapter 2

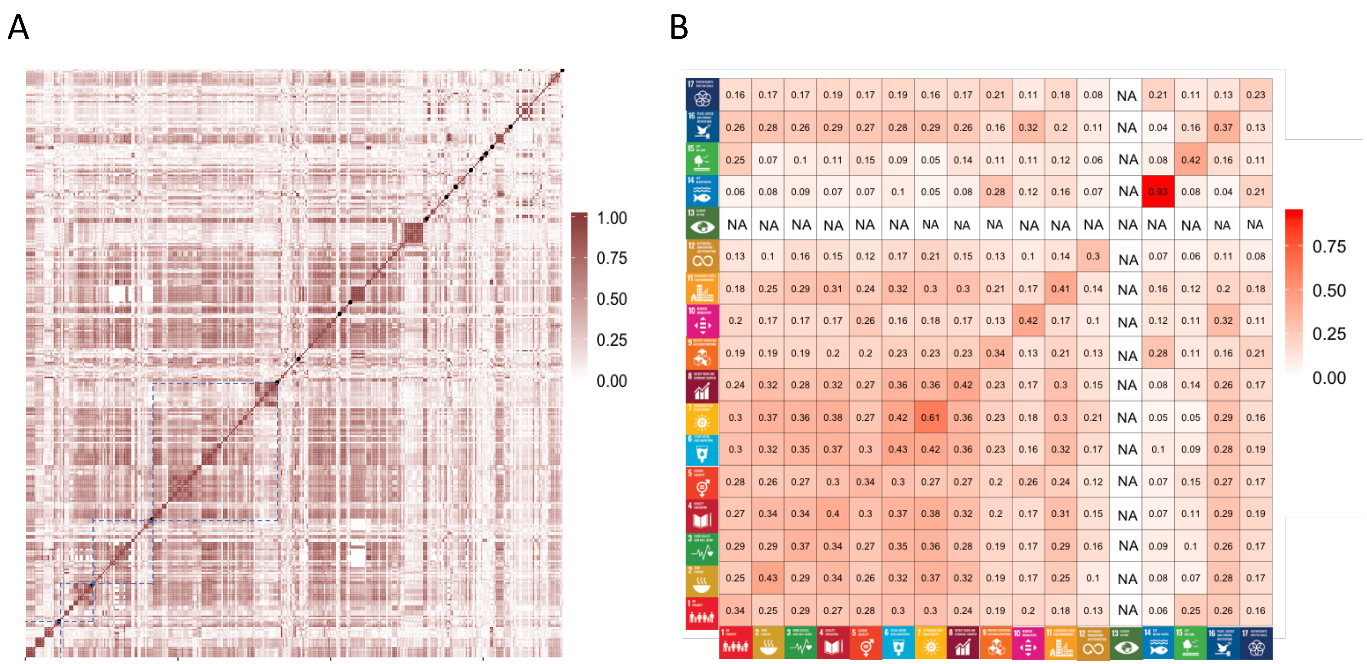

Figure A-1. Correlation between pairs of SDG indicators and SDGs.
(A) Heatmap of absolute Pearson correlation coefficients between pairs of the 351 SDG indicators. (B) Heatmap of average correlation coefficients of indicators between different SDGs. “NA” means there is no data for indicators in SDG 13 (Climate Action).

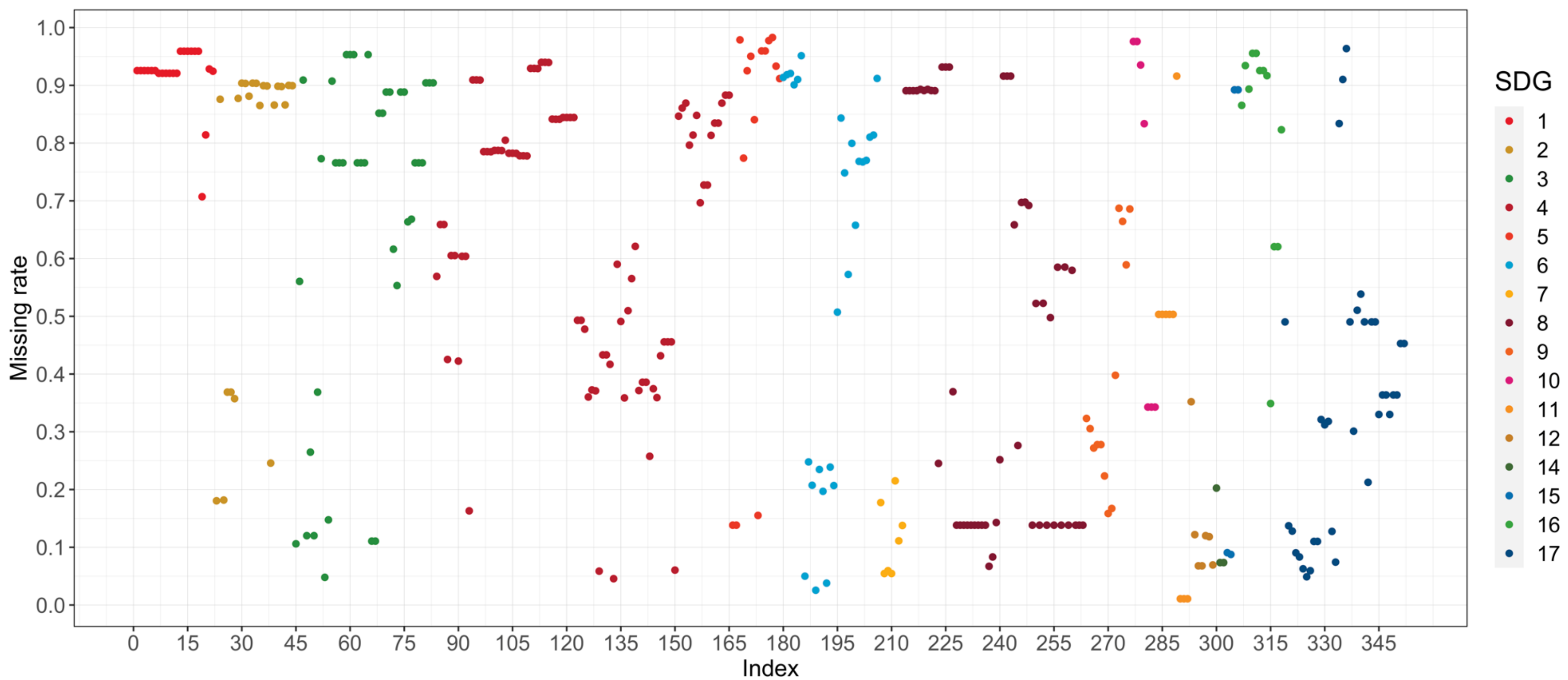

Figure A-2. Missing rates of the 351 SDG indicators (indexed by x-axis) between 2000 and 2017.

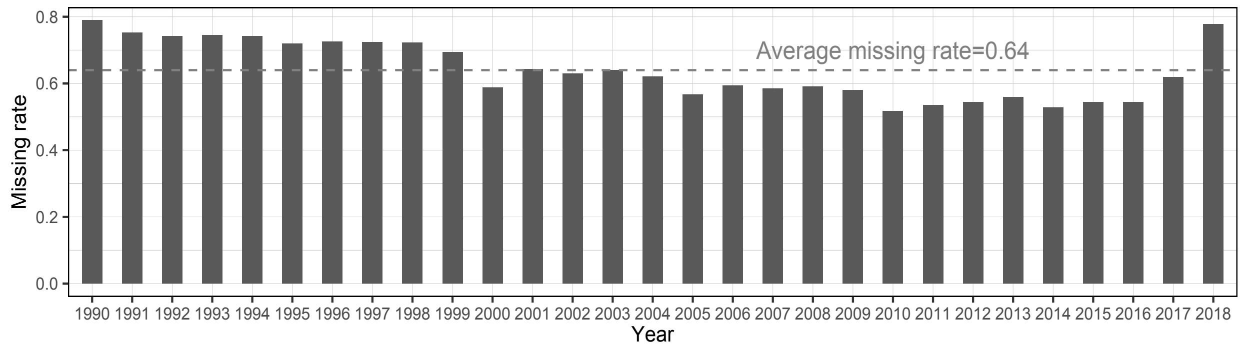

Figure A-3. Annual missing rate of all SDG indicators across countries and regions.
Dash line shows the average. Note that I exclude the 2018 data due to high missing rate.

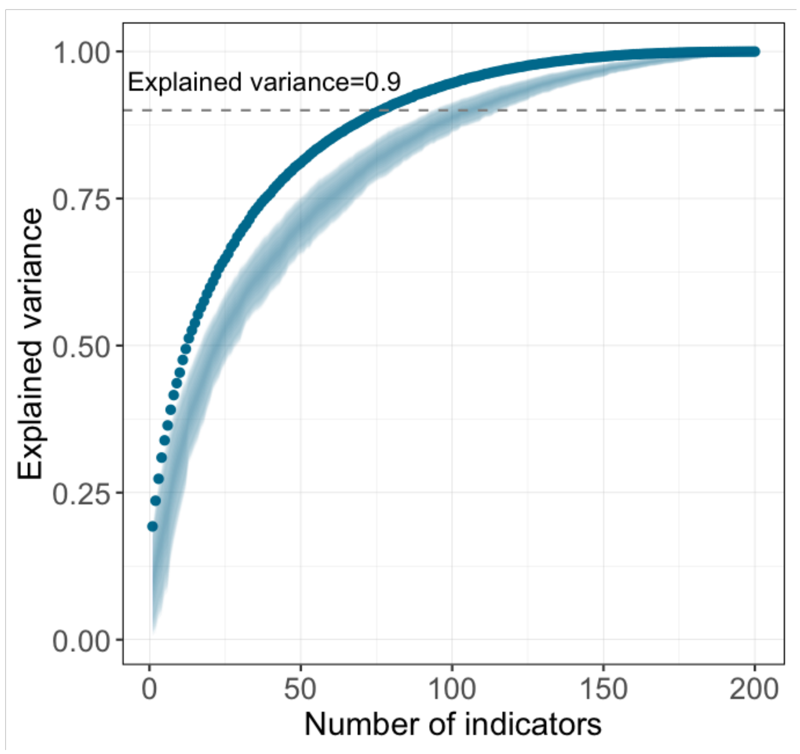

Figure A-4. Explained variance of subsets of indicators on the entire dataset without considering the difficulty of data collection.
The dotted line represents the explained variance of principal indicators in various sizes, and the shade indicates the explained variance of randomly selected indicators. Note that the explained variance of the principal indicators is substantially higher than that explained by randomly selected indicators with the same size. Overall, 77 principal indicators are needed to explain at least 90% variance for all SDG indicators from 2000 to 2017 without considering the difficulty of data collection.

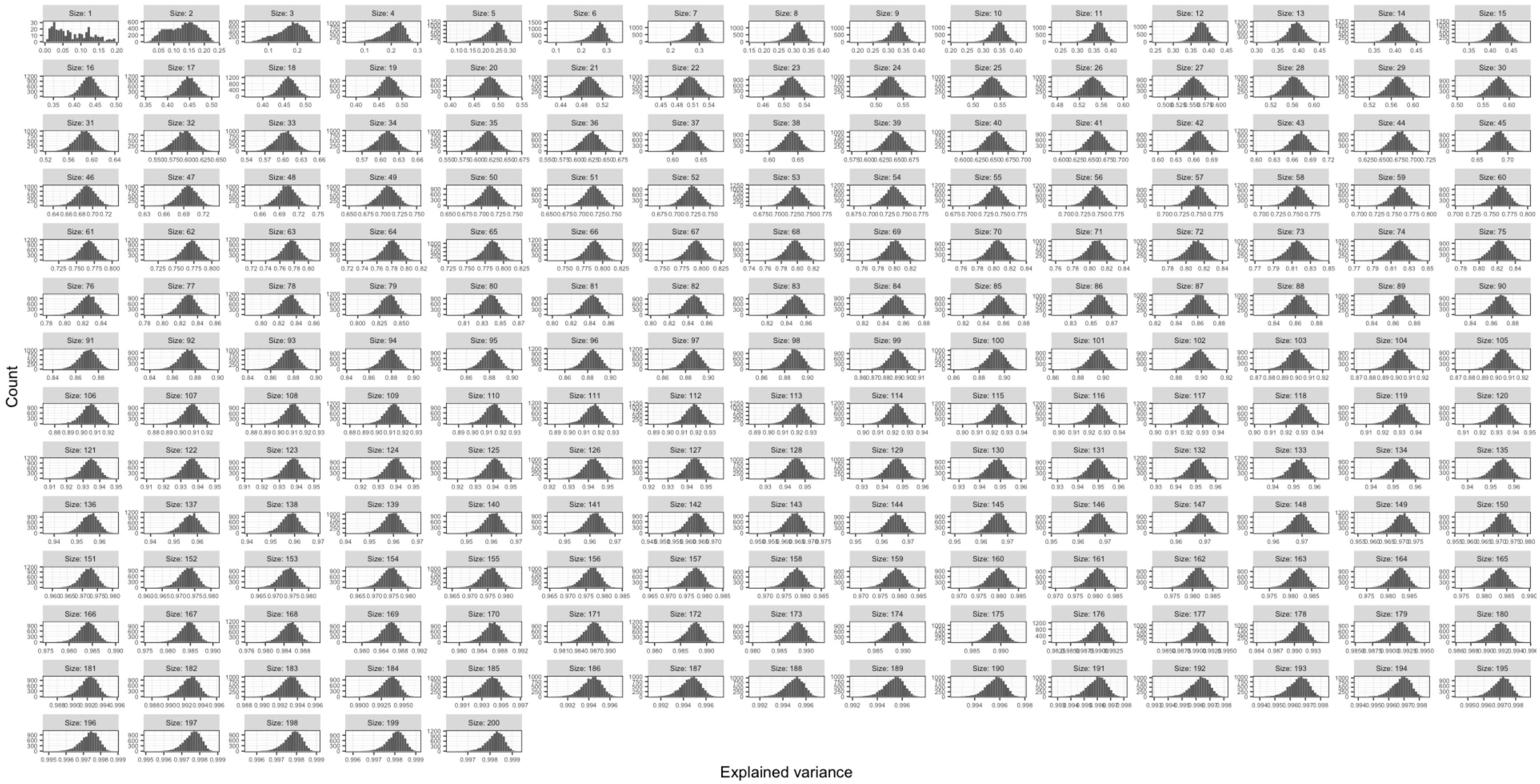

Figure A-5. Distribution of explained variance of randomly selected indicators at different size without considering the difficulty of data collection.

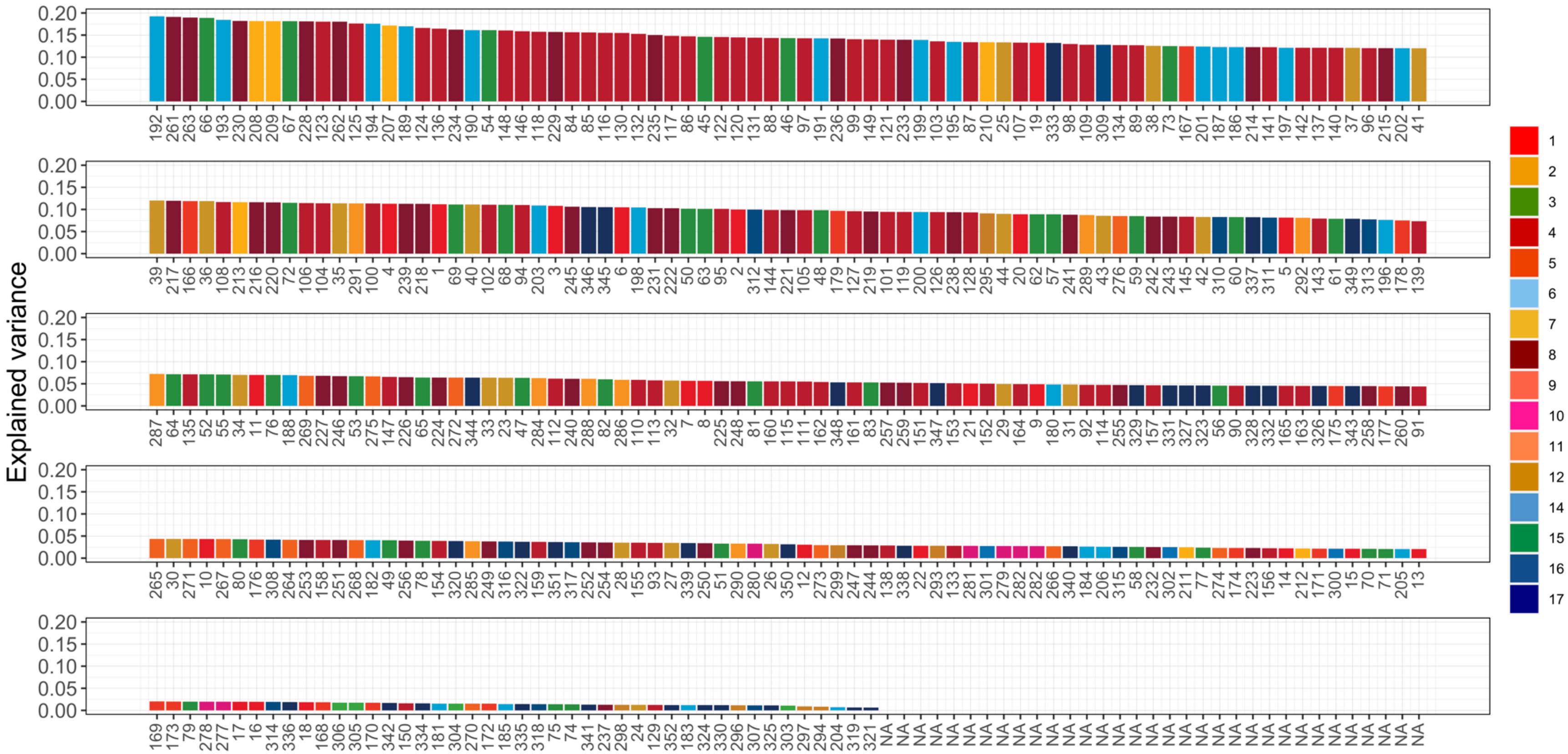

Figure A-6. SDG indicators ranked in explained variance on the full dataset without considering the difficulty of data collection.

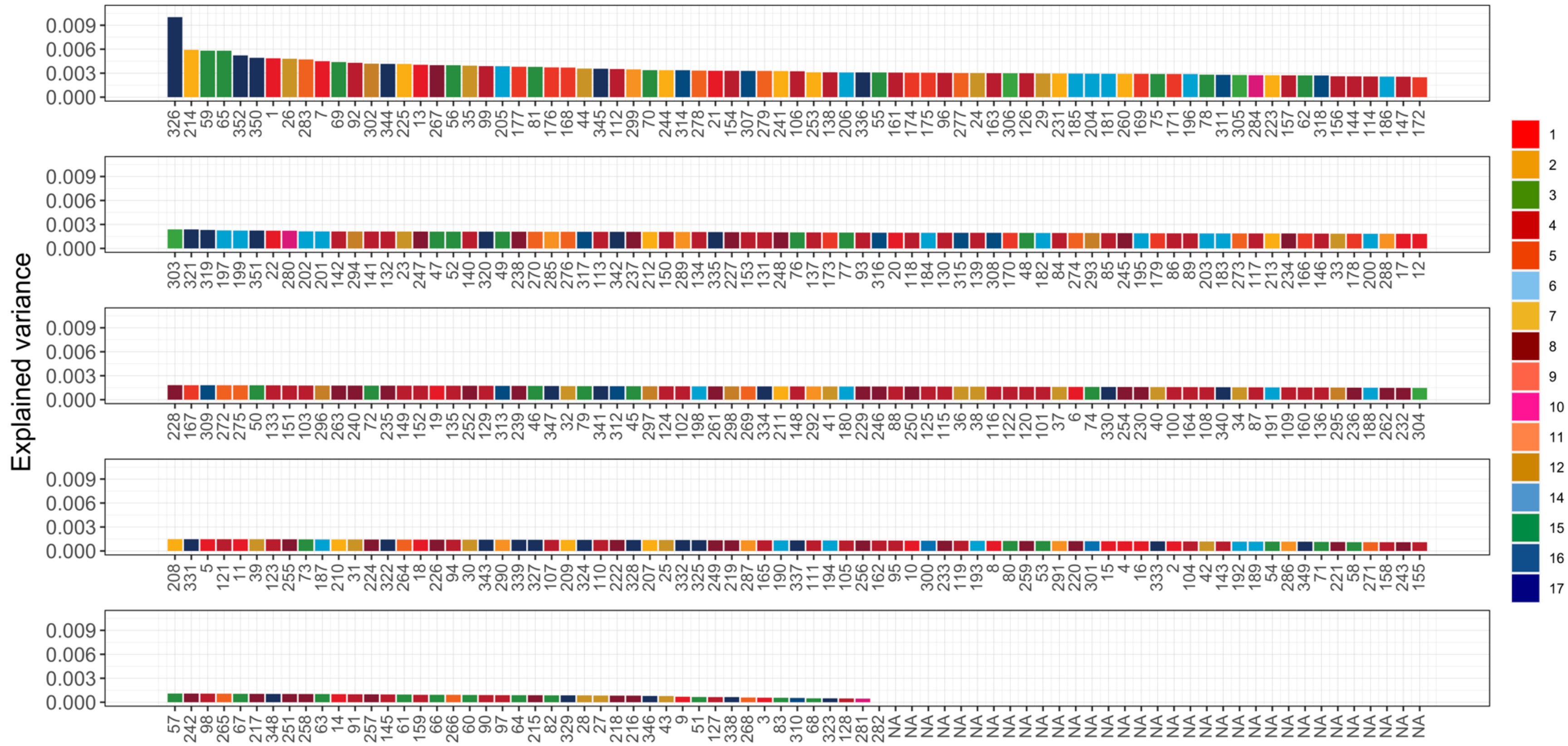

Figure A-7. SDG indicators ranked in marginal explained variance on the entire dataset without considering the difficulty of data collection. Note that only five of the 77 of top-ranked indicators are selected as principal indicators. This is due to the strong correlations among these top-ranked indicators, which makes the marginal explained variance of one additional top-ranked indicator declines quickly. As a result, the amount of variance that can be explained by these top-ranked indicators is substantially smaller than what can be explained by the same number of principal indicators (65% vs. 90%). The top 77 indicators explain the most marginal variance are exactly the 77 principal indicators. This further validates my method.

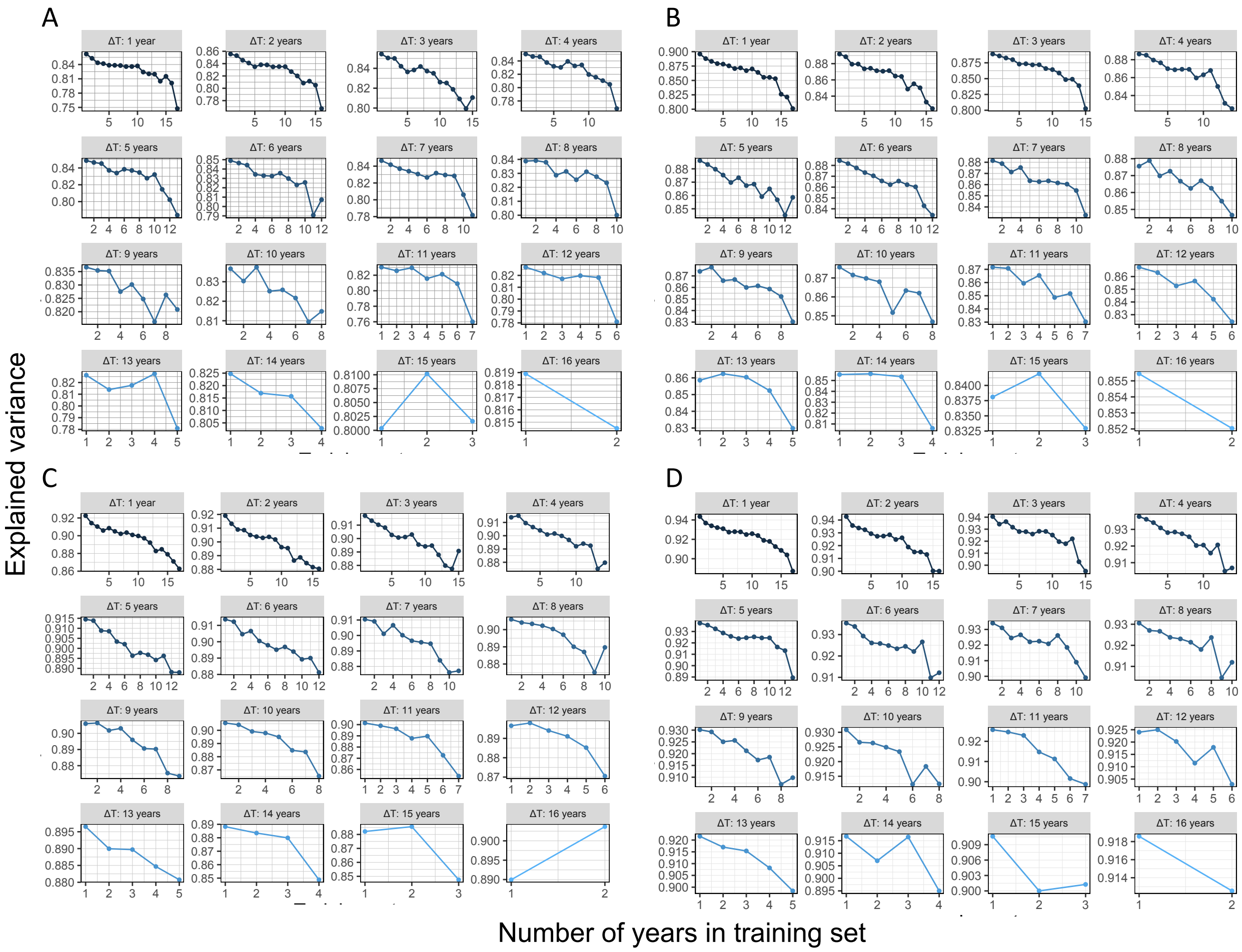

Figure A-8. Explained variances of a certain number of principal indicators identified from various training sets on test sets of each of the future years without considering the difficulty of data collection.
ΔT indicates the period between the test set year and the last year of the training set. (A)-(D) 60, 70, 80, and 90 principal indicators, respectively. These results show the year 2017 is the best dataset to identify the principal indicators to monitor future SDG progress without considering the difficulty of data collection.

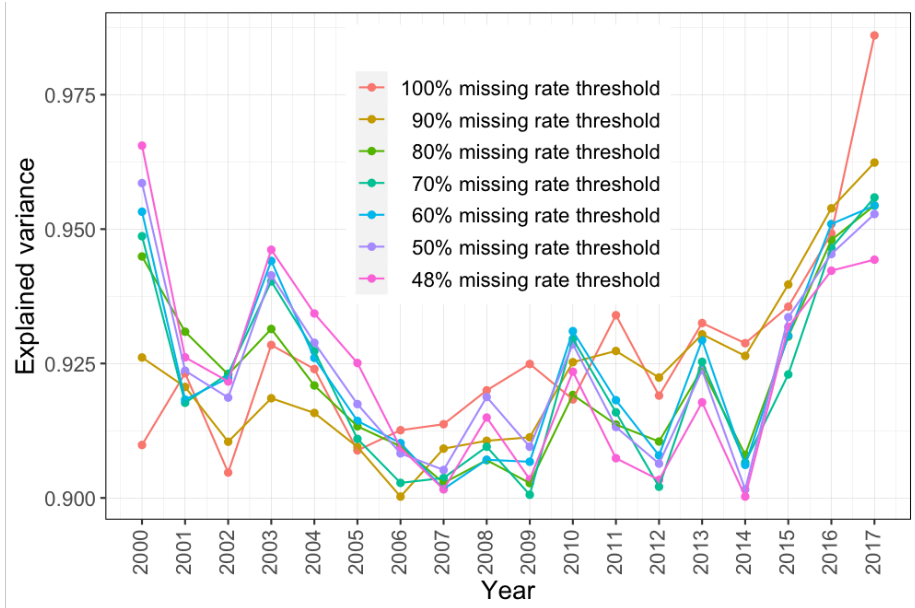

Figure A-9. Explained variances of the least number of principal indicators identified under various missing rate thresholds to explain at least 90% of the annual variances of the dataset.

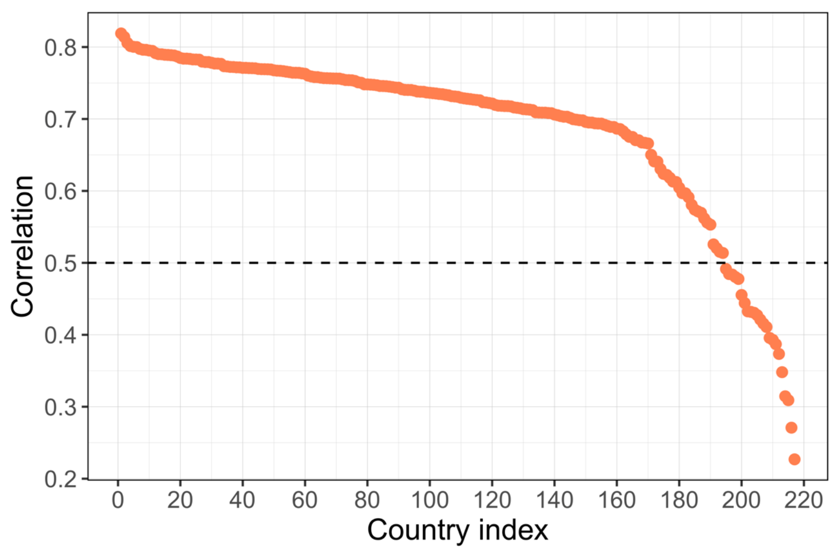

Figure A-10. Correlation coefficient between the structure of indicator missing rates for each country or region and that of the latest year.

Result shows that 193 countries or regions have very similar missing rate structure (correlation coefficient above 0.5) with the latest year.

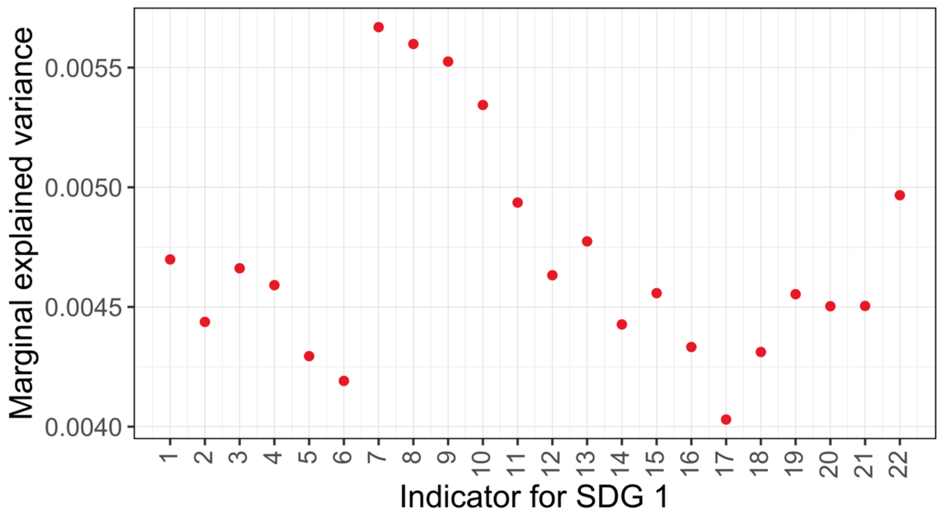

Figure A-11. Marginal explained variances of individual SDG 1 indicators added to the existing 147 principal indicators.

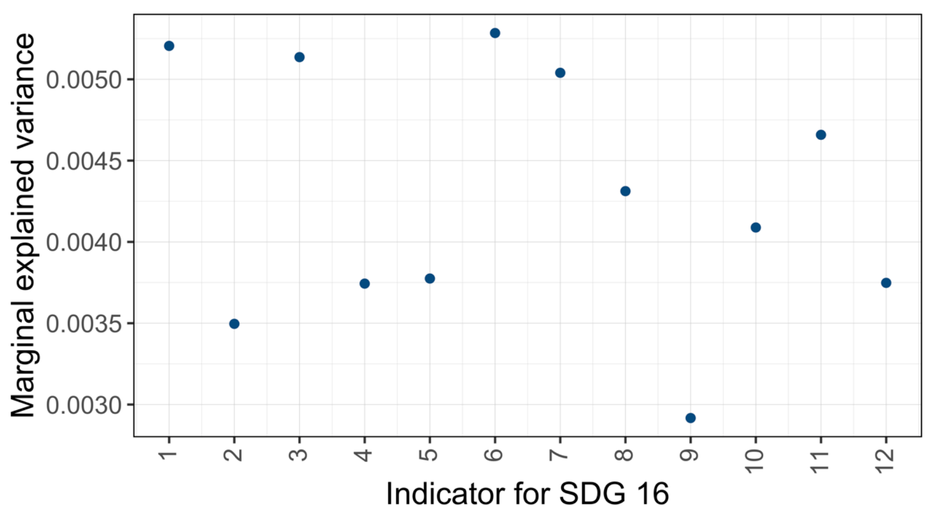

Figure A-12. Marginal explained variances of individual SDG 16 indicators added to the existing 147 principal indicators

Table A-1. The 351 SDG indicators from the World Bank dataset.

Number	SDG indicator	SDG
1	Coverage of social insurance programs (% of population)	1
2	Coverage of social insurance programs in 2nd quintile (% of population)	1
3	Coverage of social insurance programs in 3rd quintile (% of population)	1
4	Coverage of social insurance programs in 4th quintile (% of population)	1
5	Coverage of social insurance programs in poorest quintile (% of population)	1
6	Coverage of social insurance programs in richest quintile (% of population)	1
7	Coverage of social safety net programs (% of population)	1
8	Coverage of social safety net programs in 2nd quintile (% of population)	1
9	Coverage of social safety net programs in 3rd quintile (% of population)	1
10	Coverage of social safety net programs in 4th quintile (% of population)	1
11	Coverage of social safety net programs in poorest quintile (% of population)	1
12	Coverage of social safety net programs in richest quintile (% of population)	1
13	Coverage of unemployment benefits and ALMP (% of population)	1
14	Coverage of unemployment benefits and ALMP in 2nd quintile (% of population)	1
15	Coverage of unemployment benefits and ALMP in 3rd quintile (% of population)	1
16	Coverage of unemployment benefits and ALMP in 4th quintile (% of population)	1
17	Coverage of unemployment benefits and ALMP in poorest quintile (% of population)	1
18	Coverage of unemployment benefits and ALMP in richest quintile (% of population)	1
19	Poverty headcount ratio at $1.90 a day (2011 PPP) (% of population)	1
20	Poverty headcount ratio at national poverty lines (% of population)	1
21	Rural poverty headcount ratio at national poverty lines (% of rural population)	1
22	Urban poverty headcount ratio at national poverty lines (% of urban population)	1
23	Cereal yield (kg per hectare)	2
24	Exclusive breastfeeding (% of children under 6 months)	2

25	Prevalence of anemia among women of reproductive age (% of women ages 15-49)	2
26	Prevalence of HIV, female (% ages 15-24)	2
27	Prevalence of HIV, male (% ages 15-24)	2
28	Prevalence of HIV, total (% of population ages 15-49)	2
29	Prevalence of overweight, weight for height (% of children under 5)	2
30	Prevalence of overweight, weight for height, female (% of children under 5)	2
31	Prevalence of overweight, weight for height, male (% of children under 5)	2
32	Prevalence of severe wasting, weight for height (% of children under 5)	2
33	Prevalence of severe wasting, weight for height, female (% of children under 5)	2
34	Prevalence of severe wasting, weight for height, male (% of children under 5)	2
35	Prevalence of stunting, height for age (% of children under 5)	2
36	Prevalence of stunting, height for age, female (% of children under 5)	2
37	Prevalence of stunting, height for age, male (% of children under 5)	2
38	Prevalence of undernourishment (% of population)	2
39	Prevalence of underweight, weight for age (% of children under 5)	2
40	Prevalence of underweight, weight for age, female (% of children under 5)	2
41	Prevalence of underweight, weight for age, male (% of children under 5)	2
42	Prevalence of wasting, weight for height (% of children under 5)	2
43	Prevalence of wasting, weight for height, female (% of children under 5)	2
44	Prevalence of wasting, weight for height, male (% of children under 5)	2
45	Adolescent fertility rate (births per 1,000 women ages 15-19)	3
46	Births attended by skilled health staff (% of total)	3
47	Demand for family planning satisfied by modern methods (% of married women with demand for family planning)	3
48	Immunization, DPT (% of children ages 12-23 months)	3
49	Immunization, HepB3 (% of one-year-old children)	3
50	Immunization, measles (% of children ages 12-23 months)	3

51	Incidence of HIV (per 1,000 uninfected population ages 15-49)	3
52	Incidence of malaria (per 1,000 population at risk)	3
53	Incidence of tuberculosis (per 100,000 people)	3
54	Maternal mortality ratio (modeled estimate, per 100,000 live births)	3
55	Mortality caused by road traffic injury (per 100,000 people)	3
56	Mortality from CVD, cancer, diabetes or CRD between exact ages 30 and 70 (%)	3
57	Mortality from CVD, cancer, diabetes or CRD between exact ages 30 and 70, female (%)	3
58	Mortality from CVD, cancer, diabetes or CRD between exact ages 30 and 70, male (%)	3
59	Mortality rate attributed to household and ambient air pollution, age-standardized (per 100,000 population)	3
60	Mortality rate attributed to household and ambient air pollution, age-standardized, female (per 100,000 female population)	3
61	Mortality rate attributed to household and ambient air pollution, age-standardized, male (per 100,000 male population)	3
62	Mortality rate attributed to unintentional poisoning (per 100,000 population)	3
63	Mortality rate attributed to unintentional poisoning, female (per 100,000 female population)	3
64	Mortality rate attributed to unintentional poisoning, male (per 100,000 male population)	3
65	Mortality rate attributed to unsafe water, unsafe sanitation and lack of hygiene (per 100,000 population)	3
66	Mortality rate, neonatal (per 1,000 live births)	3
67	Mortality rate, under-5 (per 1,000 live births)	3
68	Mortality rate, under-5, female (per 1,000 live births)	3
69	Mortality rate, under-5, male (per 1,000 live births)	3
70	Number of people spending more than 10% of household consumption or income on out-of-pocket health care expenditure	3
71	Number of people spending more than 25% of household consumption or income on out-of-pocket health care expenditure	3
72	Nurses and midwives (per 1,000 people)	3
73	Physicians (per 1,000 people)	3
74	Proportion of population spending more than 10% of household consumption or income on out-of-pocket health care expenditure (%)	3
75	Proportion of population spending more than 25% of household consumption or income on out-of-pocket health care expenditure (%)	3
76	Smoking prevalence, females (% of adults)	3

77	Smoking prevalence, males (% of adults)	3
78	Suicide mortality rate (per 100,000 population)	3
79	Suicide mortality rate, female (per 100,000 female population)	3
80	Suicide mortality rate, male (per 100,000 male population)	3
81	Total alcohol consumption per capita (liters of pure alcohol, projected estimates, 15+ years of age)	3
82	Total alcohol consumption per capita, female (liters of pure alcohol, projected estimates, female 15+ years of age)	3
83	Total alcohol consumption per capita, male (liters of pure alcohol, projected estimates, male 15+ years of age)	3
84	Adolescents out of school (% of lower secondary school age)	4
85	Adolescents out of school, female (% of female lower secondary school age)	4
86	Adolescents out of school, male (% of male lower secondary school age)	4
87	Children out of school (% of primary school age)	4
88	Children out of school, female (% of female primary school age)	4
89	Children out of school, male (% of male primary school age)	4
90	Children out of school, primary	4
91	Children out of school, primary, female	4
92	Children out of school, primary, male	4
93	Compulsory education, duration (years)	4
94	Educational attainment, at least Bachelor's or equivalent, population 25+, female (%) (cumulative)	4
95	Educational attainment, at least Bachelor's or equivalent, population 25+, male (%) (cumulative)	4
96	Educational attainment, at least Bachelor's or equivalent, population 25+, total (%) (cumulative)	4
97	Educational attainment, at least completed lower secondary, population 25+, female (%) (cumulative)	4
98	Educational attainment, at least completed lower secondary, population 25+, male (%) (cumulative)	4
99	Educational attainment, at least completed lower secondary, population 25+, total (%) (cumulative)	4
100	Educational attainment, at least completed post-secondary, population 25+, female (%) (cumulative)	4
101	Educational attainment, at least completed post-secondary, population 25+, male (%) (cumulative)	4
102	Educational attainment, at least completed post-secondary, population 25+, total (%) (cumulative)	4

103	Educational attainment, at least completed primary, population 25+ years, male (%) (cumulative)	4
104	Educational attainment, at least completed short-cycle tertiary, population 25+, female (%) (cumulative)	4
105	Educational attainment, at least completed short-cycle tertiary, population 25+, male (%) (cumulative)	4
106	Educational attainment, at least completed short-cycle tertiary, population 25+, total (%) (cumulative)	4
107	Educational attainment, at least completed upper secondary, population 25+, female (%) (cumulative)	4
108	Educational attainment, at least completed upper secondary, population 25+, male (%) (cumulative)	4
109	Educational attainment, at least completed upper secondary, population 25+, total (%) (cumulative)	4
110	Educational attainment, at least Master's or equivalent, population 25+, female (%) (cumulative)	4
111	Educational attainment, at least Master's or equivalent, population 25+, male (%) (cumulative)	4
112	Educational attainment, at least Master's or equivalent, population 25+, total (%) (cumulative)	4
113	Educational attainment, Doctoral or equivalent, population 25+, female (%) (cumulative)	4
114	Educational attainment, Doctoral or equivalent, population 25+, male (%) (cumulative)	4
115	Educational attainment, Doctoral or equivalent, population 25+, total (%) (cumulative)	4
116	Literacy rate, adult female (% of females ages 15 and above)	4
117	Literacy rate, adult male (% of males ages 15 and above)	4
118	Literacy rate, adult total (% of people ages 15 and above)	4
119	Literacy rate, youth (ages 15-24), gender parity index (GPI)	4
120	Literacy rate, youth female (% of females ages 15-24)	4
121	Literacy rate, youth male (% of males ages 15-24)	4
122	Literacy rate, youth total (% of people ages 15-24)	4
123	Lower secondary completion rate, female (% of relevant age group)	4
124	Lower secondary completion rate, male (% of relevant age group)	4
125	Lower secondary completion rate, total (% of relevant age group)	4
126	Over-age students, primary (% of enrollment)	4
127	Over-age students, primary, female (% of female enrollment)	4
128	Over-age students, primary, male (% of male enrollment)	4

129	Preprimary education, duration (years)	4
130	Primary completion rate, female (% of relevant age group)	4
131	Primary completion rate, male (% of relevant age group)	4
132	Primary completion rate, total (% of relevant age group)	4
133	Primary education, duration (years)	4
134	Pupil-teacher ratio, lower secondary	4
135	Pupil-teacher ratio, preprimary	4
136	Pupil-teacher ratio, primary	4
137	Pupil-teacher ratio, secondary	4
138	Pupil-teacher ratio, tertiary	4
139	Pupil-teacher ratio, upper secondary	4
140	School enrollment, preprimary (% gross)	4
141	School enrollment, preprimary, female (% gross)	4
142	School enrollment, preprimary, male (% gross)	4
143	School enrollment, primary (gross), gender parity index (GPI)	4
144	School enrollment, primary and secondary (gross), gender parity index (GPI)	4
145	School enrollment, secondary (gross), gender parity index (GPI)	4
146	School enrollment, tertiary (% gross)	4
147	School enrollment, tertiary (gross), gender parity index (GPI)	4
148	School enrollment, tertiary, female (% gross)	4
149	School enrollment, tertiary, male (% gross)	4
150	Secondary education, duration (years)	4
151	Trained teachers in lower secondary education (% of total teachers)	4
152	Trained teachers in lower secondary education, female (% of female teachers)	4
153	Trained teachers in lower secondary education, male (% of male teachers)	4
154	Trained teachers in preprimary education (% of total teachers)	4

155	Trained teachers in preprimary education, female (% of female teachers)	4
156	Trained teachers in preprimary education, male (% of male teachers)	4
157	Trained teachers in primary education (% of total teachers)	4
158	Trained teachers in primary education, female (% of female teachers)	4
159	Trained teachers in primary education, male (% of male teachers)	4
160	Trained teachers in secondary education (% of total teachers)	4
161	Trained teachers in secondary education, female (% of female teachers)	4
162	Trained teachers in secondary education, male (% of male teachers)	4
163	Trained teachers in upper secondary education (% of total teachers)	4
164	Trained teachers in upper secondary education, female (% of female teachers)	4
165	Trained teachers in upper secondary education, male (% of male teachers)	4
166	Contributing family workers, female (% of female employment) (modeled ILO estimate)	5
167	Contributing family workers, male (% of male employment) (modeled ILO estimate)	5
168	Female genital mutilation prevalence (%)	5
169	Female share of employment in senior and middle management (%)	5
170	Firms with female participation in ownership (% of firms)	5
171	Firms with female top manager (% of firms)	5
172	Nondiscrimination clause mentions gender in the constitution (1=yes; 0=no)	5
173	Proportion of seats held by women in national parliaments (%)	5
174	Proportion of time spent on unpaid domestic and care work, female (% of 24 hour day)	5
175	Proportion of time spent on unpaid domestic and care work, male (% of 24 hour day)	5
176	Proportion of women subjected to physical and/or sexual violence in the last 12 months (% of women age 15-49)	5
177	Women making their own informed decisions regarding sexual relations, contraceptive use and reproductive health care (% of women age 15-49)	5
178	Women who were first married by age 15 (% of women ages 20-24)	5
179	Women who were first married by age 18 (% of women ages 20-24)	5

180	Annual freshwater withdrawals, agriculture (% of total freshwater withdrawal)	6
181	Annual freshwater withdrawals, domestic (% of total freshwater withdrawal)	6
182	Annual freshwater withdrawals, industry (% of total freshwater withdrawal)	6
183	Annual freshwater withdrawals, total (% of internal resources)	6
184	Annual freshwater withdrawals, total (billion cubic meters)	6
185	Level of water stress: freshwater withdrawal as a proportion of available freshwater resources	6
186	People practicing open defecation (% of population)	6
187	People practicing open defecation, rural (% of rural population)	6
188	People practicing open defecation, urban (% of urban population)	6
189	People using at least basic drinking water services (% of population)	6
190	People using at least basic drinking water services, rural (% of rural population)	6
191	People using at least basic drinking water services, urban (% of urban population)	6
192	People using at least basic sanitation services (% of population)	6
193	People using at least basic sanitation services, rural (% of rural population)	6
194	People using at least basic sanitation services, urban (% of urban population)	6
195	People using safely managed drinking water services (% of population)	6
196	People using safely managed drinking water services, rural (% of rural population)	6
197	People using safely managed drinking water services, urban (% of urban population)	6
198	People using safely managed sanitation services (% of population)	6
199	People using safely managed sanitation services, rural (% of rural population)	6
200	People using safely managed sanitation services, urban (% of urban population)	6
201	People with basic handwashing facilities including soap and water (% of population)	6
202	People with basic handwashing facilities including soap and water, rural (% of rural population)	6
203	People with basic handwashing facilities including soap and water, urban (% of urban population)	6
204	Renewable internal freshwater resources per capita (cubic meters)	6
205	Renewable internal freshwater resources, total (billion cubic meters)	6

206	Water productivity, total (constant 2010 US$ GDP per cubic meter of total freshwater withdrawal)	6
207	Access to clean fuels and technologies for cooking (% of population)	7
208	Access to electricity (% of population)	7
209	Access to electricity, rural (% of rural population)	7
210	Access to electricity, urban (% of urban population)	7
211	Energy intensity level of primary energy (MJ/$2011 PPP GDP)	7
212	Renewable electricity output (% of total electricity output)	7
213	Renewable energy consumption (% of total final energy consumption)	7
214	Account ownership at a financial institution or with a mobile-money-service provider (% of population ages 15+)	8
215	Account ownership at a financial institution or with a mobile-money-service provider, female (% of population ages 15+)	8
216	Account ownership at a financial institution or with a mobile-money-service provider, male (% of population ages 15+)	8
217	Account ownership at a financial institution or with a mobile-money-service provider, older adults (% of population ages 25+)	8
218	Account ownership at a financial institution or with a mobile-money-service provider, poorest 40% (% of population ages 15+)	8
219	Account ownership at a financial institution or with a mobile-money-service provider, primary education or less (% of population ages 15+)	8
220	Account ownership at a financial institution or with a mobile-money-service provider, richest 60% (% of population ages 15+)	8
221	Account ownership at a financial institution or with a mobile-money-service provider, secondary education or more (% of population ages 15+)	8
222	Account ownership at a financial institution or with a mobile-money-service provider, young adults (% of population ages 15-24)	8
223	Agriculture, forestry, and fishing, value added per worker (constant 2010 US$)	8
224	Children in employment, female (% of female children ages 7-14)	8
225	Children in employment, male (% of male children ages 7-14)	8
226	Children in employment, total (% of children ages 7-14)	8
227	Commercial bank branches (per 100,000 adults)	8
228	Employment in agriculture (% of total employment) (modeled ILO estimate)	8
229	Employment in agriculture, female (% of female employment) (modeled ILO estimate)	8
230	Employment in agriculture, male (% of male employment) (modeled ILO estimate)	8

231	Employment in industry (% of total employment) (modeled ILO estimate)	8
232	Employment in industry, female (% of female employment) (modeled ILO estimate)	8
233	Employment in industry, male (% of male employment) (modeled ILO estimate)	8
234	Employment in services (% of total employment) (modeled ILO estimate)	8
235	Employment in services, female (% of female employment) (modeled ILO estimate)	8
236	Employment in services, male (% of male employment) (modeled ILO estimate)	8
237	GDP growth (annual %)	8
238	GDP per capita (constant 2010 US$)	8
239	GDP per person employed (constant 2011 PPP $)	8
240	Industry (including construction), value added per worker (constant 2010 US$)	8
241	Informal employment (% of total non-agricultural employment)	8
242	Informal employment, female (% of total non-agricultural employment)	8
243	Informal employment, male (% of total non-agricultural employment)	8
244	New business density (new registrations per 1,000 people ages 15-64)	8
245	Services, value added per worker (constant 2010 US$)	8
246	Share of youth not in education, employment or training, female (% of female youth population)	8
247	Share of youth not in education, employment or training, male (% of male youth population)	8
248	Share of youth not in education, employment or training, total (% of youth population)	8
249	Unemployment, female (% of female labor force) (modeled ILO estimate)	8
250	Unemployment, female (% of female labor force) (national estimate)	8
251	Unemployment, male (% of male labor force) (modeled ILO estimate)	8
252	Unemployment, male (% of male labor force) (national estimate)	8
253	Unemployment, total (% of total labor force) (modeled ILO estimate)	8
254	Unemployment, total (% of total labor force) (national estimate)	8
255	Unemployment, youth female (% of female labor force ages 15-24) (modeled ILO estimate)	8
256	Unemployment, youth female (% of female labor force ages 15-24) (national estimate)	8

257	Unemployment, youth male (% of male labor force ages 15-24) (modeled ILO estimate)	8
258	Unemployment, youth male (% of male labor force ages 15-24) (national estimate)	8
259	Unemployment, youth total (% of total labor force ages 15-24) (modeled ILO estimate)	8
260	Unemployment, youth total (% of total labor force ages 15-24) (national estimate)	8
261	Wage and salaried workers, female (% of female employment) (modeled ILO estimate)	8
262	Wage and salaried workers, male (% of male employment) (modeled ILO estimate)	8
263	Wage and salaried workers, total (% of total employment) (modeled ILO estimate)	8
264	Air transport, freight (million ton-km)	9
265	Air transport, passengers carried	9
266	CO2 emissions (kg per 2010 US$ of GDP)	9
267	CO2 emissions (kg per 2011 PPP $ of GDP)	9
268	CO2 emissions (kg per PPP $ of GDP)	9
269	CO2 emissions (metric tons per capita)	9
270	Manufacturing, value added (% of GDP)	9
271	Manufacturing, value added (current US$)	9
272	Medium and high-tech Industry (including construction) (% manufacturing value added)	9
273	Railways, goods transported (million ton-km)	9
274	Railways, passengers carried (million passenger-km)	9
275	Research and development expenditure (% of GDP)	9
276	Researchers in R&D (per million people)	9
277	Annualized average growth rate in per capita real survey mean consumption or income, bottom 40% of population (%)	10
278	Annualized average growth rate in per capita real survey mean consumption or income, total population (%)	10
279	Average transaction cost of sending remittances from a specific country (%)	10
280	Average transaction cost of sending remittances to a specific country (%)	10
281	Net official development assistance received (constant 2015 US$)	10
282	Net official development assistance received (current US$)	10

283	PM2.5 air pollution, mean annual exposure (micrograms per cubic meter)	11
284	PM2.5 air pollution, population exposed to levels exceeding WHO guideline value (% of total)	11
285	PM2.5 pollution, population exposed to levels exceeding WHO Interim Target-1 value (% of total)	11
286	PM2.5 pollution, population exposed to levels exceeding WHO Interim Target-2 value (% of total)	11
287	PM2.5 pollution, population exposed to levels exceeding WHO Interim Target-3 value (% of total)	11
288	Population living in slums (% of urban population)	11
289	Urban population	11
290	Urban population (% of total population)	11
291	Urban population growth (annual %)	11
292	Adjusted net savings, excluding particulate emission damage (% of GNI)	12
293	Coal rents (% of GDP)	12
294	Forest rents (% of GDP)	12
295	Mineral rents (% of GDP)	12
296	Natural gas rents (% of GDP)	12
297	Oil rents (% of GDP)	12
298	Total natural resources rents (% of GDP)	12
299	Aquaculture production (metric tons)	14
300	Capture fisheries production (metric tons)	14
301	Total fisheries production (metric tons)	14
302	Forest area (% of land area)	15
303	Forest area (sq. km)	15
304	Terrestrial and marine protected areas (% of total territorial area)	15
305	Terrestrial protected areas (% of total land area)	15
306	Battle-related deaths (number of people)	16
307	Bribery incidence (% of firms experiencing at least one bribe payment request)	16
308	Completeness of birth registration (%)	16

309	Completeness of birth registration, female (%)	16
310	Completeness of birth registration, male (%)	16
311	Completeness of birth registration, rural (%)	16
312	Completeness of birth registration, urban (%)	16
313	Firms expected to give gifts in meetings with tax officials (% of firms)	16
314	Intentional homicides (per 100,000 people)	16
315	Intentional homicides, female (per 100,000 female)	16
316	Intentional homicides, male (per 100,000 male)	16
317	Primary government expenditures as a proportion of original approved budget (%)	16
318	Debt service (PPG and IMF only, % of exports of goods, services and primary income)	17
319	Exports of goods and services (% of GDP)	17
320	Foreign direct investment, net inflows (% of GDP)	17
321	Foreign direct investment, net inflows (BoP, current US$)	17
322	GDP (constant 2010 US$)	17
323	GDP (constant LCU)	17
324	GDP (current LCU)	17
325	GDP (current US$)	17
326	GDP, PPP (constant 2011 international $)	17
327	GDP, PPP (current international $)	17
328	GNI (constant 2010 US$)	17
329	GNI (constant LCU)	17
330	GNI, PPP (constant 2011 international $)	17
331	GNI, PPP (current international $)	17
332	Individuals using the Internet (% of population)	17
333	Investment in energy with private participation (current US$)	17
334	Investment in transport with private participation (current US$)	17

335	Investment in water and sanitation with private participation (current US$)	17
336	Methodology assessment of statistical capacity (scale 0 - 100)	17
337	Net official development assistance and official aid received (current US$)	17
338	Patent applications, nonresidents	17
339	Patent applications, residents	17
340	Periodicity and timeliness assessment of statistical capacity (scale 0 - 100)	17
341	Personal remittances, received (% of GDP)	17
342	Source data assessment of statistical capacity (scale 0 - 100)	17
343	Statistical Capacity score (Overall average)	17
344	Tariff rate, applied, simple mean, all products (%)	17
345	Tariff rate, applied, simple mean, manufactured products (%)	17
346	Tariff rate, applied, simple mean, primary products (%)	17
347	Tariff rate, applied, weighted mean, all products (%)	17
348	Tariff rate, applied, weighted mean, manufactured products (%)	17
349	Tariff rate, applied, weighted mean, primary products (%)	17
350	Tax revenue (% of GDP)	17
351	Tax revenue (current LCU)	17

Table A-2. The 147 identified principal indicators.

Index	SDG indicator	SDG
23	Cereal yield (kg per hectare)	2
25	Prevalence of anemia among women of reproductive age (% of women ages 15-49)	2
26	Prevalence of HIV, female (% ages 15-24)	2
27	Prevalence of HIV, male (% ages 15-24)	2
28	Prevalence of HIV, total (% of population ages 15-49)	2
38	Prevalence of undernourishment (% of population)	2
45	Adolescent fertility rate (births per 1,000 women ages 15-19)	3
48	Immunization, DPT (% of children ages 12-23 months)	3
49	Immunization, HepB3 (% of one-year-old children)	3
50	Immunization, measles (% of children ages 12-23 months)	3
51	Incidence of HIV (per 1,000 uninfected population ages 15-49)	3
53	Incidence of tuberculosis (per 100,000 people)	3
54	Maternal mortality ratio (modeled estimate, per 100,000 live births)	3
55	Mortality caused by road traffic injury (per 100,000 people)	3
57	Mortality from CVD, cancer, diabetes or CRD between exact ages 30 and 70, female (%)	3
58	Mortality from CVD, cancer, diabetes or CRD between exact ages 30 and 70, male (%)	3
60	Mortality rate attributed to household and ambient air pollution, age-standardized, female (per 100,000 female population)	3
61	Mortality rate attributed to household and ambient air pollution, age-standardized, male (per 100,000 male population)	3
62	Mortality rate attributed to unintentional poisoning (per 100,000 population)	3
63	Mortality rate attributed to unintentional poisoning, female (per 100,000 female population)	3
65	Mortality rate attributed to unsafe water, unsafe sanitation and lack of hygiene (per 100,000 population)	3
66	Mortality rate, neonatal (per 1,000 live births)	3
67	Mortality rate, under-5 (per 1,000 live births)	3
76	Smoking prevalence, females (% of adults)	3
77	Smoking prevalence, males (% of adults)	3

78	Suicide mortality rate (per 100,000 population)	3
79	Suicide mortality rate, female (per 100,000 female population)	3
80	Suicide mortality rate, male (per 100,000 male population)	3
87	Children out of school (% of primary school age)	4
90	Children out of school, primary	4
93	Compulsory education, duration (years)	4
123	Lower secondary completion rate, female (% of relevant age group)	4
124	Lower secondary completion rate, male (% of relevant age group)	4
126	Over-age students, primary (% of enrollment)	4
127	Over-age students, primary, female (% of female enrollment)	4
129	Preprimary education, duration (years)	4
130	Primary completion rate, female (% of relevant age group)	4
132	Primary completion rate, total (% of relevant age group)	4
133	Primary education, duration (years)	4
136	Pupil-teacher ratio, primary	4
140	School enrollment, preprimary (% gross)	4
141	School enrollment, preprimary, female (% gross)	4
143	School enrollment, primary (gross), gender parity index (GPI)	4
144	School enrollment, primary and secondary (gross), gender parity index (GPI)	4
145	School enrollment, secondary (gross), gender parity index (GPI)	4
146	School enrollment, tertiary (% gross)	4
147	School enrollment, tertiary (gross), gender parity index (GPI)	4
148	School enrollment, tertiary, female (% gross)	4
149	School enrollment, tertiary, male (% gross)	4
150	Secondary education, duration (years)	4
166	Contributing family workers, female (% of female employment) (modeled ILO estimate)	5

167	Contributing family workers, male (% of male employment) (modeled ILO estimate)	5
172	Nondiscrimination clause mentions gender in the constitution (1=yes; 0=no)	5
173	Proportion of seats held by women in national parliaments (%)	5
186	People practicing open defecation (% of population)	6
187	People practicing open defecation, rural (% of rural population)	6
188	People practicing open defecation, urban (% of urban population)	6
189	People using at least basic drinking water services (% of population)	6
190	People using at least basic drinking water services, rural (% of rural population)	6
191	People using at least basic drinking water services, urban (% of urban population)	6
192	People using at least basic sanitation services (% of population)	6
193	People using at least basic sanitation services, rural (% of rural population)	6
194	People using at least basic sanitation services, urban (% of urban population)	6
195	People using safely managed drinking water services (% of population)	6
207	Access to clean fuels and technologies for cooking (% of population)	7
208	Access to electricity (% of population)	7
209	Access to electricity, rural (% of rural population)	7
210	Access to electricity, urban (% of urban population)	7
211	Energy intensity level of primary energy (MJ/$2011 PPP GDP)	7
212	Renewable electricity output (% of total electricity output)	7
213	Renewable energy consumption (% of total final energy consumption)	7
223	Agriculture, forestry, and fishing, value added per worker (constant 2010 US$)	8
227	Commercial bank branches (per 100,000 adults)	8
230	Employment in agriculture, male (% of male employment) (modeled ILO estimate)	8
231	Employment in industry (% of total employment) (modeled ILO estimate)	8
232	Employment in industry, female (% of female employment) (modeled ILO estimate)	8
233	Employment in industry, male (% of male employment) (modeled ILO estimate)	8

235	Employment in services, female (% of female employment) (modeled ILO estimate)	8
236	Employment in services, male (% of male employment) (modeled ILO estimate)	8
237	GDP growth (annual %)	8
238	GDP per capita (constant 2010 US$)	8
239	GDP per person employed (constant 2011 PPP $)	8
240	Industry (including construction), value added per worker (constant 2010 US$)	8
244	New business density (new registrations per 1,000 people ages 15-64)	8
245	Services, value added per worker (constant 2010 US$)	8
249	Unemployment, female (% of female labor force) (modeled ILO estimate)	8
251	Unemployment, male (% of male labor force) (modeled ILO estimate)	8
255	Unemployment, youth female (% of female labor force ages 15-24) (modeled ILO estimate)	8
259	Unemployment, youth total (% of total labor force ages 15-24) (modeled ILO estimate)	8
261	Wage and salaried workers, female (% of female employment) (modeled ILO estimate)	8
262	Wage and salaried workers, male (% of male employment) (modeled ILO estimate)	8
264	Air transport, freight (million ton-km)	9
265	Air transport, passengers carried	9
266	CO2 emissions (kg per 2010 US$ of GDP)	9
268	CO2 emissions (kg per PPP $ of GDP)	9
270	Manufacturing, value added (% of GDP)	9
271	Manufacturing, value added (current US$)	9
272	Medium and high-tech Industry (including construction) (% manufacturing value added)	9
281	Net official development assistance received (constant 2015 US$)	10
283	PM2.5 air pollution, mean annual exposure (micrograms per cubic meter)	11
284	PM2.5 air pollution, population exposed to levels exceeding WHO guideline value (% of total)	11
285	PM2.5 pollution, population exposed to levels exceeding WHO Interim Target-1 value (% of total)	11
286	PM2.5 pollution, population exposed to levels exceeding WHO Interim Target-2 value (% of total)	11

287	PM2.5 pollution, population exposed to levels exceeding WHO Interim Target-3 value (% of total)	11
289	Urban population	11
290	Urban population (% of total population)	11
291	Urban population growth (annual %)	11
292	Adjusted net savings, excluding particulate emission damage (% of GNI)	12
293	Coal rents (% of GDP)	12
294	Forest rents (% of GDP)	12
295	Mineral rents (% of GDP)	12
296	Natural gas rents (% of GDP)	12
297	Oil rents (% of GDP)	12
298	Total natural resources rents (% of GDP)	12
299	Aquaculture production (metric tons)	14
300	Capture fisheries production (metric tons)	14
301	Total fisheries production (metric tons)	14
302	Forest area (% of land area)	15
303	Forest area (sq. km)	15
304	Terrestrial and marine protected areas (% of total territorial area)	15
305	Terrestrial protected areas (% of total land area)	15
318	Debt service (PPG and IMF only, % of exports of goods, services and primary income)	17
319	Exports of goods and services (% of GDP)	17
320	Foreign direct investment, net inflows (% of GDP)	17
321	Foreign direct investment, net inflows (BoP, current US$)	17
322	GDP (constant 2010 US$)	17
323	GDP (constant LCU)	17
324	GDP (current LCU)	17
328	GNI (constant 2010 US$)	17

329	GNI (constant LCU)	17
330	GNI, PPP (constant 2011 international $)	17
331	GNI, PPP (current international $)	17
332	Individuals using the Internet (% of population)	17
336	Methodology assessment of statistical capacity (scale 0 - 100)	17
337	Net official development assistance and official aid received (current US$)	17
338	Patent applications, nonresidents	17
340	Periodicity and timeliness assessment of statistical capacity (scale 0 - 100)	17
341	Personal remittances, received (% of GDP)	17
342	Source data assessment of statistical capacity (scale 0 - 100)	17
343	Statistical Capacity score (Overall average)	17
344	Tariff rate, applied, simple mean, all products (%)	17
346	Tariff rate, applied, simple mean, primary products (%)	17
347	Tariff rate, applied, weighted mean, all products (%)	17
348	Tariff rate, applied, weighted mean, manufactured products (%)	17
349	Tariff rate, applied, weighted mean, primary products (%)	17
350	Tax revenue (% of GDP)	17
351	Tax revenue (current LCU)	17

Table A-3. Average missing rate of indicators for each country or region.

Country or region	Average missing rate of all indicators	Average missing rate of principal indicators	Country or region	Average missing rate of all indicators	Average missing rate of principal indicators
Afghanistan	63.40%	32.10%	**Latvia**	46.50%	17.70%
Albania	49.20%	17.30%	**Lebanon**	62.10%	29.90%
Algeria	51.80%	18.40%	**Lesotho**	53.50%	19.50%
American Samoa	90.90%	79.90%	**Liberia**	60.10%	30.20%
Andorra	79.40%	64.80%	**Libya**	70.40%	42.60%
Angola	59.90%	24.90%	**Liechtenstein**	84.80%	71.60%
Antigua and Barbuda	71.60%	47.00%	**Lithuania**	47.90%	19.60%
Argentina	47.20%	14.20%	**Luxembourg**	50.80%	20.00%
Armenia	46.40%	18.60%	**Macao SAR, China**	57.50%	39.10%
Aruba	76.70%	61.10%	**Madagascar**	53.50%	17.90%
Australia	54.60%	26.60%	**Malawi**	54.00%	20.70%
Austria	50.30%	21.80%	**Malaysia**	47.00%	14.10%
Azerbaijan	47.50%	20.00%	**Maldives**	58.00%	27.20%
Bahamas, The	59.10%	31.10%	**Mali**	51.40%	19.40%
Bahrain	59.20%	32.00%	**Malta**	53.20%	23.80%
Bangladesh	44.10%	18.20%	**Marshall Islands**	76.20%	52.30%
Barbados	58.20%	29.70%	**Mauritania**	54.90%	23.60%
Belarus	46.00%	14.50%	**Mauritius**	49.10%	13.00%
Belgium	50.40%	21.00%	**Mexico**	38.90%	12.40%
Belize	49.30%	15.80%	**Micronesia, Fed. Sts.**	80.20%	59.90%
Benin	55.10%	20.10%	**Moldova**	41.80%	13.00%
Bermuda	77.50%	62.70%	**Monaco**	86.00%	72.90%
Bhutan	53.40%	22.00%	**Mongolia**	45.20%	17.00%
Bolivia	47.80%	18.40%	**Montenegro**	57.90%	29.80%

Bosnia and Herzegovina	54.90%	27.40%	**Morocco**	48.10%	13.20%
Botswana	54.60%	19.30%	**Mozambique**	50.50%	17.70%
Brazil	45.50%	19.80%	**Myanmar**	53.90%	23.20%
British Virgin Islands	87.50%	78.30%	**Namibia**	62.00%	29.60%
Brunei Darussalam	60.30%	29.70%	**Nauru**	82.70%	66.50%
Bulgaria	45.80%	15.20%	**Nepal**	49.70%	16.10%
Burkina Faso	49.10%	16.10%	**Netherlands**	51.80%	22.40%
Burundi	55.00%	22.40%	**New Caledonia**	86.30%	73.10%
Cabo Verde	52.20%	18.20%	**New Zealand**	55.80%	24.00%
Cambodia	48.10%	17.50%	**Nicaragua**	53.10%	22.40%
Cameroon	53.20%	19.70%	**Niger**	49.30%	19.10%
Canada	57.70%	27.70%	**Nigeria**	52.80%	21.40%
Cayman Islands	84.30%	78.30%	**North Macedonia**	49.00%	16.80%
Central African Republic	62.10%	29.70%	**Northern Mariana Islands**	91.50%	80.80%
Chad	57.70%	26.20%	**Norway**	49.80%	18.40%
Channel Islands	89.60%	80.10%	**Oman**	57.60%	24.10%
Chile	46.90%	16.30%	**Pakistan**	44.30%	16.30%
China	53.70%	23.40%	**Palau**	78.20%	56.00%
Colombia	38.20%	13.80%	**Panama**	45.00%	17.00%
Comoros	63.10%	30.40%	**Papua New Guinea**	64.60%	31.50%
Congo, Dem. Rep.	58.30%	28.30%	**Paraguay**	45.80%	16.70%
Congo, Rep.	57.30%	21.70%	**Peru**	42.00%	14.60%
Costa Rica	44.60%	16.30%	**Philippines**	44.60%	15.30%
Cote d'Ivoire	52.90%	21.40%	**Poland**	45.80%	19.50%
Croatia	48.20%	15.60%	**Portugal**	48.10%	19.20%
Cuba	52.20%	24.50%	**Puerto Rico**	71.00%	52.70%

Curacao	91.70%	83.10%	**Qatar**	57.00%	30.80%
Cyprus	50.10%	20.10%	**Romania**	43.70%	14.90%
Czech Republic	48.60%	17.90%	**Russian Federation**	51.00%	21.00%
Denmark	51.00%	19.20%	**Rwanda**	52.10%	20.30%
Djibouti	62.00%	33.10%	**Samoa**	61.90%	29.40%
Dominica	68.80%	47.70%	**San Marino**	84.60%	70.30%
Dominican Republic	41.10%	17.80%	**Sao Tome and Principe**	61.50%	31.90%
Ecuador	40.10%	16.50%	**Saudi Arabia**	58.70%	30.40%
Egypt, Arab Rep.	48.20%	16.30%	**Senegal**	47.30%	17.00%
El Salvador	42.00%	14.70%	**Serbia**	46.70%	22.10%
Equatorial Guinea	65.00%	34.70%	**Seychelles**	68.50%	42.50%
Eritrea	59.50%	33.00%	**Sierra Leone**	56.90%	26.90%
Estonia	48.50%	16.00%	**Singapore**	56.30%	30.50%
Eswatini	55.30%	21.90%	**Sint Maarten (Dutch part)**	93.90%	86.80%
Ethiopia	51.90%	20.40%	**Slovak Republic**	45.40%	14.40%
Faroe Islands	91.90%	82.10%	**Slovenia**	47.90%	20.00%
Fiji	60.30%	27.90%	**Solomon Islands**	65.30%	35.60%
Finland	48.30%	17.20%	**Somalia**	75.10%	52.60%
France	48.70%	18.10%	**South Africa**	46.10%	18.00%
French Polynesia	86.40%	73.80%	**South Sudan**	77.50%	58.90%
Gabon	62.50%	28.10%	**Spain**	48.20%	18.20%
Gambia, The	56.10%	23.90%	**Sri Lanka**	49.40%	17.60%
Georgia	47.50%	16.60%	**St. Kitts and Nevis**	72.70%	51.10%
Germany	50.30%	20.10%	**St. Lucia**	56.30%	25.00%
Ghana	46.90%	18.40%	**St. Martin (French part)**	98.40%	96.40%

Gibraltar	90.10%	82.00%	**St. Vincent and the Grenadines**	62.70%	32.00%
Greece	50.60%	21.70%	**Sudan**	58.60%	24.90%
Greenland	85.40%	72.10%	**Suriname**	57.40%	25.70%
Grenada	70.00%	46.20%	**Sweden**	49.10%	20.90%
Guam	83.10%	66.40%	**Switzerland**	53.00%	20.90%
Guatemala	47.30%	16.70%	**Syrian Arab Republic**	62.90%	35.70%
Guinea	56.60%	24.50%	**Tajikistan**	52.10%	19.90%
Guinea-Bissau	63.40%	31.60%	**Tanzania**	50.70%	19.00%
Guyana	58.10%	26.50%	**Thailand**	50.00%	19.20%
Haiti	62.70%	30.30%	**Timor-Leste**	61.40%	31.60%
Honduras	47.40%	19.30%	**Togo**	53.60%	19.40%
Hong Kong SAR, China	60.70%	38.50%	**Tonga**	64.30%	32.20%
Hungary	45.30%	15.50%	**Trinidad and Tobago**	61.40%	34.80%
Iceland	54.30%	21.50%	**Tunisia**	50.90%	16.00%
India	50.30%	18.80%	**Turkey**	42.20%	22.10%
Indonesia	42.60%	15.40%	**Turkmenistan**	67.40%	40.60%
Iran, Islamic Rep.	49.60%	17.30%	**Turks and Caicos Islands**	90.10%	81.10%
Iraq	61.00%	36.30%	**Tuvalu**	79.20%	61.90%
Ireland	52.90%	19.30%	**Uganda**	52.10%	19.70%
Isle of Man	92.50%	84.70%	**Ukraine**	48.10%	15.30%
Israel	50.00%	16.40%	**United Arab Emirates**	62.80%	35.30%
Italy	49.40%	18.00%	**United Kingdom**	51.70%	22.20%
Jamaica	52.10%	19.10%	**United States**	50.30%	24.40%
Japan	59.00%	31.10%	**Uruguay**	45.20%	18.40%

Jordan	49.10%	15.50%	**Uzbekistan**	51.50%	19.30%
Kazakhstan	48.30%	15.60%	**Vanuatu**	63.00%	28.40%
Kenya	52.70%	20.30%	**Venezuela, RB**	54.70%	27.80%
Kiribati	75.90%	53.40%	**Vietnam**	48.10%	19.50%
Korea, Dem. People's Rep.	76.10%	56.00%	**Virgin Islands (U.S.)**	84.00%	68.20%
Korea, Rep.	54.10%	24.40%	**West Bank and Gaza**	48.80%	30.30%
Kosovo	87.20%	80.90%	**Yemen, Rep.**	58.90%	26.10%
Kuwait	51.80%	26.90%	**Zambia**	56.50%	22.50%
Kyrgyz Republic	44.20%	13.50%	**Zimbabwe**	58.00%	27.10%
Lao PDR	47.90%	17.30%			

Appendix B. Supporting Information for Chapter 3

Supplementary Note B-1: Example of scale-independent comparison.

The development of nations is measured by a variety of economic, social, and environmental indicators. Many of these indicators that characterize the size of stocks in a country, such as gross domestic product (GDP), are often normalized by population to compare among countries. Whether such comparison using population-normalized indicators is appropriate, however, has long been debated [121, 122]. The underlying assumption that warrants the comparison using population-normalized indicators is that those indicators scale linearly with population. For some indicators, this assumption may not be valid because it ignores the aggregation effect resulting from non-linear interactions in socioeconomic systems [25, 96]. To better evaluate and monitor the development of countries, it is imperative to understand how some of the stock indicators scale non-linearly with population and develop adjusted normalization approaches to taking into account the non-linear relationship.

I re-rank countries using scale-independent GDP (Methods) and compare the ranking using GDP per capita. The results show distinct differences for most countries in the two rankings (Table S3). Note that China has the largest improvement using scale-independent GDP (53th) from the ranking using per-capita GDP (77th), and Nauru owns the largest decline (113th from 78th). These results provide important insights on how a country performs in national development compared to what it should be given its population.

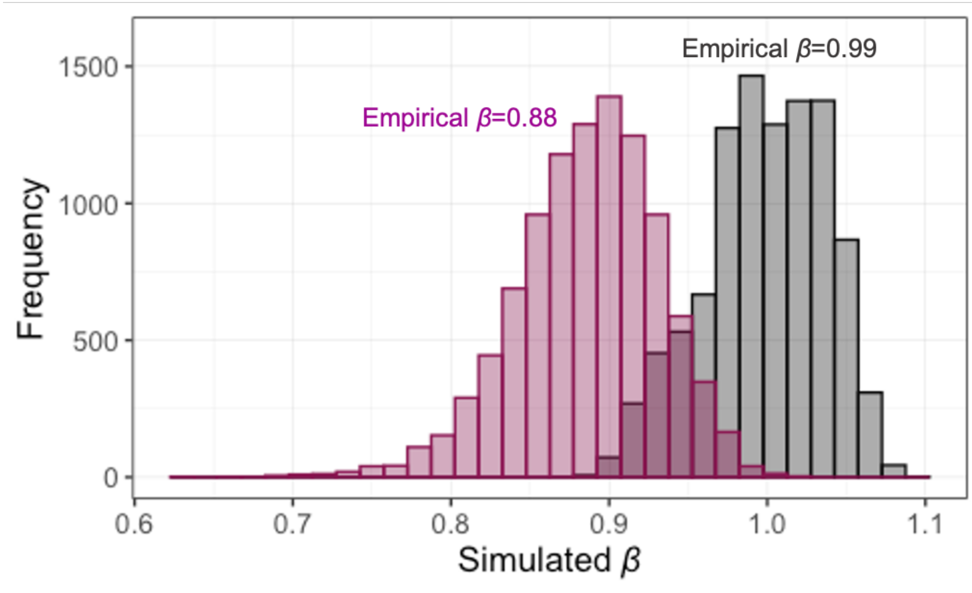

Figure B-1. Histogram of β for socioeconomic activity indicators and individual need indicators in countries from 10,000 simulations.
Median value of the simulated β for socioeconomic activity (red) and individual need (grey) indicators are 0.89 and 0.99, respectively. Distribution of parameters is uniform.

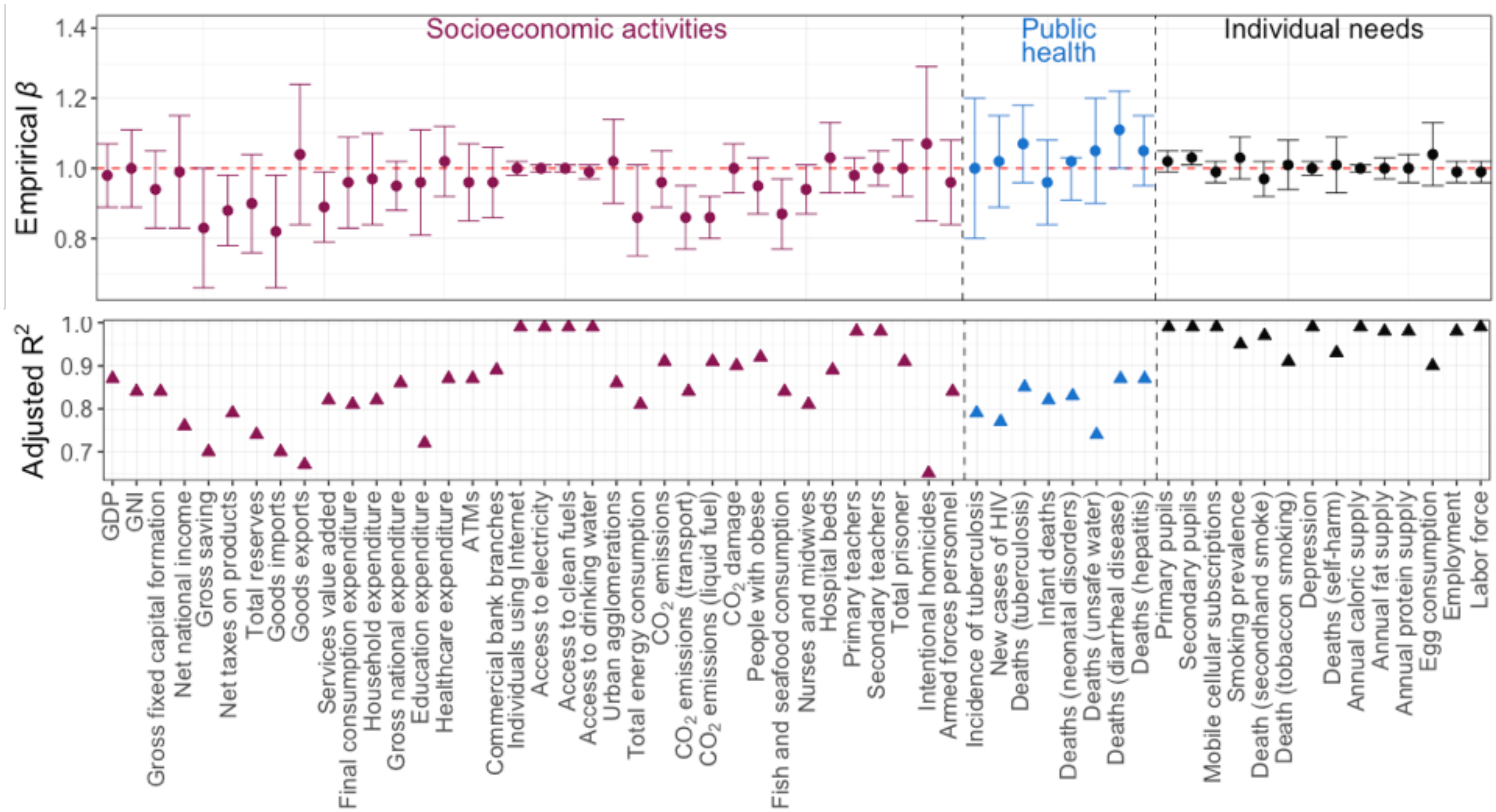

Figure B-2. Empirical scaling exponents for indicators of socioeconomic activities, public health, and individual needs for highly urbanized countries. Dash line shows the linear scaling ($\beta = 1$).

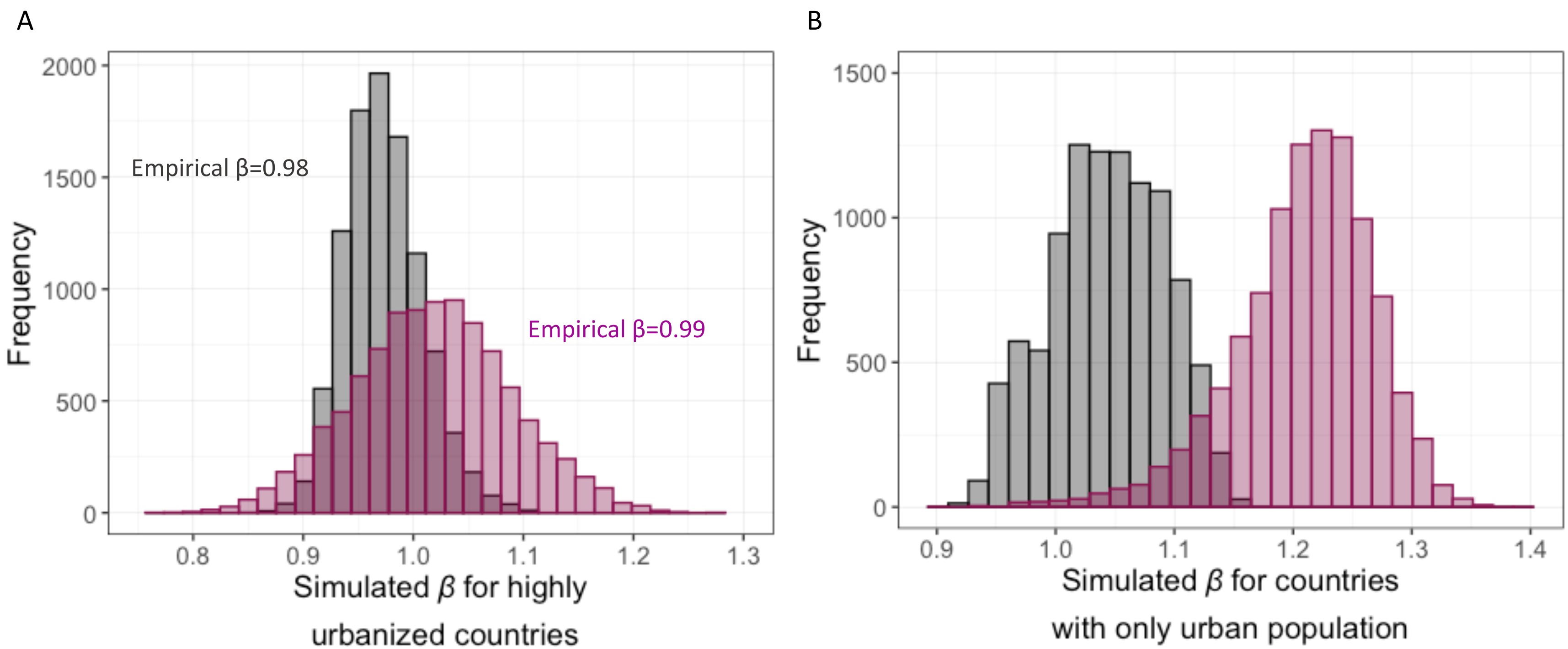

Figure B-3. Histogram of simulated β for socioeconomic activity indicators and individual need indicators from 10,000 simulations. Distributions of parameters in is uniform. (A) result for highly urbanized countries. (B) result for all countries with only urban population.

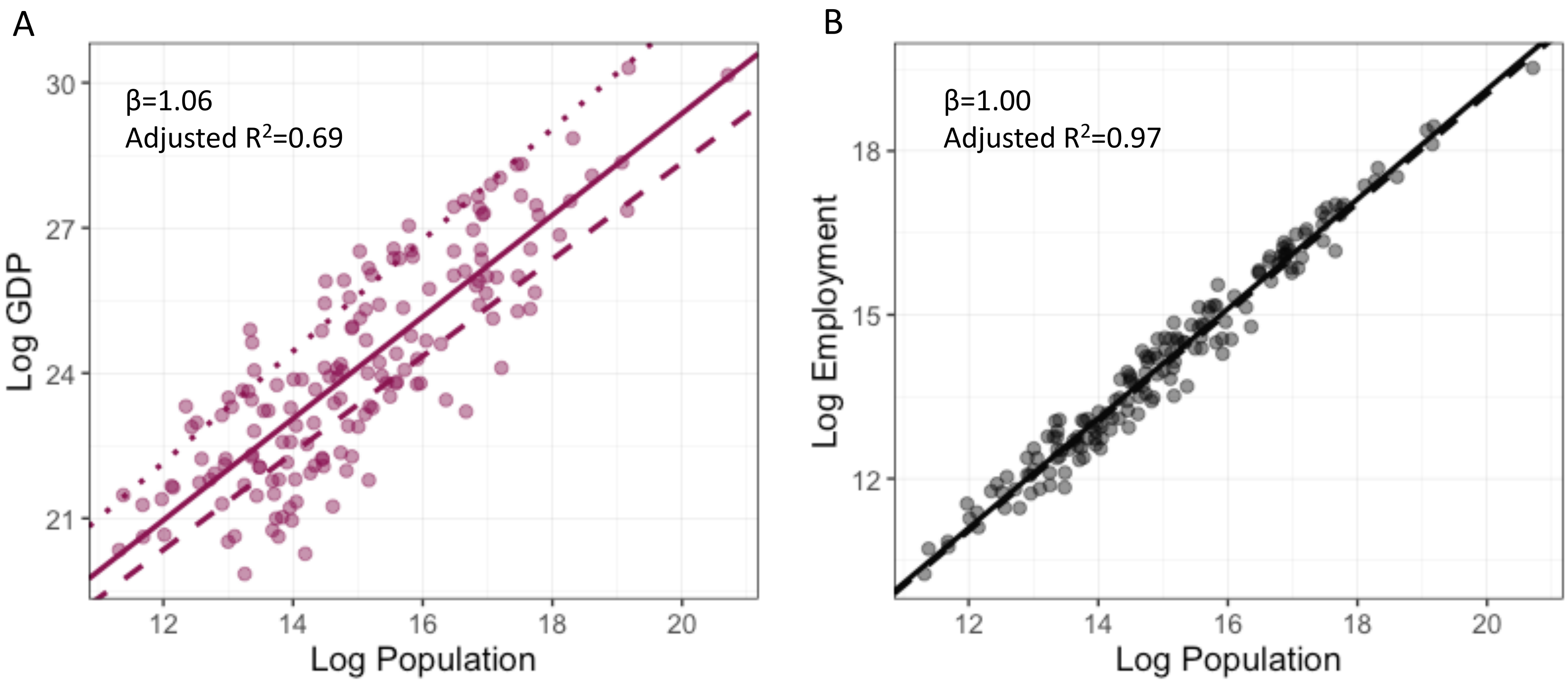

Figure B-4. Scaling relations in countries with only cities.
Solid line shows the best-fit relation, dash line shows the linear scaling, and dotted line shows the scaling in cities. (A) GDP vs. population in 2019. (B) employment vs. population in 2019. Data are from Oxford Economics database [84].

Table B-1. List of 213 countries and regions.

Afghanistan	Dominican Republic	Lesotho	Sao Tome and Principe
Albania	Ecuador	Liberia	Saudi Arabia
Algeria	Egypt, Arab Rep.	Libya	Senegal
Andorra	El Salvador	Liechtenstein	Serbia
Angola	Equatorial Guinea	Lithuania	Seychelles
Antigua and Barbuda	Eritrea	Luxembourg	Sierra Leone
Argentina	Estonia	Macao SAR, China	Singapore
Armenia	Ethiopia	Macedonia, FYR	Sint Maarten (Dutch part)
Aruba	Faroe Islands	Madagascar	Slovak Republic
Australia	Fiji	Malawi	Slovenia
Austria	Finland	Malaysia	Solomon Islands
Azerbaijan	France	Maldives	Somalia
Bahamas, The	French Polynesia	Mali	South Africa
Bahrain	Gabon	Malta	South Sudan
Bangladesh	Gambia, The	Marshall Islands	Spain
Barbados	Georgia	Mauritania	Sri Lanka
Belarus	Germany	Mauritius	St. Kitts and Nevis
Belgium	Ghana	Mexico	St. Lucia
Belize	Gibraltar	Micronesia, Fed. Sts.	St. Vincent and the Grenadines
Benin	Greece	Moldova	Sudan
Bermuda	Greenland	Monaco	Suriname
Bhutan	Grenada	Mongolia	Swaziland
Bolivia	Guam	Montenegro	Sweden
Bosnia and Herzegovina	Guatemala	Morocco	Switzerland
Botswana	Guinea	Mozambique	Syrian Arab Republic
Brazil	Guinea-Bissau	Myanmar	Tajikistan
British Virgin Islands	Guyana	Namibia	Tanzania
Brunei Darussalam	Haiti	Nauru	Thailand
Bulgaria	Honduras	Nepal	Timor-Leste
Burkina Faso	Hong Kong SAR, China	Netherlands	Togo
Burundi	Hungary	New Caledonia	Tonga
Cabo Verde	Iceland	New Zealand	Trinidad and Tobago
Cambodia	India	Nicaragua	Tunisia
Cameroon	Indonesia	Niger	Turkey
Canada	Iran, Islamic Rep.	Nigeria	Turkmenistan

Cayman Islands	Iraq	Northern Mariana Islands	Turks and Caicos Islands
Central African Republic	Ireland	Norway	Tuvalu
Chad	Isle of Man	Oman	Uganda
Chile	Israel	Pakistan	Ukraine
China	Italy	Palau	United Arab Emirates
Colombia	Jamaica	Panama	United Kingdom
Comoros	Japan	Papua New Guinea	United States
Congo, Dem. Rep.	Jordan	Paraguay	Uruguay
Congo, Rep.	Kazakhstan	Peru	Uzbekistan
Costa Rica	Kenya	Philippines	Vanuatu
Cote d'Ivoire	Kiribati	Poland	Venezuela, RB
Croatia	Korea, Dem. People's Rep.	Portugal	Vietnam
Cuba	Korea, Rep.	Puerto Rico	Virgin Islands (U.S.)
Curacao	Kosovo	Qatar	Yemen, Rep.
Cyprus	Kuwait	Romania	Zambia
Czech Republic	Kyrgyz Republic	Russian Federation	Zimbabwe
Denmark	Lao PDR	Rwanda	
Djibouti	Latvia	Samoa	
Dominica	Lebanon	San Marino	

Table B-2. Summary of empirical exponents for national development indicators.

	National development indicator (unit)	Exponent	90% CI	Adj-R^2	Observation	Year
Socioeconomic activities	GDP (constant 2010 US$)	0.88	[0.76, 0.96]	0.64	183	2019
	GNI (constant 2010 US$)	0.92	[0.77, 1.03]	0.58	138	2019
	Gross fixed capital formation (constant 2010 US$)	0.85	[0.74, 0.96]	0.51	144	2018
	Net national income (constant 2010 US$)	0.92	[0.80, 1.04]	0.52	131	2018
	Gross saving (constant 2010 US$)	0.87	[0.74, 0.99]	0.51	145	2018
	Net taxes on products (constant 2010 US$)	0.78	[0.69, 0.87]	0.55	171	2018
	Total reserves (constant 2010 US$)	0.85	[0.75, 0.95]	0.55	153	2018
	Goods imports (constant 2010 US$)	0.77	[0.66, 0.88]	0.53	138	2019
	Goods exports (constant 2010 US$)	0.94	[0.80, 1.04]	0.5	138	2019
	Services, value added (constant 2010 US$)	0.83	[0.72, 0.94]	0.53	158	2019
	Final consumption expenditure (constant 2010 US$)	0.86	[0.75, 0.97]	0.52	137	2019
	Household expenditure (constant 2010 US$)	0.88	[0.77, 0.99]	0.54	146	2019
	Gross national expenditure (constant 2010 US$)	0.88	[0.78, 0.98]	0.58	145	2018
	Education expenditure (constant 2010 US$)	0.86	[0.76, 0.96]	0.53	175	2018
	Healthcare expenditure (constant 2010 US$)	0.87	[0.79, 0.95]	0.6	181	2018
	ATMs (number)	0.91	[0.82, 1.00]	0.66	142	2019
	Commercial bank branches (number)	0.88	[0.82, 0.94]	0.77	142	2019
	Individuals using the Internet (number)	0.94	[0.90, 0.98]	0.89	205	2019
	Access to electricity (person)	0.96	[0.94, 0.98]	0.95	211	2018
	Access to clean fuels (person)	0.92	[0.85, 0.99]	0.69	186	2016
	Access to drinking water (person)	0.93	[0.89, 0.97]	0.93	106	2017
	Urban agglomerations (person)	0.89	[0.81, 0.97]	0.75	121	2019
	Total energy consumption (kg)	0.82	[0.73, 0.91]	0.65	134	2014
	CO2 emissions (kg)	0.88	[0.80, 0.96]	0.66	200	2016
	CO2 emissions from transport (kg)	0.74	[0.64, 0.84]	0.57	140	2014

	CO2 emissions from liquid fuel (kg)	0.76	[0.70, 0.82]	0.67	199	2016
	CO2 damage (constant 2010 US$)	0.93	[0.85, 1.01]	0.67	188	2018
	People with obese (person)	0.88	[0.84, 0.92]	0.84	178	2016
	Fish and seafood consumption (kg)	0.87	[0.80, 0.94]	0.72	169	2017
	Nurses and midwives (person)	0.88	[0.80, 0.96]	0.78	111	2018
	Hospital beds (number)	0.93	[0.87, 0.99]	0.84	100	2017
	Primary teachers (person)	0.94	[0.92, 0.96]	0.97	130	2018
	Secondary teachers (person)	0.94	[0.90, 0.98]	0.94	111	2018
	Total prisoner (person)	0.92	[0.90,0.94]	0.84	144	2014
	Intentional homicides (number)	0.92	[0.81, 1.03]	0.65	111	2017
	Armed forces personnel (person)	0.93	[0.86, 0.99]	0.71	165	2018
Public health	Incidence of tuberculosis (person)	1.13	[1.06, 1.20]	0.75	201	2019
	New cases of HIV (person)	1.05	[0.95, 1.15]	0.62	189	2017
	Deaths from tuberculosis (person)	1.15	[1.06, 1.24]	0.68	189	2017
	Infant deaths (person)	1.05	[0.97, 1.13]	0.69	189	2019
	Deaths from neonatal disorders (person)	1.07	[0.99, 1.15]	0.71	189	2017
	Deaths from unsafe water	1.16	[1.03, 1.29]	0.51	189	2017
	Deaths from diarrheal diseases	1.17	[1.07, 1.27]	0.66	186	2017
	Deaths from hepatitis (person)	1.09	[1.02, 1.16]	0.8	186	2017
Individual needs	Primary pupils (person)	1.01	[0.99, 1.03]	0.96	143	2018
	Secondary pupils (person)	1.01	[0.99, 1.03]	0.98	128	2018
	Mobile cellular subscriptions (number)	0.99	[0.97, 1.01]	0.96	206	2019
	Smoking prevalence (person)	0.98	[0.93, 1.03]	0.9	141	2018
	Death from secondhand smoke (person)	0.97	[0.94, 1.01]	0.91	189	2017
	Death from tobacco smoking (person)	0.97	[0.92, 1.03]	0.85	186	2017
	Prevalence of depression (person)	1.01	[1.00, 1.02]	0.99	187	2017

	Death from self-harm	0.97	[0.94, 1.01]	0.92	190	2017
	Annual caloric supply (kc)	1	[0.99, 1.01]	0.99	169	2013
	Annual fat supply (g)	0.97	[0.94, 1.00]	0.95	167	2013
	Annual protein supply (g)	1	[0.98, 1.02]	0.98	171	2017
	Egg consumption (kg)	1.02	[0.95, 1.08]	0.77	169	2017
	Employment (person)	0.99	[0.98, 1.00]	0.98	183	2019
	Labor force (person)	0.99	[0.98, 1.00]	0.98	183	2019

Table B-3. Rankings of countries by scale-independent GDP and GDP per capita in year 2019.

Country	Rank based on scaling-independent GDP	Rank based on GDP per capita	Difference
Norway	1	3	-2
United States	2	11	-9
Luxembourg	3	1	2
Switzerland	4	5	-1
Ireland	5	4	1
Japan	6	18	-12
Australia	7	10	-3
Denmark	8	6	2
Germany	9	20	-11
Canada	10	14	-4
Netherlands	11	12	-1
Sweden	12	9	3
France	13	22	-9
Singapore	14	8	6
United Kingdom	15	23	-8
Qatar	16	7	9
Bermuda	17	2	15
Austria	18	16	2
Belgium	19	19	0
Finland	20	17	3
Italy	21	27	-6
United Arab Emirates	22	24	-2
Spain	23	29	-6
Macao SAR, China	24	13	11
Hong Kong SAR, China	25	26	-1
New Zealand	26	25	1
Israel	27	28	-1
Iceland	28	15	13
Korea, Rep.	29	34	-5
Kuwait	30	30	0
Andorra	31	21	10
Portugal	32	37	-5
Czech Republic	33	38	-5
Puerto Rico	34	35	-1

Greece	35	39	-4
Saudi Arabia	36	44	-8
Slovenia	37	36	1
Brunei Darussalam	38	31	7
Malta	39	32	7
Poland	40	47	-7
Bahamas, The	41	33	8
Slovak Republic	42	41	1
Turkey	43	54	-11
Cyprus	44	40	4
Hungary	45	46	-1
Bahrain	46	42	4
Russian Federation	47	60	-13
Estonia	48	43	5
Chile	49	55	-6
Lithuania	50	45	5
Brazil	51	65	-14
Croatia	52	50	2
China	53	77	-24
Malaysia	54	59	-5
Oman	55	56	-1
Mexico	56	67	-11
Latvia	57	49	8
Romania	58	61	-3
Uruguay	59	58	1
Kazakhstan	60	64	-4
Trinidad and Tobago	61	53	8
Argentina	62	69	-7
Panama	63	63	0
Barbados	64	51	13
Colombia	65	83	-18
St. Kitts and Nevis	66	48	18
Costa Rica	67	68	-1
Antigua and Barbuda	68	52	16
South Africa	69	84	-15
Seychelles	70	57	13
Bulgaria	71	74	-3

Mauritius	72	66	6
Thailand	73	89	-16
Dominican Republic	74	82	-8
Libya	75	79	-4
Gabon	76	73	3
Iran, Islamic Rep.	77	93	-16
Peru	78	90	-12
Equatorial Guinea	79	71	8
Serbia	80	85	-5
Botswana	81	80	1
Belarus	82	88	-6
Iraq	83	98	-15
Indonesia	84	108	-24
Montenegro	85	75	10
Maldives	86	76	10
Azerbaijan	87	94	-7
St. Lucia	88	70	18
Suriname	89	81	8
Palau	90	62	28
Algeria	91	106	-15
Lebanon	92	95	-3
Bosnia and Herzegovina	93	91	2
Ecuador	94	101	-7
Grenada	95	72	23
Paraguay	96	99	-3
Namibia	97	96	1
North Macedonia	98	97	1
Guyana	99	92	7
Albania	100	100	0
Tunisia	101	109	-8
Georgia	102	102	0
Philippines	103	121	-18
Sri Lanka	104	113	-9
Jamaica	105	103	2
Armenia	106	105	1
St. Vincent and the Grenadines	107	87	20
Morocco	108	119	-11

Egypt, Arab Rep.	109	126	-17
Dominica	110	86	24
India	111	134	-23
Mongolia	112	111	1
Nauru	113	78	35
Ukraine	114	125	-11
Kosovo	115	107	8
Guatemala	116	120	-4
Fiji	117	104	13
Angola	118	124	-6
El Salvador	119	118	1
Nigeria	120	131	-11
Jordan	121	122	-1
Moldova	122	117	5
Uzbekistan	123	130	-7
Belize	124	112	12
Cabo Verde	125	115	10
Vietnam	126	135	-9
Bolivia	127	128	-1
Tonga	128	110	18
Papua New Guinea	129	129	0
Samoa	130	116	14
Bhutan	131	123	8
Honduras	132	132	0
Ghana	133	136	-3
Congo, Rep.	134	133	1
Myanmar	135	144	-9
Cote d'Ivoire	136	142	-6
Vanuatu	137	127	10
Sudan	138	147	-9
Bangladesh	139	151	-12
Tuvalu	140	114	26
Zambia	141	143	-2
Lao PDR	142	137	5
Pakistan	143	156	-13
Cameroon	144	146	-2
Senegal	145	145	0

Nicaragua	146	139	7
Mauritania	147	140	7
Kenya	148	155	-7
Cambodia	149	152	-3
Benin	150	153	-3
Solomon Islands	151	141	10
Haiti	152	154	-2
Zimbabwe	153	157	-4
Tanzania	154	162	-8
Uganda	155	161	-6
Lesotho	156	149	7
Tajikistan	157	158	-1
Kiribati	158	138	20
Kyrgyz Republic	159	159	0
Comoros	160	148	12
Nepal	161	165	-4
Guinea	162	163	-1
Rwanda	163	164	-1
Burkina Faso	164	166	-2
Chad	165	168	-3
Mali	166	169	-3
Sao Tome and Principe	167	150	17
Ethiopia	168	173	-5
Timor-Leste	169	160	9
Yemen, Rep.	170	172	-2
Gambia, The	171	167	4
Togo	172	170	2
Afghanistan	173	175	-2
Mozambique	174	174	0
Niger	175	176	-1
Malawi	176	177	-1
Madagascar	177	179	-2
Congo, Dem. Rep.	178	181	-3
Guinea-Bissau	179	171	8
Liberia	180	178	2
Sierra Leone	181	180	1
Central African Republic	182	182	0

Burundi	183	183	0

Appendix C. Supporting Information for Chapter 4

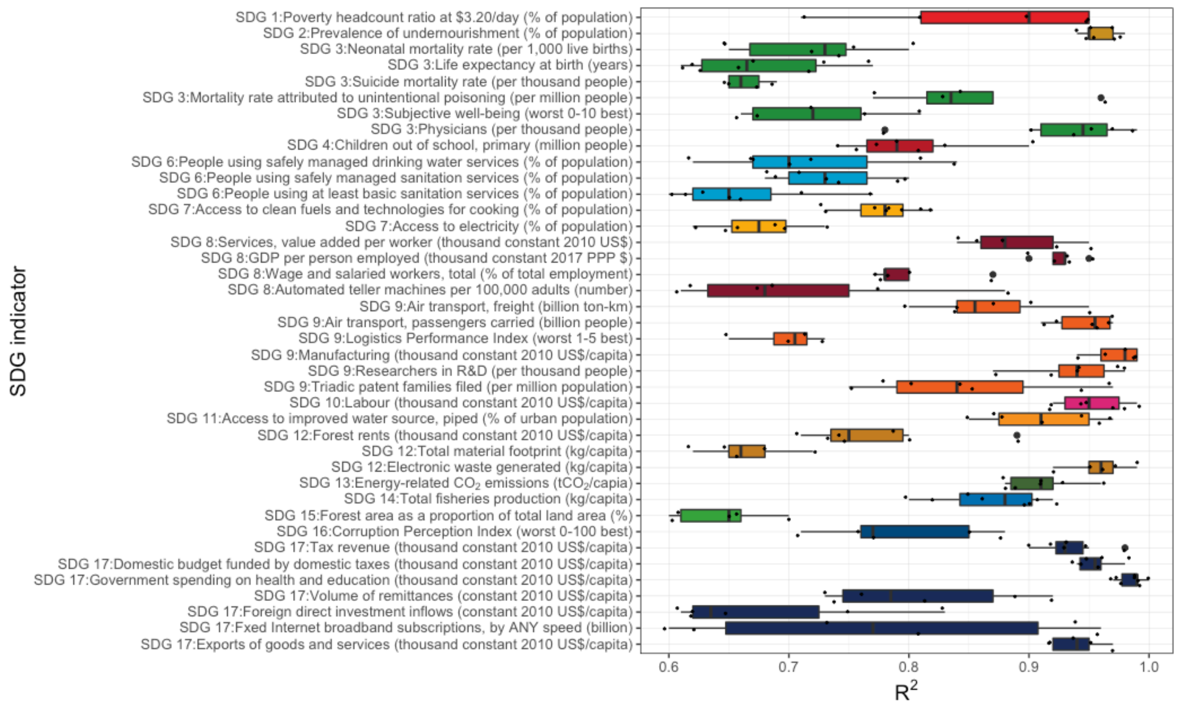

Figure C-1. R^2 of SDG indicators on test sets at the global level.
In each box plot, the central rectangle box spans the first quartile to the third quartile; the central line segment inside the rectangle represents the median value. Note that indicators “SDG 8: GDP growth (%)” and “SDG 8:GDP per capita (thousand constant 2010 US$)” are not included in this plot as they are model inputs based on the IMF prediction [98, 99].

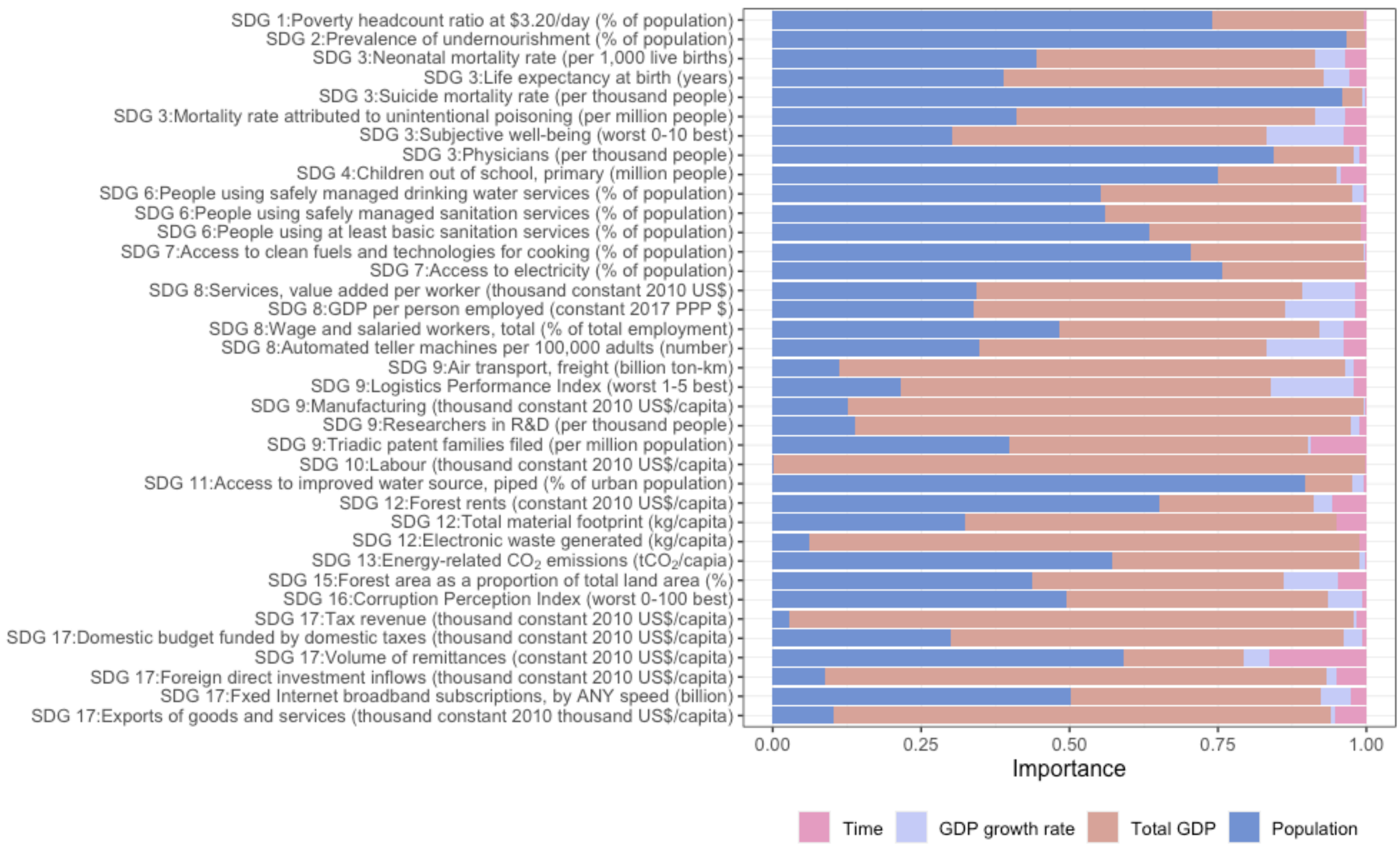

Figure C-2. Relative importance of the predictors used in the prediction models for each SDG indicator.

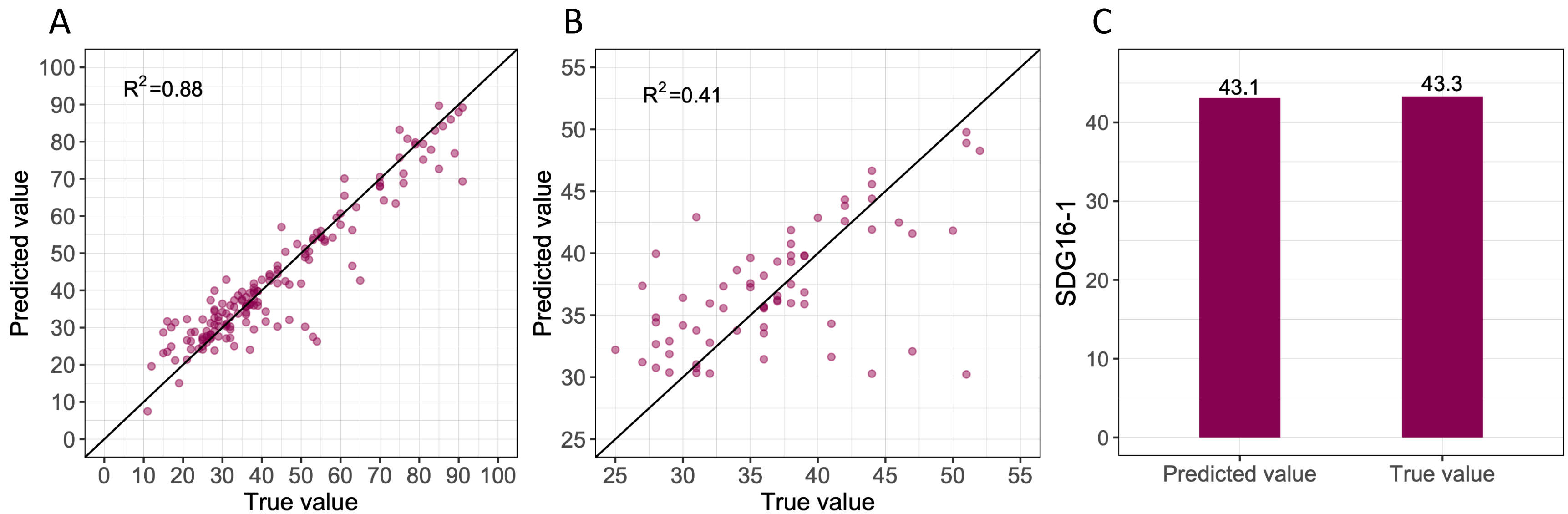

Figure C-3. Sample of prediction comparison between global and country level.
(A) R^2 of SDG16-1 “Corruption Perception Index (worst 0-100 best)” on a test set. (B) R^2 of indicator “Corruption Perception Index (worst 0-100 best)” on the same test set with a specific range (25 < True value < 55). (C) Comparison of the prediction and true value at the global level.

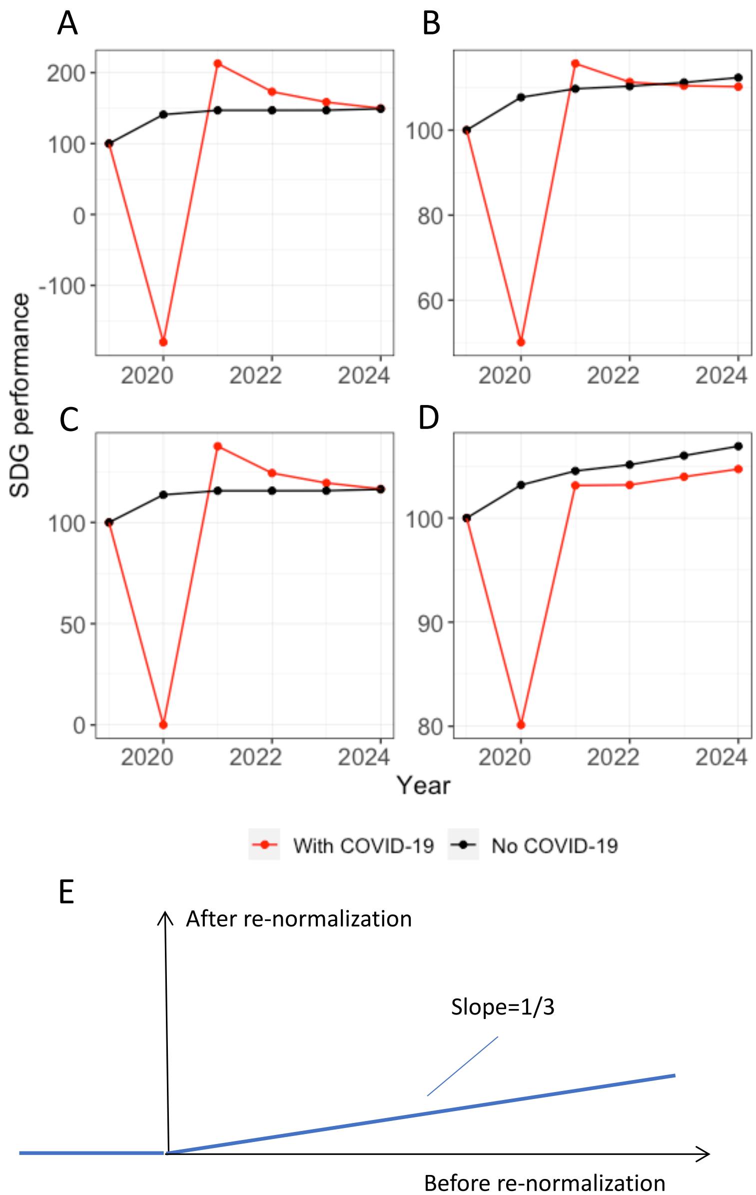

Figure C-4. Re-normalization of the indicator “GDP growth (%)”.
(A) Normalized performance of the indicator “GDP growth (%)” under the COVID-19 and no-COVID-19 scenarios. (B) Normalized performance of SDG 8 under the COVID-19 and no-COVID-19 scenarios. (C) Re-normalized performance of the indicator “GDP growth (%)” under the COVID-19 and no-COVID-19 scenarios. (D) Re-normalized performance of SDG 8 under the COVID-19 and no-COVID-19 scenarios. (E) Piecewise function for re-normalizing the SDG indicator performance.

Figure C-5. Comparison of SDG indicator performances in 2020 between the COVID-19 and no-COVID-19 scenarios at the global level. Note that SDG performance in 2019 is 100.

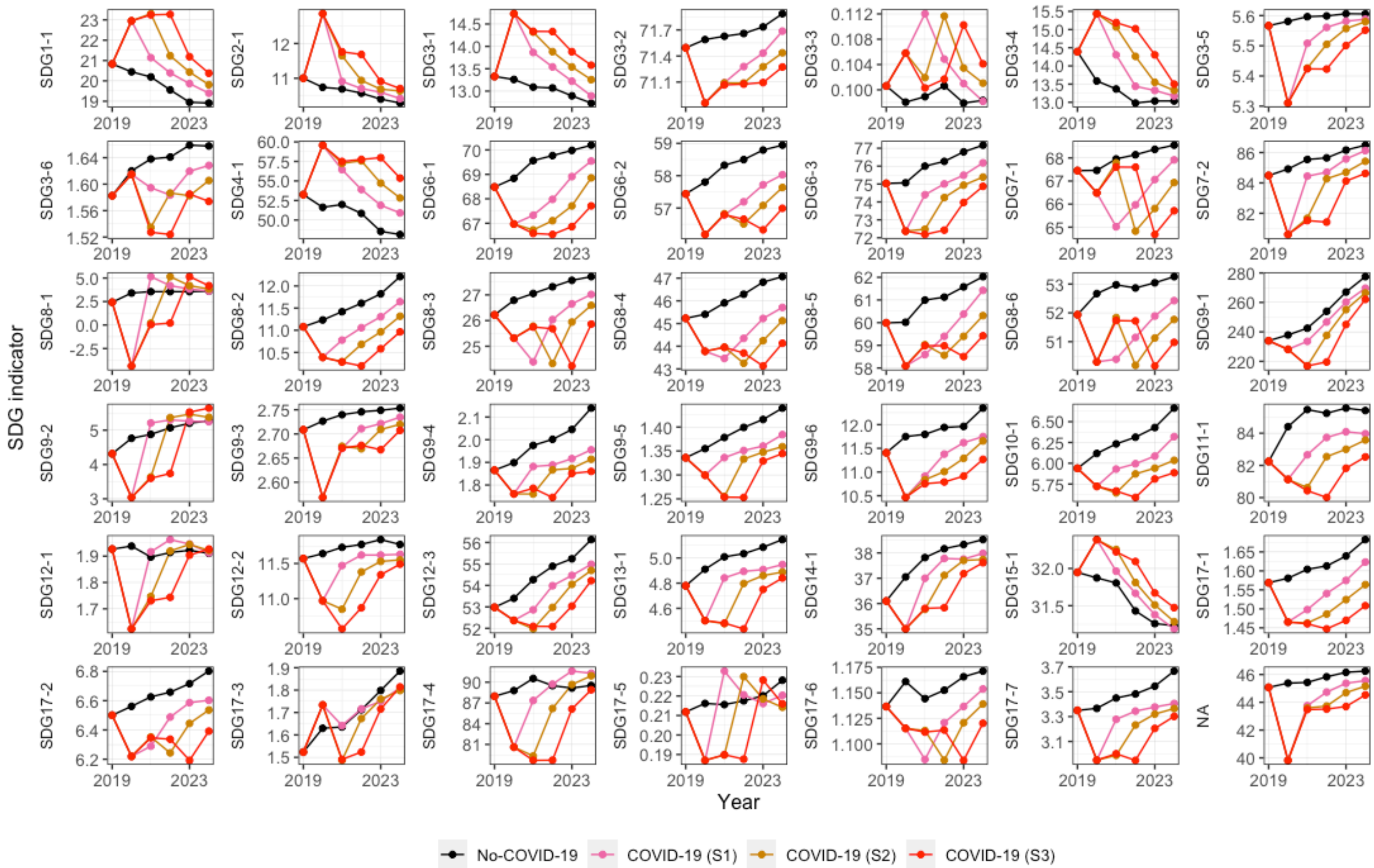

Figure C-6. Comparison of SDG indicators between the three COVID-19 and no-COVID-19 scenarios at the global level from 2019 to 2024.

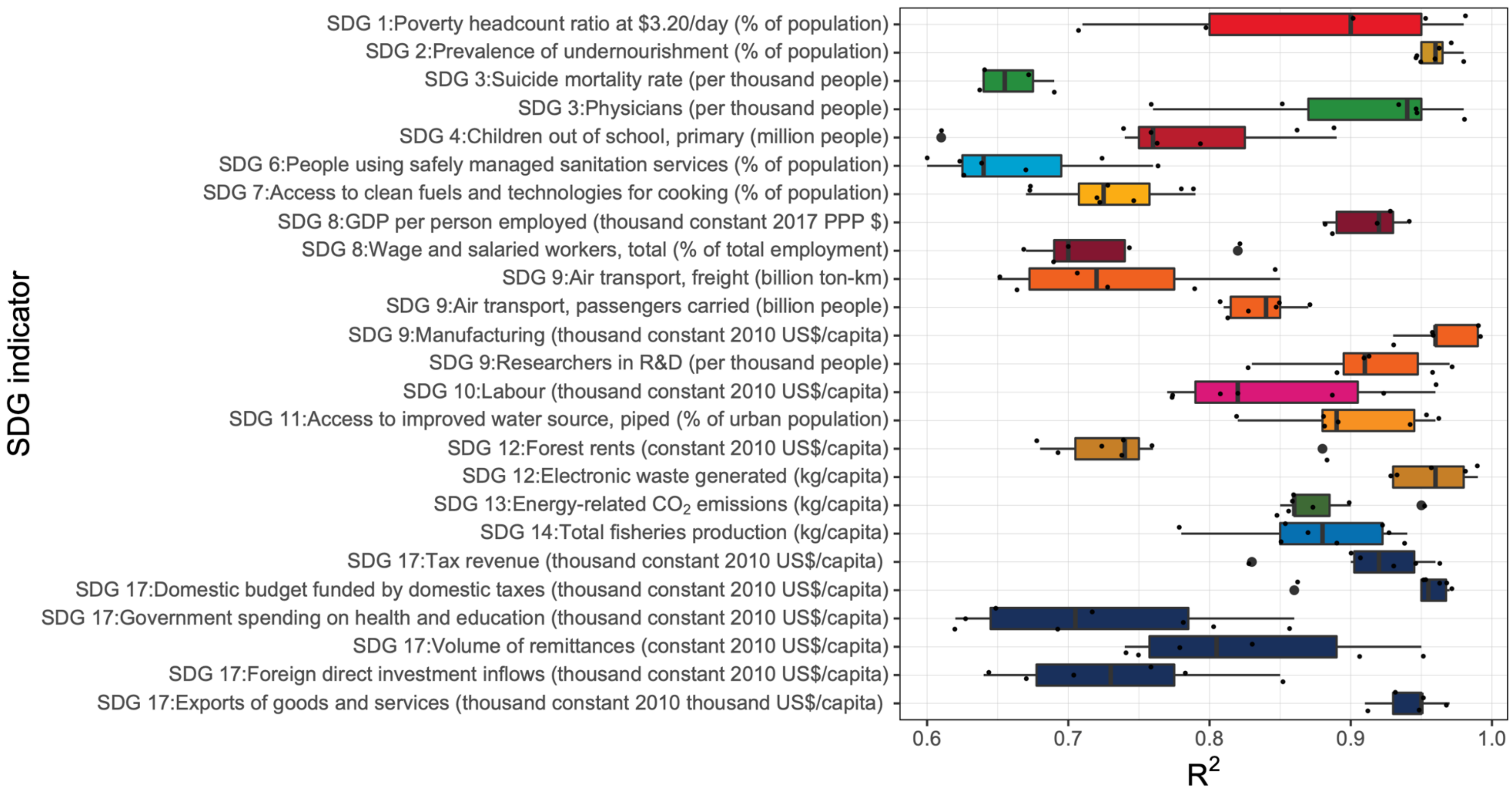

Figure C-7. R^2 of SDG indicators on test sets for the EMDE countries.

In each box plot, the central rectangle box spans the first quartile to the third quartile; the central line segment inside the rectangle represents the median value. Note that indicators "SDG 8: GDP growth (%)" and "SDG 8:GDP per capita (thousand constant 2010 US$)" are not included in this plot as they are model inputs based on the IMF prediction[98, 99].

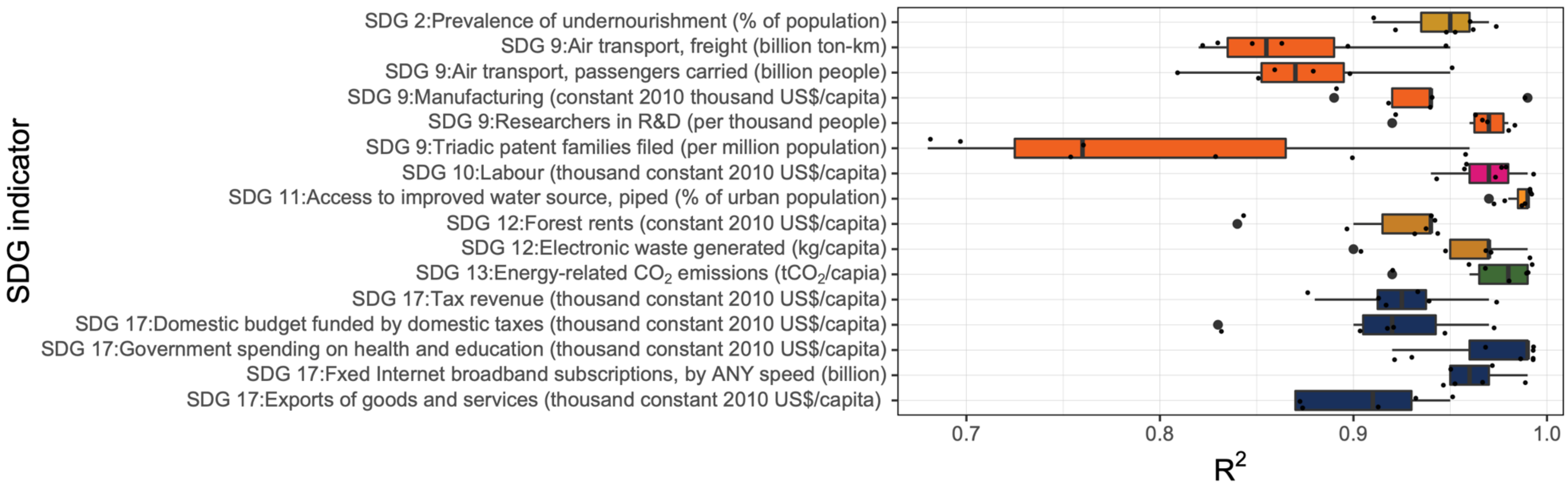

Figure C-8. R^2 of SDG indicators on test sets for AE countries.

In each box plot, the central rectangle box spans the first quartile to the third quartile; the central line segment inside the rectangle represents the median value. Note that indicators “SDG 8: GDP growth (%)” and “SDG 8:GDP per capita (thousand constant 2010 US$)” are not included in this plot as they are model inputs based on the IMF prediction [98, 99].

Table C-1. List of SDG indicators included in this study, their data sources, and the best model prediction model. Note that XGBoost is short for extreme gradient boosting, RF means random forest and SVR means support vector regression.

SDG indicator (unit)	Target	Index	Data source	Best model
SDG 1:Poverty headcount ratio at $3.20/day (% of population)	Target 1.2	SDG1-1	SDR2020 (Sachs et al., 2020a)	XGBoost
SDG 2:Prevalence of undernourishment (% of population)	Target 2.1	SDG2-1	UN (United Nations, 2020a)	XGBoost
SDG 3:Neonatal mortality rate (per 1,000 live births)	Target 3.2	SDG3-1	UN (United Nations, 2020a)	RF
SDG 3:Life expectancy at birth (years)	Target 3.4	SDG3-2	SDR2020 (Sachs et al., 2020a)	RF
SDG 3:Suicide mortality rate (per thousand people)	Target 3.4	SDG3-3	UN (United Nations, 2020a)	RF
SDG 3:Mortality rate attributed to unintentional poisoning (per million people)	Target 3.9	SDG3-4	UN (United Nations, 2020a)	XGBoost
SDG 3:Subjective well-being (worst 0-10 best)	Target 3.b	SDG3-5	SDR2020 (Sachs et al., 2020a)	RF
SDG 3:Physicians (per thousand people)	Target 3.c	SDG3-6	WD (World Bank, 2020)	XGBoost
SDG 4:Children out of school, primary (million people)	Target 4.1	SDG4-1	WD (World Bank, 2020)	XGBoost
SDG 6:People using safely managed drinking water services (% of population)	Target 6.1	SDG6-1	UN (United Nations, 2020a)	XGBoost
SDG 6:People using safely managed sanitation services (% of population)	Target 6.2	SDG6-2	UN (United Nations, 2020a)	RF
SDG 6:People using at least basic sanitation services (% of population)	Target 6.2	SDG6-3	WD (World Bank, 2020)	XGBoost
SDG 7:Access to clean fuels and technologies for cooking (% of population)	Target 7.1	SDG7-1	WD (World Bank, 2020)	RF
SDG 7:Access to electricity (% of population)	Target 7.1	SDG7-2	UN (United Nations, 2020a)	XGBoost
SDG 8:GDP growth rate (%)	Target 8.1	SDG8-1	UN (United Nations, 2020a)	IMF
SDG 8:GDP per capita (thousand constant 2010 US$)	Target 8.2	SDG8-2	WD (World Bank, 2020)	IMF
SDG 8:Services, value added per worker (thousand constant 2010 US$)	Target 8.2	SDG8-3	WD (World Bank, 2020)	RF
SDG 8:GDP per person employed (thousand constant 2017 PPP $)	Target 8.2	SDG8-4	WD (World Bank, 2020)	RF
SDG 8:Wage and salaried workers, total (% of total employment)	Target 8.5	SDG8-5	WD (World Bank, 2020)	RF
SDG 8:Automated teller machines per 100,000 adults (number)	Target 8.10	SDG8-6	UN (United Nations, 2020a)	RF
SDG 9:Air transport, freight (billion ton-km)	Target 9.1	SDG9-1	WD (World Bank, 2020)	RF
SDG 9:Air transport, passengers carried (billion people)	Target 9.1	SDG9-2	WD (World Bank, 2020)	XGBoost
SDG 9:Logistics Performance Index (worst 1-5 best)	Target 9.1	SDG9-3	SDR2020 (Sachs et al., 2020a)	RF
SDG 9:Manufacturing (thousand constant 2010 US$/capita)	Target 9.2	SDG9-4	UN (United Nations, 2020a)	EGB

SDG 9:Researchers in R&D (per thousand people)	Target 9.5	SDG9-5	UN (United Nations, 2020a)	RF
SDG 9:Triadic patent families filed (per million population)	Target 9.a	SDG9-6	SDR2020 (Sachs et al., 2020a)	XGBoost
SDG 10:Labour (thousand constant 2010 US$/capita)	Target 10.4	SDG10-2	WD (World Bank, 2020)	XGBoost
SDG 11:Access to improved water source, piped (% of urban population)	Target 11.2	SDG11-1	SDR2020 (Sachs et al., 2020a)	XGBoost
SDG 12:Forest rents (constant 2010 US$/capita)	Target 12.2	SDG12-1	WD (World Bank, 2020)	XGBoost
SDG 12:Total material footprint (kg/capita)	Target 12.2	SDG12-2	UN (United Nations, 2020a)	RF
SDG 12:Electronic waste generated (million tons)	Target 12.4	SDG12-3	UN (United Nations, 2020a)	XGBoost
SDG 13:Energy-related CO_2 emissions (kg/capita)	Target 13.2	SDG13-1	SDR2020 (Sachs et al., 2020a)	RF
SDG 14:Total fisheries production (kg/capita)	Target 14.4	SDG14-1	WD (World Bank, 2020)	SVR
SDG 15:Forest area as a proportion of total land area (%)	Target 15.1	SDG15-1	UN (United Nations, 2020a)	RF
SDG 16:Corruption Perception Index (worst 0-100 best)	Target 16.5	SDG16-1	SDR2020 (Sachs et al., 2020a)	XGBoost
SDG 17:Tax revenue (thousand constant 2010 US$/capita)	Target 17.1	SDG17-1	WD (World Bank, 2020)	RF
SDG 17:Domestic budget funded by domestic taxes (thousand constant 2010 US$/capita)	Target 17.1	SDG17-2	UN (United Nations, 2020a)	RF
SDG 17:Government spending on health and education (thousand constant 2010 US$/capita)	Target 17.1	SDG17-3	SDGR2020 (Sachs et al., 2020a)	SVM
SDG 17:Volume of remittances (constant 2010 US$/capita)	Target 17.3	SDG17-4	UN (United Nations, 2020a)	XGBoost
SDG 17:Foreign direct investment inflows (thousand constant 2010 US$/capita)	Target 17.3	SDG17-5	WD (World Bank, 2020)	RF
SDG 17:Fxed Internet broadband subscriptions, by ANY speed (billion)	Target 17.6	SDG17-6	UN (United Nations, 2020a)	RF
SDG 17:Exports of goods and services (thousand constant 2010 US$/capita)	Target 17.11	SDG17-7	UN (United Nations, 2020a)	XGBoost

Table C-2. List of 187 countries and regions.
Note that AE means advanced economy and EMDE means emerging market and developing economy.

Country and region	Group	Country and region	Group	Country and region	Group
Australia	AE	Cabo Verde	EMDE	Montenegro	EMDE
Austria	AE	Cambodia	EMDE	Morocco	EMDE
Belgium	AE	Cameroon	EMDE	Mozambique	EMDE
Canada	AE	Central African Republic	EMDE	Myanmar	EMDE
China, Hong Kong Special Administrative Region	AE	Chad	EMDE	Namibia	EMDE
China, Macao Special Administrative Region	AE	Chile	EMDE	Nauru	EMDE
Cyprus	AE	China	EMDE	Nepal	EMDE
Czechia	AE	Colombia	EMDE	Nicaragua	EMDE
Denmark	AE	Comoros	EMDE	Niger	EMDE
Estonia	AE	Congo	EMDE	Nigeria	EMDE
Finland	AE	Costa Rica	EMDE	North Macedonia	EMDE
France	AE	Côte d'Ivoire	EMDE	Oman	EMDE
Germany	AE	Croatia	EMDE	Pakistan	EMDE
Greece	AE	Democratic Republic of the Congo	EMDE	Palau	EMDE
Iceland	AE	Dominica	EMDE	Panama	EMDE
Ireland	AE	Dominican Republic	EMDE	Papua New Guinea	EMDE
Israel	AE	Ecuador	EMDE	Paraguay	EMDE
Italy	AE	Egypt	EMDE	Peru	EMDE
Japan	AE	El Salvador	EMDE	Philippines	EMDE
Latvia	AE	Equatorial Guinea	EMDE	Poland	EMDE
Lithuania	AE	Eritrea	EMDE	Qatar	EMDE
Luxembourg	AE	Eswatini	EMDE	Republic of Moldova	EMDE
Malta	AE	Ethiopia	EMDE	Romania	EMDE
Netherlands	AE	Fiji	EMDE	Russian Federation	EMDE
New Zealand	AE	Gabon	EMDE	Rwanda	EMDE

Norway	AE	Gambia	EMDE	Saint Kitts and Nevis	EMDE
Portugal	AE	Georgia	EMDE	Saint Lucia	EMDE
Puerto Rico	AE	Ghana	EMDE	Saint Vincent and the Grenadines	EMDE
Republic of Korea	AE	Grenada	EMDE	Samoa	EMDE
San Marino	AE	Guatemala	EMDE	Sao Tome and Principe	EMDE
Singapore	AE	Guinea	EMDE	Saudi Arabia	EMDE
Slovakia	AE	Guinea-Bissau	EMDE	Senegal	EMDE
Slovenia	AE	Guyana	EMDE	Serbia	EMDE
Spain	AE	Haiti	EMDE	Seychelles	EMDE
Sweden	AE	Honduras	EMDE	Sierra Leone	EMDE
Switzerland	AE	Hungary	EMDE	Solomon Islands	EMDE
United Kingdom of Great Britain and Northern Ireland	AE	India	EMDE	South Africa	EMDE
United States of America	AE	Indonesia	EMDE	Sri Lanka	EMDE
Afghanistan	EMDE	Iraq	EMDE	Sudan	EMDE
Albania	EMDE	Islamic Republic of Iran	EMDE	Suriname	EMDE
Algeria	EMDE	Jamaica	EMDE	Tajikistan	EMDE
Angola	EMDE	Jordan	EMDE	Tanzania	EMDE
Antigua and Barbuda	EMDE	Kazakhstan	EMDE	Thailand	EMDE
Argentina	EMDE	Kenya	EMDE	Timor-Leste	EMDE
Armenia	EMDE	Kiribati	EMDE	Togo	EMDE
Aruba	EMDE	Kuwait	EMDE	Tonga	EMDE
Azerbaijan	EMDE	Kyrgyzstan	EMDE	Trinidad and Tobago	EMDE
Bahamas	EMDE	Lao People's Democratic Republic	EMDE	Tunisia	EMDE
Bahrain	EMDE	Lesotho	EMDE	Turkey	EMDE
Bangladesh	EMDE	Liberia	EMDE	Turkmenistan	EMDE
Barbados	EMDE	Libya	EMDE	Tuvalu	EMDE
Belarus	EMDE	Madagascar	EMDE	Uganda	EMDE

Belize	EMDE	Malawi	EMDE	Ukraine	EMDE
Benin	EMDE	Malaysia	EMDE	United Arab Emirates	EMDE
Bhutan	EMDE	Maldives	EMDE	Uruguay	EMDE
Bolivia (Plurinational State of)	EMDE	Mali	EMDE	Uzbekistan	EMDE
Bosnia and Herzegovina	EMDE	Marshall Islands	EMDE	Vanuatu	EMDE
Botswana	EMDE	Mauritania	EMDE	Viet Nam	EMDE
Brazil	EMDE	Mauritius	EMDE	Yemen	EMDE
Brunei Darussalam	EMDE	Mexico	EMDE	Zambia	EMDE
Bulgaria	EMDE	Micronesia	EMDE	Zimbabwe	EMDE
Burkina Faso	EMDE	Moldova	EMDE	Zimbabwe	EMDE
Burundi	EMDE	Mongolia	EMDE		

Table C-3. Formula of calculating performance for SDG indicator.
Where x is the value of a given SDG indicator in a given year, x2019 stands for value of the indicator in 2019, and "direction" means the directional relationship between the indicator and its performance (i.e., "negative" means higher indicator value yields lower indicator performance).

SDG indicator (unit)	Index	Direction	Formula for performance
SDG 1:Poverty headcount ratio at $3.20/day (% of population)	SDG1-1	Negative	$x_{2019}/x*100$
SDG 2:Prevalence of undernourishment (% of population)	SDG2-1	Negative	$x_{2019}/x*100$
SDG 3:Neonatal mortality rate (per 1,000 live births)	SDG3-1	Negative	$x_{2019}/x*100$
SDG 3:Life expectancy at birth (years)	SDG3-2	Positive	$x/x_{2019}*100$
SDG 3:Suicide mortality rate (per thousand people)	SDG3-3	Negative	$x_{2019}/x*100$
SDG 3:Mortality rate attributed to unintentional poisoning (per million people)	SDG3-4	Negative	$x_{2019}/x*100$
SDG 3:Subjective well-being (worst 0-10 best)	SDG3-5	Positive	$x/x_{2019}*100$
SDG 3:Physicians (per thousand people)	SDG3-6	Positive	$x/x_{2019}*100$
SDG 4:Children out of school, primary (million people)	SDG4-1	Negative	$x_{2019}/x*100$
SDG 6:People using safely managed drinking water services (% of population)	SDG6-1	Positive	$x/x_{2019}*100$
SDG 6:People using safely managed sanitation services (% of population)	SDG6-2	Positive	$x/x_{2019}*100$
SDG 6:People using at least basic sanitation services (% of population)	SDG6-3	Positive	$x/x_{2019}*100$
SDG 7:Access to clean fuels and technologies for cooking (% of population)	SDG7-1	Positive	$x/x_{2019}*100$
SDG 7:Access to electricity (% of population)	SDG7-2	Positive	$x/x_{2019}*100$
SDG 8:GDP growth rate (%)	SDG8-1	Piecewise	$(x/x_{2019}*100-100)/3+100$
SDG 8:GDP per capita (thousand constant 2010 US$)	SDG8-2	Positive	$x/x_{2019}*100$
SDG 8:Services, value added per worker (thousand constant 2010 US$)	SDG8-3	Positive	$x/x_{2019}*100$
SDG 8:GDP per person employed (thousand constant 2017 PPP $)	SDG8-4	Positive	$x/x_{2019}*100$
SDG 8:Wage and salaried workers, total (% of total employment)	SDG8-5	Positive	$x/x_{2019}*100$
SDG 8:Automated teller machines per 100,000 adults (number)	SDG8-6	Positive	$x/x_{2019}*100$
SDG 9:Air transport, freight (billion ton-km)	SDG9-1	Positive	$x/x_{2019}*100$
SDG 9:Air transport, passengers carried (billion people)	SDG9-2	Positive	$x/x_{2019}*100$
SDG 9:Logistics Performance Index (worst 1-5 best)	SDG9-3	Positive	$x/x_{2019}*100$
SDG 9:Manufacturing (thousand constant 2010 US$/capita)	SDG9-4	Positive	$x/x_{2019}*100$
SDG 9:Researchers in R&D (per thousand people)	SDG9-5	Positive	$x/x_{2019}*100$

SDG 9:Triadic patent families filed (per million population)	SDG9-6	Positive	$x/x_{2019}*100$
SDG 10:Labour (thousand constant 2010 US$/capita)	SDG10-2	Positive	$x/x_{2019}*100$
SDG 11:Access to improved water source, piped (% of urban population)	SDG11-1	Positive	$x/x_{2019}*100$
SDG 12:Forest rents (constant 2010 US$/capita)	SDG12-1	Negative	$x_{2019}/x*100$
SDG 12:Total material footprint (kg/capita)	SDG12-2	Negative	$x_{2019}/x*100$
SDG 12:Electronic waste generated (million tons)	SDG12-3	Negative	$x_{2019}/x*100$
SDG 13:Energy-related CO_2 emissions (kg/capita)	SDG13-1	Negative	$x_{2019}/x*100$
SDG 14:Total fisheries production (kg/capita)	SDG14-1	Negative	$x_{2019}/x*100$
SDG 15:Forest area as a proportion of total land area (%)	SDG15-1	Positive	$x/x_{2019}*100$
SDG 16:Corruption Perception Index (worst 0-100 best)	SDG16-1	Positive	$x/x_{2019}*100$
SDG 17:Tax revenue (thousand constant 2010 US$/capita)	SDG17-1	Positive	$x/x_{2019}*100$
SDG 17:Domestic budget funded by domestic taxes (thousand constant 2010 US$/capita)	SDG17-2	Positive	$x/x_{2019}*100$
SDG 17:Government spending on health and education (thousand constant 2010 US$/capita)	SDG17-3	Positive	$x/x_{2019}*100$
SDG 17:Volume of remittances (constant 2010 US$/capita)	SDG17-4	Positive	$x/x_{2019}*100$
SDG 17:Foreign direct investment inflows (thousand constant 2010 US$/capita)	SDG17-5	Positive	$x/x_{2019}*100$
SDG 17:Fixed Internet broadband subscriptions, by ANY speed (billion)	SDG17-6	Positive	$x/x_{2019}*100$
SDG 17:Exports of goods and services (thousand constant 2010 US$/capita)	SDG17-7	Positive	$x/x_{2019}*100$

References

1. Parpart, J. L.; Veltmeyer, H., The development project in theory and practice: a review of its shifting dynamics. *Canadian Journal of Development Studies/Revue canadienne d'études du développement* **2004,** *25*, (1), 39-59.

2. Tipps, D. C., Modernization theory and the comparative study of national societies: A critical perspective. *Comparative studies in society and history* **1973,** *15*, (2), 199-226.

3. Cardoso, F. H., Dependency and development in Latin America. In *Introduction to the Sociology of "Developing Societies"*, Springer: 1982; pp 112-127.

4. Ahiakpor, J. C., The success and failure of dependency theory: The experience of Ghana. *International Organization* **1985,** *39*, (3), 535-552.

5. Sen, A., Development as freedom (1999). *The globalization and development reader: Perspectives on development and global change* **2014,** *525*.

6. Anand, S.; Sen, A., Human Development Index: Methodology and Measurement. **1994**.

7. Keeble, B. R., The Brundtland report:'Our common future'. *Medicine and war* **1988,** *4*, (1), 17-25.

8. United Nations, Millennium Development Goals, https://www.un.org/millenniumgoals. **2000**.

9. United Nations, Transforming our world: The 2030 agenda for sustainable development. *General Assembley 70 session* **2015**.

10. World Bank, DataBank: Sustainable Development Goals. **2018**.

11. The State of Development Data, https://opendatawatch.com/the-state-of-development-data-2016/. **2016**.

12. United Nation System, TOTAL EXPENDITURE. *http://www.unsystem.org/content/FS-F00-05* **2017**.

13. Fodor, I. K. *A survey of dimension reduction techniques*; Lawrence Livermore National Lab., CA (US): 2002.

14. Wold, S.; Esbensen, K.; Geladi, P., Principal component analysis. *Chemometrics and intelligent laboratory systems* **1987,** *2*, (1-3), 37-52.

15. Lamichhane, S.; Eğilmez, G.; Gedik, R.; Bhutta, M. K. S.; Erenay, B., Benchmarking OECD countries' sustainable development performance: A goal-specific principal component analysis approach. *Journal of Cleaner Production* **2021,** *287*, 125040.

16. Jiang, Q.; Liu, Z.; Liu, W.; Li, T.; Cong, W.; Zhang, H.; Shi, J., A principal component analysis based three-dimensional sustainability assessment model to evaluate corporate sustainable performance. *Journal of Cleaner Production* **2018,** *187*, 625-637.

17. Rummel, R. J., *Applied factor analysis*. Northwestern University Press: 1988.

18. Laurett, R.; Paco, A.; Mainardes, E. W., Measuring sustainable development, its antecedents, barriers and consequences in agriculture: An exploratory factor analysis. *Environmental Development* **2021,** *37*, 100583.

19. Joliffe, I.; Morgan, B., Principal component analysis and exploratory factor analysis. *Statistical methods in medical research* **1992,** *1*, (1), 69-95.

20. Bettencourt, L. M.; Lobo, J.; Strumsky, D.; West, G. B., Urban scaling and its deviations: Revealing the structure of wealth, innovation and crime across cities. *PloS one* **2010,** *5*, (11).

21. Keuschnigg, M., Scaling trajectories of cities. *Proceedings of the National Academy of Sciences* **2019,** *116*, (28), 13759-13761.

22. Keuschnigg, M.; Mutgan, S.; Hedström, P., Urban scaling and the regional divide. *Science advances* **2019,** *5*, (1), eaav0042.

23. Bettencourt, L. M.; Lobo, J.; Helbing, D.; Kühnert, C.; West, G. B., Growth, innovation, scaling, and the pace of life in cities. *Proceedings of the national academy of sciences* **2007,** *104*, (17), 7301-7306.

24. Bettencourt, L. M.; Lobo, J., Urban scaling in Europe. *Journal of The Royal Society Interface* **2016,** *13*, (116), 20160005.

25. Bettencourt, L. M.; Lobo, J.; Strumsky, D.; West, G. B., Urban scaling and its deviations: Revealing the structure of wealth, innovation and crime across cities. *PloS one* **2010,** *5*, (11), e13541.

26. Bettencourt, L. M., Urban growth and the emergent statistics of cities. *Science advances* **2020,** *6*, (34), eaat8812.

27. Sahasranaman, A.; Bettencourt, L. M., Urban geography and scaling of contemporary Indian cities. *Journal of the Royal Society Interface* **2019,** *16*, (152), 20180758.

28. United Nations, The Sustainable Development Goals Report 2020, https://www.un.org/sustainabledevelopment/progress-report. **2020**.

29. United Nations, Socioeconomic Impact of COVID-19: Policy Briefs, https://www.unescwa.org/publications/socioeconomic-impact-covid-19-policy-briefs. **2020**.

30. Sachs, J.; Schmidt-Traub, G.; Kroll, C. L., Guillaume Fuller, Grayson; Woelm, F., Sustainable Development Report 2020, https://sdgindex.org. **2020**.

31. Naidoo, R.; Fisher, B., Reset sustainable development goals for a pandemic world. *Nature* **2020,** *583*, 198-201.

32. Sachs, J.; Schmidt-Traub, G.; Lafortune, G., Speaking truth to power about the SDGs. *Nature* **2020,** *584*, (7821), 344-344.

33. Laborde, D.; Martin, W.; Swinnen, J.; Vos, R., COVID-19 risks to global food security. *Science* **2020,** *369*, (6503), 500-502.

34. Barbier, E. B.; Burgess, J. C., Sustainability and development after COVID-19. *World Development* **2020,** *135*, 105082.

35. Khetrapal, S.; Bhatia, R., Impact of COVID-19 pandemic on health system & Sustainable Development Goal 3. *The Indian Journal of Medical Research* **2020,** *151*, (5), 395.

36. Nundy, S.; Ghosh, A.; Mesloub, A.; Albaqawy, G. A.; Alnaim, M. M., Impact of COVID-19 pandemic on socio-economic, energy-environment and transport sector globally and sustainable development goal (SDG). *Journal of Cleaner Production* **2021**, 127705.

37. Storm, H.; Baylis, K.; Heckelei, T., Machine learning in agricultural and applied economics. *European Review of Agricultural Economics* **2020,** *47*, (3), 849-892.

38. Podgorski, J.; Berg, M., Global threat of arsenic in groundwater. *Science* **2020,** *368*, (6493), 845-850.

39. Porciello, J.; Ivanina, M.; Islam, M.; Einarson, S.; Hirsh, H., Accelerating evidence-informed decision-making for the Sustainable Development Goals using machine learning. *Nature Machine Intelligence* **2020,** *2*, (10), 559-565.

40. Alova, G.; Trotter, P. A.; Money, A., A machine-learning approach to predicting Africa's electricity mix based on planned power plants and their chances of success. *Nature Energy* **2021,** *6*, (2), 158-166.

41. Jean, N.; Burke, M.; Xie, M.; Davis, W. M.; Lobell, D. B.; Ermon, S., Combining satellite imagery and machine learning to predict poverty. *Science* **2016,** *353*, (6301), 790-794.

42. Zhao, B.; Yu, L.; Wang, C.; Shuai, C.; Zhu, J.; Qu, S.; Taiebat, M.; Xu, M., Urban Air Pollution Mapping Using Fleet Vehicles as Mobile Monitors and Machine Learning. *Environmental Science & Technology* **2021,** *55*, (8), 5579-5588.

43. Shuai, C.; Yu, L.; Chen, X.; Zhao, B.; Qu, S.; Zhu, J.; Liu, J.; Miller, S. A.; Xu, M., Principal indicators to monitor Sustainable Development Goals. *Environmental Research Letters* **2021**.

44. Fritz, S.; See, L.; Carlson, T.; Haklay, M. M.; Oliver, J. L.; Fraisl, D.; Mondardini, R.; Brocklehurst, M.; Shanley, L. A.; Schade, S., Citizen science and the United Nations Sustainable Development Goals. *Nature Sustainability* **2019,** *2*, (10), 922-930.

45. Hsu, A.; Malik, O.; Johnson, L.; Esty, D. C., Development: Mobilize citizens to track sustainability. *Nature News* **2014,** *508*, (7494), 33.

46. Liu, J.; Mooney, H.; Hull, V.; Davis, S. J.; Gaskell, J.; Hertel, T.; Lubchenco, J.; Seto, K. C.; Gleick, P.; Kremen, C., Systems integration for global sustainability. *Science* **2015,** *347*, (6225), 1258832.

47. Global Partership for Sustainable Development Data, http://www.data4sdgs.org/partner-listing.

48. Lyytimäki, J., Seeking SDG indicators. *Nature Sustainability* **2019,** *2*, (8), 646-646.

49. Xu, Z.; Chau, S. N.; Chen, X.; Zhang, J.; Li, Y.; Dietz, T.; Wang, J.; Winkler, J. A.; Fan, F.; Huang, B., Assessing progress towards sustainable development over space and time. *Nature* **2020,** *577*, (7788), 74-78.

50. Nilsson, M.; Griggs, D.; Visbeck, M., Policy: map the interactions between Sustainable Development Goals. *Nature News* **2016,** *534*, (7607), 320.

51. Lusseau, D.; Mancini, F., Income-based variation in Sustainable Development Goal interaction networks. *Nature Sustainability* **2019,** *2*, (3), 242.

52. Liu, J.; Hull, V.; Godfray, H. C. J.; Tilman, D.; Gleick, P.; Hoff, H.; Pahl-Wostl, C.; Xu, Z.; Chung, M. G.; Sun, J., Nexus approaches to global sustainable development. *Nature Sustainability* **2018,** *1*, (9), 466.

53. Nerini, F. F.; Sovacool, B.; Hughes, N.; Cozzi, L.; Cosgrave, E.; Howells, M.; Tavoni, M.; Tomei, J.; Zerriffi, H.; Milligan, B., Connecting climate action with other Sustainable Development Goals. *Nature Sustainability* **2019**, 1.

54. Haines, A.; Amann, M.; Borgford-Parnell, N.; Leonard, S.; Kuylenstierna, J.; Shindell, D., Short-lived climate pollutant mitigation and the Sustainable Development Goals. *Nature climate change* **2017,** *7*, (12), 863.

55. Cadima, J. F.; Jolliffe, I. T., Variable selection and the interpretation of principal subspaces. *Journal of Agricultural, Biological, and Environmental Statistics* **2001,** *6*, (1), 62.

56. Steinmann, Z. J.; Schipper, A. M.; Hauck, M.; Huijbregts, M. A., How many environmental impact indicators are needed in the evaluation of product life cycles? *Environmental science & technology* **2016,** *50*, (7), 3913-3919.

57. Jolliffe, I., *Principal component analysis*. Springer: 2011.

58. Thomas, K., The lmf R package. **2015**.

59. Cadima, J.; Cerdeira, J. O.; Minhoto, M., Computational aspects of algorithms for variable selection in the context of principal components. *Computational Statistics & Data Analysis* **2004,** *47*, (2), 225-236.

60. Cadima, J.; Cerdeira, J. O.; Silva, P. D.; Minhoto, M., The subselect R package. **2012**.

61. Capps, K. A.; Bentsen, C. N.; Ramírez, A., Poverty, urbanization, and environmental degradation: urban streams in the developing world. *Freshwater Science* **2016,** *35*, (1), 429-435.

62. Carter, R. C.; Danert, K., The private sector and water and sanitation services—policy and poverty issues. *Journal of International Development* **2003,** *15*, (8), 1067-1072.

63. United Nations, A World That Counts: Mobilising The Data Revolution for Sustainable Development, https://www.undatarevolution.org/report/. **2015**.

64. OECD, Global Outlook on Financing for Sustainable Development 2019: Time to Face the Challenge. **2019**.

65. Mealy, P.; Farmer, J. D.; Teytelboym, A., Interpreting economic complexity. *Science advances* **2019,** *5*, (1), eaau1705.

66. Bloom, D. E.; Khoury, A.; Subbaraman, R., The promise and peril of universal health care. *Science* **2018,** *361*, (6404), eaat9644.

67. Smith, T. M.; Goodkind, A. L.; Kim, T.; Pelton, R. E.; Suh, K.; Schmitt, J., Subnational mobility and consumption-based environmental accounting of US corn in animal protein and ethanol supply chains. *Proceedings of the National Academy of Sciences* **2017,** *114*, (38), E7891-E7899.

68. Bettencourt, L. M., The origins of scaling in cities. *science* **2013,** *340*, (6139), 1438-1441.

69. Ortman, S. G.; Cabaniss, A. H.; Sturm, J. O.; Bettencourt, L. M., Settlement scaling and increasing returns in an ancient society. *Science Advances* **2015,** *1*, (1), e1400066.

70. Li, R.; Dong, L.; Zhang, J.; Wang, X.; Wang, W.-X.; Di, Z.; Stanley, H. E., Simple spatial scaling rules behind complex cities. *Nature communications* **2017,** *8*, (1), 1841.

71. Depersin, J.; Barthelemy, M., From global scaling to the dynamics of individual cities. *Proceedings of the National Academy of Sciences* **2018,** *115*, (10), 2317-2322.

72. Bloom, D. E.; Canning, D.; Fink, G., Urbanization and the wealth of nations. *Science* **2008,** *319*, (5864), 772-775.

73. World Bank, Data. https://data.worldbank.org/indicator.

74. Our World in Data, https://dataunodc.un.org.

75. United Nations, United Nations Office on Drugs and Crime. https://dataunodc.un.org.

76. West, G. B., *Scale: the universal laws of growth, innovation, sustainability, and the pace of life in organisms, cities, economies, and companies*. Penguin: 2017.

77. Jiao, L.; Xu, Z.; Xu, G.; Zhao, R.; Liu, J.; Wang, W., Assessment of urban land use efficiency in China: A perspective of scaling law. *Habitat International* **2020,** *99*, 102172.

78. Li, R.; Dong, L.; Zhang, J.; Wang, X.; Wang, W.-X.; Di, Z.; Stanley, H. E., Simple spatial scaling rules behind complex cities. *Nature communications* **2017,** *8*, (1), 1-7.

79. Soo, K. T., Zipf's Law for cities: a cross-country investigation. *Regional science and urban Economics* **2005,** *35*, (3), 239-263.

80. Verbavatz, V.; Barthelemy, M., The growth equation of cities. *Nature* **2020,** *587*, (7834), 397-401.

81. Giesen, K.; Südekum, J., Zipf's law for cities in the regions and the country. *Journal of economic geography* **2011,** *11*, (4), 667-686.

82. Jiang, B.; Yin, J.; Liu, Q., Zipf's law for all the natural cities around the world. *International Journal of Geographical Information Science* **2015,** *29*, (3), 498-522.

83. United Nations, Population Dynamics.https://population.un.org/wup.

84. Oxford Economics, Cities and Regions Services. https://www.oxfordeconomics.com/cities-and-regions.

85. West, G. B.; Brown, J. H.; Enquist, B. J., A general model for the origin of allometric scaling laws in biology. *Science* **1997,** *276*, (5309), 122-126.

86. Daepp, M. I.; Hamilton, M. J.; West, G. B.; Bettencourt, L. M., The mortality of companies. *Journal of The Royal Society Interface* **2015,** *12*, (106), 20150120.

87. Henderson, V., The urbanization process and economic growth: The so-what question. *Journal of Economic growth* **2003,** *8*, (1), 47-71.

88. Our World in Data, Urban population vs. GDP per capita, https://ourworldindata.org/grapher/urbanization-vs-gdp. **2016**.

89. United Nations, Human Development Index.http://hdr.undp.org/en/content/human-development-index-hdi.

90. Hoekstra, A. Y.; Wiedmann, T. O., Humanity's unsustainable environmental footprint. *Science* **2014,** *344*, (6188), 1114-1117.

91. Hong, C.; Burney, J. A.; Pongratz, J.; Nabel, J. E.; Mueller, N. D.; Jackson, R. B.; Davis, S. J., Global and regional drivers of land-use emissions in 1961–2017. *Nature* **2021,** *589*, (7843), 554-561.

92. Smith, M. R.; Myers, S. S., Impact of anthropogenic CO 2 emissions on global human nutrition. *Nature Climate Change* **2018,** *8*, (9), 834-839.

93. Meng, J.; Mi, Z.; Guan, D.; Li, J.; Tao, S.; Li, Y.; Feng, K.; Liu, J.; Liu, Z.; Wang, X., The rise of South–South trade and its effect on global CO 2 emissions. *Nature communications* **2018,** *9*, (1), 1-7.

94. Crippa, M.; Solazzo, E.; Guizzardi, D.; Monforti-Ferrario, F.; Tubiello, F.; Leip, A., Food systems are responsible for a third of global anthropogenic GHG emissions. *Nature Food* **2021**, 1-12.

95. Beerling, D. J.; Kantzas, E. P.; Lomas, M. R.; Wade, P.; Eufrasio, R. M.; Renforth, P.; Sarkar, B.; Andrews, M. G.; James, R. H.; Pearce, C. R., Potential for large-scale CO 2 removal via enhanced rock weathering with croplands. *Nature* **2020,** *583*, (7815), 242-248.

96. Katz, J. S., Scale-independent indicators and research evaluation. *Science and Public Policy* **2000,** *27*, (1), 23-36.

97. World Health Organization, WHO Corronavirus Disease (COVID-19) Dashboard. *https://covid19.who.int* **2020**.

98. International Monetary Fund, World Economic Outlook Database, https://www.imf.org/en/Publications/WEO/weo-database/2020/October. **2020**.

99. International Monetary Fund, World Economic Outlook Database, https://www.imf.org/en/Publications/WEO/weo-database/2019/October. **2019**.

100. Kissler, S. M.; Tedijanto, C.; Goldstein, E.; Grad, Y. H.; Lipsitch, M., Projecting the transmission dynamics of SARS-CoV-2 through the postpandemic period. *Science* **2020,** *368*, (6493), 860-868.

101. Dinda, S., Environmental Kuznets curve hypothesis: a survey. *Ecological economics* **2004,** *49*, (4), 431-455.

102. Liu, J.; Mooney, H.; Hull, V.; Davis, S. J.; Gaskell, J.; Hertel, T.; Lubchenco, J.; Seto, K. C.; Gleick, P.; Kremen, C., Systems integration for global sustainability. *Science* **2015,** *347*, (6225).

103. Liu, J.; Hull, V.; Godfray, H. C. J.; Tilman, D.; Gleick, P.; Hoff, H.; Pahl-Wostl, C.; Xu, Z.; Chung, M. G.; Sun, J., Nexus approaches to global sustainable development. *Nature Sustainability* **2018,** *1*, (9), 466-476.

104. Xu, Z., On the causality between export growth and GDP growth: an empirical reinvestigation. *Review of International Economics* **1996,** *4*, (2), 172-184.

105. United Nations, SDGindicators, https://unstats.un.org/sdgs/indicators/database/. **2020**.

106. World Bank, Sustainable Development Goals, http://datatopics.worldbank.org/sdgs/. **2020**.

107. World Bank, World Bank Open Data, https://data.worldbank.org. **2019**.

108. United Nations, World Population Prospects 2019, https://population.un.org/wpp/Download/Standard/Population/. **2019**.

109. Meyer, D.; Dimitriadou, E.; Hornik, K.; Weingessel, A.; Leisch, F.; Chang, C.-C.; Lin, C.-C.; Meyer, M. D., Package 'e1071'. *The R Journal* **2019**.

110. RColorBrewer, S.; Liaw, M. A., Package 'randomForest'. *University of California, Berkeley: Berkeley, CA, USA* **2018**.

111. Chen, T.; He, T.; Benesty, M.; Khotilovich, V., Package 'xgboost'. In 2020.

112. DiCiccio, T. J.; Efron, B., Bootstrap confidence intervals. *Statistical science* **1996**, 189-212.

113. Xu, Z.; Li, Y.; Chau, S. N.; Dietz, T.; Li, C.; Wan, L.; Zhang, J.; Zhang, L.; Li, Y.; Chung, M. G., Impacts of international trade on global sustainable development. *Nature Sustainability* **2020**, 1-8.

114. Schmidt-Traub, G.; Kroll, C.; Teksoz, K.; Durand-Delacre, D.; Sachs, J. D., National baselines for the Sustainable Development Goals assessed in the SDG Index and Dashboards. *Nature geoscience* **2017,** *10*, (8), 547-555.

115. Gallego, J. A.; Prem, M.; Vargas, J. F., Corruption in the Times of Pandemia. *Available at SSRN 3600572* **2020**.

116. Iacus, S. M.; Natale, F.; Santamaria, C.; Spyratos, S.; Vespe, M., Estimating and projecting air passenger traffic during the COVID-19 coronavirus outbreak and its socio-economic impact. *Safety Science* **2020**, 104791.

117. Liu, Z.; Ciais, P.; Deng, Z.; Lei, R.; Davis, S. J.; Feng, S.; Zheng, B.; Cui, D.; Dou, X.; Zhu, B., Near-real-time monitoring of global CO 2 emissions reveals the effects of the COVID-19 pandemic. *Nature communications* **2020,** *11*, (1), 1-12.

118. Carbon Monitor, https://carbonmonitor.org. **2020**.

119. McCollum, D. L.; Zhou, W.; Bertram, C.; De Boer, H.-S.; Bosetti, V.; Busch, S.; Després, J.; Drouet, L.; Emmerling, J.; Fay, M., Energy investment needs for fulfilling the Paris Agreement and achieving the Sustainable Development Goals. *Nature Energy* **2018,** *3*, (7), 589-599.

120. Andrijevic, M.; Schleussner, C.-F.; Gidden, M. J.; McCollum, D. L.; Rogelj, J., COVID-19 recovery funds dwarf clean energy investment needs. *Science* **2020,** *370*, (6514), 298-300.

121. Uslaner, E. M., The pitfalls of per capita. *American Journal of Political Science* **1976**, 125-133.

122. Lyons, W., Per capita index construction: a defense. *American Journal of Political Science* **1977**, 177-182.