ENVIRONMENTAL RESEARCH
LETTERS

LETTER • **OPEN ACCESS**

Principal indicators to monitor sustainable development goals

IOP Publishing Environ. Res. Lett. 16 (2021) 124015 https://doi.org/10.1088/1748-9326/ac3697

ENVIRONMENTAL RESEARCH
LETTERS

OPEN ACCESS

RECEIVED
2 August 2021

REVISED
3 November 2021

ACCEPTED FOR PUBLICATION
4 November 2021

PUBLISHED
18 November 2021

LETTER

Principal indicators to monitor sustainable development goals

Chenyang Shuai[1,2], Long Yu[3], Xi Chen[4], Bu Zhao[1,2], Shen Qu[1], Ji Zhu[5], Jianguo Liu[6], Shelie A Miller[1,7] and Ming Xu[1,7,*]

[1] School for Environment and Sustainability, University of Michigan, Ann Arbor, MI, United States of America
[2] Michigan Institute for Computational Discovery and Engineering, University of Michigan, Ann Arbor, MI, United States of America
[3] Department of Statistics, Fudan University, Shanghai, People's Republic of China
[4] College of Economics and Management, Southwest University, Chongqing, People's Republic of China
[5] Department of Statistics, University of Michigan, Ann Arbor, MI, United States of America
[6] Center for Systems Integration and Sustainability, Department of Fisheries and Wildlife, Michigan State University, East Lansing, MI, United States of America
[7] Department of Civil and Environmental Engineering, University of Michigan, Ann Arbor, MI, United States of America
[*] Author to whom any correspondence should be addressed.

E-mail: mingxu@umich.edu

Keywords: sustainable development goals, principal indicators, dimension reduction

Supplementary material for this article is available online

Abstract

Hundreds of indicators are available to monitor progress of countries and regions towards the Sustainable Development Goals (SDGs). However, the sheer number of indicators poses unprecedented challenges for data collection and compilation. Here we identify a subset of SDG indicators (principal indicators) that are relatively easy to collect data for and also are representative for all the indicators by considering the complex interrelationship among them. We find 147 principal indicators that 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. Our results can guide future investment in building the data infrastructure for SDG monitoring to give priorities to these principal indicators for global comparison.

1. Introduction

The United Nations (UN) set 17 Sustainable Development Goals (SDGs) in 2015 as a universal call to eliminate poverty, protect the environment, and ensure peace and prosperity for all people (United Nations 2016). To monitor the progress towards achieving the SDGs, the 17 goals are underpinned by 169 targets which are measured by an even larger number of indicators (Espey 2019). These indicators are tracked at the national level by various agencies and organizations.

Collecting data to regularly monitor the SDG indicators is not an easy task (Hsu *et al* 2014, Liu *et al* 2015, Fritz *et al* 2019). Such efforts need significant investment of institutional and financial resources and engagement with a vast number of stakeholders. For example, over 1200 stakeholders worldwide have contributed to data collection for SDG indicators, including governments, NGOs, research institutions, multilateral organizations, and private sectors (Global Partnership for Sustainable Development Data). 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 (Open Data Watch 2016), more than the UN's annual expenditure in 2016 (United Nation System 2017). Despite many achievements, it is still challenging, if not impractical, to annually update the sheer number of SDG indicators for all countries and regions (Lyytimäki 2019, Xu *et al* 2020). This challenge calls for alternative approaches to monitor the SDGs at a lower cost.

In general, there are three strategies to address the data challenge for SDG monitoring: (a) develop and improve the institutional capacity for SDG monitoring across all countries and regions (Fritz *et al* 2019); (b) use novel technologies or systems such as remote sensing (Watmough *et al* 2019, Yeh *et al*

2020), citizen science (Fritz *et al* 2019), and volunteered geographic information (Hsu *et al* 2014); and (c) reduce the amount of data needed for SDG monitoring using data-driven or statistical methods. The first two strategies can generate high-quality data, but require significant investment of institutional and financial resources (Open Data Watch 2016). There are already reports on declining financial investment in sustainable development, particularly in developing countries (OECD 2019). The deep global recession triggered by COVID-19 even further deteriorates future investment (World Bank 2020). Moreover, most of the proposed new technologies and systems as mentioned above are only being tested; little is known about the requirements of scaling up (Open Data Watch 2016). Given these challenges, the third strategy—reducing data demand—becomes more feasible and practical for immediate implementations. As there are only ten years left to achieve the SDGs, it is an urgent need to develop a method to reduce the demand for data but still provide sufficient information for monitoring the SDG progress across countries and regions.

One solution of reducing data demand for SDG monitoring is to identify a subset of SDG indicators as 'principal indicators', so that the changes of these principal indicators can sufficiently represent the changes of all indicators. Therefore SDG progress can be monitored by only using these principal indicators with 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 of a dataset is reduced by removing some variables without losing valuable information (i.e. variance) (Fodor 2002). Dimensionality reduction requires strong correlations between variables. Indeed, many studies as well as our analysis (supplementary figure S1 (available online at stacks.iop.org/ERL/16/124015/mmedia)) have shown that the SDG goals, targets, and indicators are highly correlated with each other (Nilsson *et al* 2016, Haines *et al* 2017, Liu *et al* 2018, Nerini *et al* 2019). Such correlation indicates that, with appropriate methods, it is possible to extract a small number of principal indicators so that their variations can sufficiently represent the variations of the entire set of SDG indicators.

The central question we aim 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 the minimal effort of data collection. To identify the principal indicators, we apply dimensionality reduction to examine a time-series data of SDG indicators. Two primary methods of dimensionality reduction are principal component analysis (PCA) (Slonim 2002, Ringnér 2008, Brosch *et al* 2018, Kashiwagi *et al* 2019, Spake *et al* 2019) and factor analysis (FA) (Elder *et al* 2018, Eisenberg *et al* 2019, Sandbrook *et al* 2019). 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 (Joliffe and Morgan 1992). For example, the research by Jiang *et al* (2018) found the first principal component can explain up to 85% variation of a set of 28 sustainable development indicators. 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 (Rummel 1988). For instance, Laurett *et al* (2021) concentrated 25 sustainable development-related variables into three factors including natural agriculture, innovation and technology, and environmental aspects using FA. Both PCA and FA identify a smaller number of new variables respectively called principal components and factors, which are linear combinations of all the original variables, to explain most of the variance of the dataset (Joliffe and Morgan 1992). However, monitoring these new variables does not reduce the work needed to collect the whole set of indicators. Therefore, our goal here is to find a subset of the original variables rather than new variables as principal indicators. Studies have already proposed qualitative criteria, such as cost effectiveness, feasibility, indispensability, and relevance, to select essential SDG variables (Reyers *et al* 2017), climate variables (World Meteorological Organization 2016) and biodiversity-related variables (Pereira *et al* 2013). In this study, we quantitatively identify the principal indicators from the whole set of SDG indicators using a hybrid approach by combining PCA and multiple regression (section 2) (Cadima and Jolliffe 2001, Steinmann *et al* 2016).

We examine a World Bank dataset of 351 SDG indicators for 217 countries and regions from 2000 to 2017 (section 2) to find principal indicators that can explain a sufficiently large amount of variance of all SDG indicators. Specifically, this dataset is approximately 42% complete with the portion of missing data ranging from 1% to 98% for individual SDG indicators and 38% to 98% for countries and regions (supplementary figure S2 and table S1). We first determine how to select the best training set from the historical data for identifying the principal indicators for future SDG progress. Next we identify principal indicators that can represent at least 90% of the variance—a benchmark criterion we choose—of all SDG indicators in the past (2000–2017) and are expected to do so in the future (2018–2030) with the lowest difficulty of data collection measured by the share of missing data (section 2).

2. Method

We adopted the 351 World Bank indicators rather than 231 UN indicators for the following two reasons and advantages. First, the two organizations have

deeply collaborated with each other on the SDGs (World Bank 2021), and their two sets of indicators highly overlap with each other (Lusseau and Mancini 2019). Second, compared with UN database, compared with the UN dataset, the World Bank dataset has a more diverse set of indicators which help minimize collinearity issues in subsequent analyses (Lusseau and Mancini 2019).

We use the World Bank dataset of SDG indicators obtained in July 2020 which includes 358 indicators for the 17 SDGs from 1990 to 2017 for 217 countries and regions (non-state entities such as Hong Kong, SAR, China) and 46 country groups (e.g. the Euro area, OECD members, and Least Developed Countries). In this research, we only use data from 2000 to 2017 because data in other years are substantially incomplete (supplementary figure S3). We also exclude seven indicators due to lack of data for 2000–2017 (supplementary table S2). Lastly we only consider data for countries or regions excluding data for country groups. As a result, we 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 (supplementary table S3). We 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.

Using this dataset, we 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 (Jolliffe 2011), we calculate the nearest PSD correlation matrix using 'nearPD' function in R (Thomas 2015). 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, we calculate the explained variance of the subset of k indicators on the entire dataset (X) using the following equation (Cadima and Jolliffe 2001):

$$\mathrm{EP}_{(k,X)} = [\mathrm{corr}\,(X, P_k X)]^2 = \frac{\mathrm{tr}\left(\left[S^2\right]_{(k)} S_k^{-1}\right)}{\mathrm{tr}\,(S)}$$

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 (Cadima *et al* 2004). These k indicators are defined as the principal indicators with size k. We then can identify the smallest number (m) of indicators for any threshold of explained variance (90% in this study). In practice, we use the 'improve' function from the R package 'subselect' (Cadima *et al* 2012) to achieve the largest explained variance. We then select the principal indicators for different missing rate thresholds. Note that we need to select all 351 SDG indicators as principal indicators if we want to represent all SDGs.

We 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 (supplementary figures S4 and S5). We also provide the marginal explained variance to validate the selection of the principal indicators.

To validate that the selected principal indicators are good proxy for the entire dataset, we examine the marginal explained variance of the principal indicators and non-principal indicator. We calculate the marginal explained variance of each individual principal indicator i on the entire dataset ($\mathrm{MEP}_{(i,X)}$), which is the difference between the explained variance of all principal indicators ($\mathrm{EP}_{(k,X)}$) and the explained variance of the principal indicators except the target one ($\mathrm{EP}_{(k-1,X)}$):

$$\mathrm{MEP}_{(i,X)} = \mathrm{EP}_{(k,X)} - \mathrm{EP}_{(k-1,X)}.$$

Similarly, the marginal explained variance of each non-principal indicator j can also be calculated. We 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 we 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 ($\mathrm{MEP}_{(j,X)}$):

$$\mathrm{MEP}_{(j,X)} = \mathrm{EP}_{(v,X)} - \mathrm{EP}_{(u,X)}.$$

An example of validation for 77 principal indicators that can explain 90% of the variance for the entire dataset when we do not consider difficulty of data

Figure 1. Explained variances of the 100 principal indicators identified from various training sets on a test set. Each plot indicates a fixed period between the test set year and the last year of the training set (ΔT). (A)–(F) are results of selecting principal indicators from indicators with missing rate less than 100%, 90%, 80%, 70%, 60%, and 50%, respectively.

collection (missing rate threshold = 100%) is shown in supplementary figures S6 and S7. 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 we do not consider the difficulty of data collection (missing rate threshold = 100%).

3. Best training set

We first examine how much future variance of the SDG indicators can be explained by principal indicators identified from various training sets. Specifically, we 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, we measure how much variance 100 principal indicators can explain for each corresponding test set as a benchmark. Then we vary the number of principal indicators to examine the impact on the explained variance.

Figure 1 shows the explained variance of selected principal indicators in each data split. Each panel (figures 1(A)–(F)) selects principal indicators only from indicators with data missing rate lower

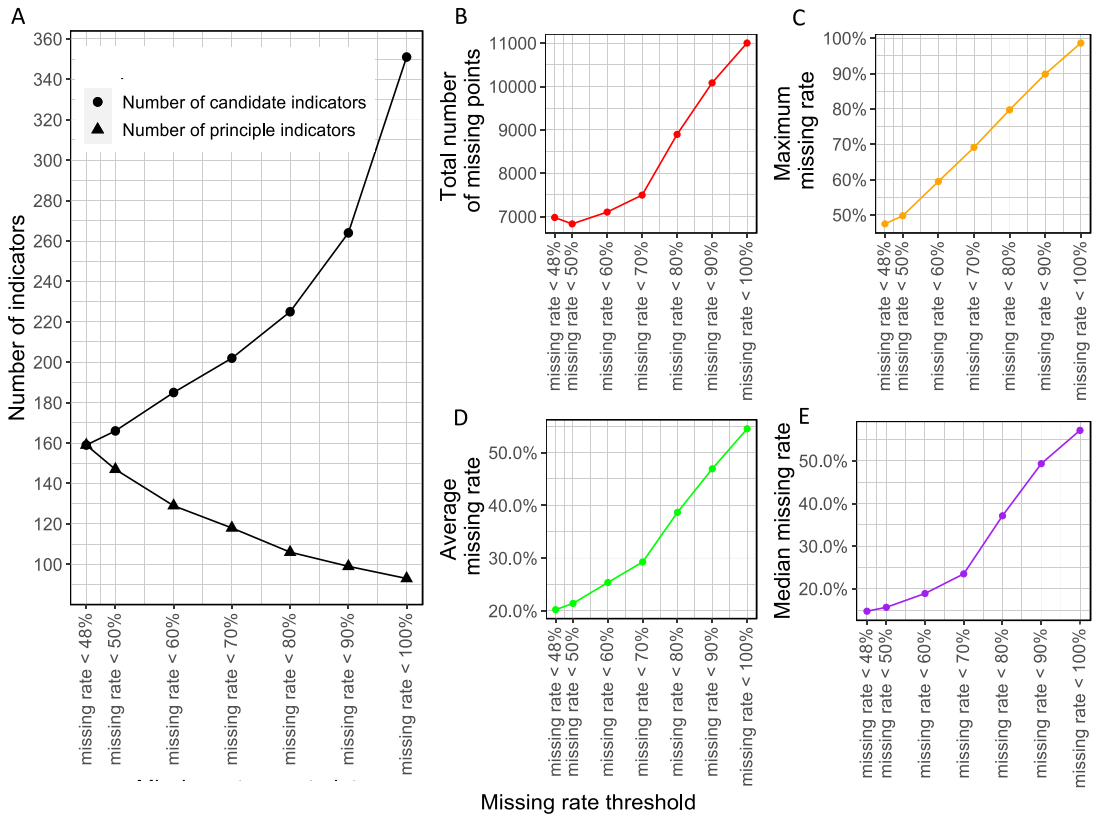

Figure 2. (A) Number of principal indicators to explain at least 90% of the annual variances from 2000 to 2017 under different missing rate thresholds. The line with dots represents the number of candidate indicators that meet the missing rate threshold requirement, and the line with triangles represents 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.

than a threshold. Therefore the threshold of 100% (figure 1(A)) means all indicators will be considered as candidates for principal indicators, implying that we 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 figures 1(B)–(F), the entire historical dataset is the best training set if we consider the difficulty of data collection (missing rate threshold $\neq$ 100%). For example, figure 1(F) shows that, when we 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. We can find similar results when varying the number of principal indicators (supplementary figure S8). Therefore we 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.

4. Principal indicators for past and future SDG progress

Using the entire historical dataset, we 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. We 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 we 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(A), when the missing rate threshold is low, we 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. We need 94 principal indicators to explain at least 90% of the variances when we do not 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 (supplementary figure S9). Note that it is

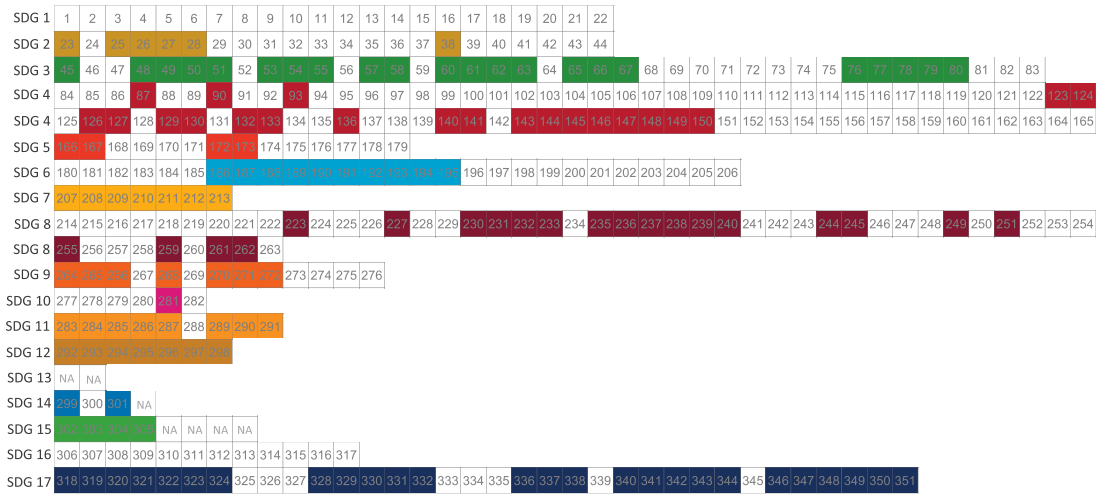

Figure 3. The 147 principal indicators highlighted among the full set of 351 SDG indicators (supplementary table S5).

not possible to explain at least 90% of the variances anymore when the missing rate threshold is less than 48% (not enough candidate indicators).

Figures 2(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 (6832). 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 (supplementary table S2). As a result, we 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) (supplementary figure S10), and has the lowest difficulty of data collection.

Figure 3 highlights the 147 principal indicators among all SDG indicators (supplementary table S5). 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.

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 is associated with poverty (Carter and Danert 2003, Capps *et al* 2016). 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)'.

We expect the SDG 13 (Climate Action) to have close relationship with many other SDGs, especially those related to the environment and quality of life (Hamdi *et al* 2020). However, there are only two indicators for SDG 13 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 (%)'.

5. Discussion and policy implications

We 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 identified principal indicators are for better cross-country (region) comparison at the global scale. In addition, individual country/region can use the identified principal indicators to monitor their SDG performance as these indicators generally have better data availability than non-principal indicators.

The UN identifies invisibility and inequality as the two big global challenges for the current state of SDG data (United Nations 2015) which are primarily due to the large amount of data to collect (high cost) and declining finance (inadequate resources) (Open Data Watch 2016, OECD 2019). Our results can help address these challenges. Specifically, we identify a small set of principal indicators to represent the whole SDG indicator set. There are two novelties in our work. First, we identify the principal indicators using a quantitative approach, while prior work does so qualitatively. Second, we select the principal indicators from the original set of indicators rather than creating new indicators like previous work does. As a result, the principal indicators in our work tend to have better data availability and can sufficiently monitor SDG progress, thus reducing the amount of data needed for effective SDG monitoring. Moreover, with limited and even declining financial resources, investment in SDG data infrastructure needs to be strategic. The principal indicators can be considered as priorities in making such investment, especially for developing countries or regions with substantial data challenges (supplementary table S1).

Our 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, our 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. Most of the countries and regions in our dataset have very similar 'pattern' of missing data rates across indicators during the study period, as 90% of them (194 out of 217) are correlated (correlation coefficients >0.5) with the missing 'pattern' of the indicators of all countries of the latest year (supplementary figure S11). 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. Note that some data are missing because the corresponding indicators are not relevant for some countries/regions (e.g. indicators about marine resources for landlocked countries/regions). However, these indicators in theory should not be considered as principal indicators anyways because they are not representative for all countries/regions.

We also recommend to examine the principal indicators for every couple of years. 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 reexamining the principal indicators every couple of years is necessary to identify those changed correlations and update the principal indicators. But we do not expect the changed correlations will happen very often and dramatic. Our results show that the average difference between 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 is only 2.5% (figure S10), which validates the statement.

To ensure the representativeness of the principal indicators for all SDGs, we can force to select at least one indicator from each SDG as principal indicators for countries or regions with relatively abundant expenditure. 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) (supplementary figures S12 and S13). Therefore we 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.

For future work, building on the principal indicators, we 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 (Xu *et al* 2020). 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.

We set explaining at least 90% of the variance as the benchmark criterion to select the principal indicators. In practice, this criterion needs to be further refined to consider the preference of stakeholders. In addition, our method is based on the correlations between SDG indicators without considering causality. Thus our results are not intended to direct

investment on SDGs themselves, but to guide investment on data infrastructure to monitor SDGs.

Data availability statement

The datasets analyzed in this study are publicly available as referenced within the article. All data and code are available from the corresponding author on request.

The data that support the findings of this study are openly available at the following URL/DOI: https://datatopics.worldbank.org/sdgs/.

Acknowledgment

This study was financially supported by the National Natural Science Foundation of China (72022004).

Conflict of interest

The authors declare no competing interests.

References

Brosch M, Kattler K, Herrmann A, von Schönfels W, Nordström K, Seehofer D, Damm G, Becker T, Zeissig S and Nehring S 2018 Epigenomic map of human liver reveals principles of zonated morphogenic and metabolic control *Nat. Commun.* **9** 4150

Cadima J F and Jolliffe I T 2001 Variable selection and the interpretation of principal subspaces *J. Agric. Biol. Environ. Stat.* **6** 62

Cadima J, Cerdeira J O and Minhoto M 2004 Computational aspects of algorithms for variable selection in the context of principal components *Comput. Stat. Data Anal.* **47** 225–36

Cadima J, Cerdeira J O, Silva P D and Minhoto M 2012 The subselect R package

Capps K A, Bentsen C N and Ramírez A 2016 Poverty, urbanization, and environmental degradation: urban streams in the developing world *Freshwater Sci.* **35** 429–35

Carter R C and Danert K 2003 The private sector and water and sanitation services—policy and poverty issues *J. Int. Dev.* **15** 1067–72

Eisenberg I W, Bissett P G, Enkavi A Z, Li J, MacKinnon D P, Marsch L A and Poldrack R A 2019 Uncovering the structure of self-regulation through data-driven ontology discovery *Nat. Commun.* **10** 2319

Elder C D, Xu X, Walker J, Schnell J L, Hinkel K M, Townsend-Small A, Arp C D, Pohlman J W, Gaglioti B V and Czimczik C I 2018 Greenhouse gas emissions from diverse Arctic Alaskan lakes are dominated by young carbon *Nat. Clim. Change* **8** 166

Espey J 2019 Sustainable development will falter without data *Nature* **571** 299–300

Fodor I K 2002 *A Survey of Dimension Reduction Techniques* (Livermore, CA: Lawrence Livermore National Laboratory)

Fritz S, See L, Carlson T, Haklay M M, Oliver J L, Fraisl D, Mondardini R, Brocklehurst M, Shanley L A and Schade S 2019 Citizen science and the United Nations sustainable development goals *Nat. Sustain.* **2** 922–30

Global partnership for sustainable development data (available at: www.data4sdgs.org/partner-listing)

Haines A, Amann M, Borgford-Parnell N, Leonard S, Kuylenstierna J and Shindell D 2017 Short-lived climate pollutant mitigation and the sustainable development goals *Nat. Clim. Change* **7** 863

Hamdi R, Kusaka H, Doan Q-V, Cai P, He H, Luo G, Kuang W, Caluwaerts S, Duchêne F and van Schaeybroek B 2020 The state-of-the-art of urban climate change modeling and observations *Earth Syst. Environ.* **4** 1–16

Hsu A, Malik O, Johnson L and Esty D C 2014 Development: mobilize citizens to track sustainability *Nature* **508** 33

Jiang Q, Liu Z, Liu W, Li T, Cong W, Zhang H and Shi J 2018 A principal component analysis based three-dimensional sustainability assessment model to evaluate corporate sustainable performance *J. Clean. Prod.* **187** 625–37

Joliffe I and Morgan B 1992 Principal component analysis and exploratory factor analysis *Stat. Methods Med. Res.* **1** 69–95

Jolliffe I 2011 *Principal Component Analysis* (Berlin: Springer)

Kashiwagi Y, Higashi T, Obashi K, Sato Y, Komiyama N H, Grant S G and Okabe S 2019 Computational geometry analysis of dendritic spines by structured illumination microscopy *Nat. Commun.* **10** 1285

Laurett R, Paco A and Mainardes E W 2021 Measuring sustainable development, its antecedents, barriers and consequences in agriculture: an exploratory factor analysis *Environ. Dev.* **37** 100583

Liu J, Hull V, Godfray H C J, Tilman D, Gleick P, Hoff H, Pahl-Wostl C, Xu Z, Chung M G and Sun J 2018 Nexus approaches to global sustainable development *Nat. Sustain.* **1** 466

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

Lusseau D and Mancini F 2019 Income-based variation in sustainable development goal interaction networks *Nat. Sustain.* **2** 242–7

Lyytimäki J 2019 Seeking SDG indicators *Nat. Sustain.* **2** 646

Nerini F F, Sovacool B, Hughes N, Cozzi L, Cosgrave E, Howells M, Tavoni M, Tomei J, Zerriffi H and Milligan B 2019 Connecting climate action with other sustainable development goals *Nat. Sustain.* **1**

Nilsson M, Griggs D and Visbeck M 2016 Policy: map the interactions between sustainable development goals *Nature* **534** 320

OECD 2019 Global outlook on financing for sustainable development 2019: time to face the challenge

Open Data Watch 2016 The state of development data 2016 (available at: https://opendatawatch.com/the-state-of-development-data-2016/) (Accessed October 2019)

Pereira H M, Ferrier S, Walters M, Geller G N, Jongman R, Scholes R J, Bruford M W, Brummitt N, Butchart S and Cardoso A 2013 Essential biodiversity variables *Science* **339** 277–8

Reyers B, Stafford-Smith M, Erb K-H, Scholes R J and Selomane O 2017 Essential variables help to focus sustainable development goals monitoring *Curr. Opin. Environ. Sustain.* **26** 97–105

Ringnér M 2008 What is principal component analysis? *Nat. Biotechnol.* **26** 303

Rummel R J 1988 *Applied Factor Analysis* (Evanston, IL: Northwestern University)

Sandbrook C, Fisher J A, Holmes G, Luque-Lora R and Keane A 2019 The global conservation movement is diverse but not divided *Nat. Sustain.* **2** 316

Slonim D K 2002 From patterns to pathways: gene expression data analysis comes of age *Nat. Genet.* **32** 502

Spake R, Bellamy C, Graham L J, Watts K, Wilson T, Norton L R, Wood C M, Schmucki R, Bullock J M and Eigenbrod F 2019 An analytical framework for spatially targeted management of natural capital *Nat. Sustain.* **2** 90

Steinmann Z J, Schipper A M, Hauck M and Huijbregts M A 2016 How many environmental impact indicators are needed in the evaluation of product life cycles? *Environ. Sci. Technol.* **50** 3913–9

Thomas K 2015 The lmf R package

United Nation System 2017 Total expenditure (available at: www.unsystem.org/content/FS-F00-05) (Accessed October 2019)

United Nations 2015 A world that counts: mobilising the data revolution for sustainable development (available at: www.undatarevolution.org/report/) (Accessed December 2019)

United Nations 2016 Transforming our world: the 2030 agenda for sustainable development

Watmough G R, Marcinko C L, Sullivan C, Tschirhart K, Mutuo P K, Palm C A and Svenning J-C 2019 Socioecologically informed use of remote sensing data to predict rural household poverty *Proc. Natl Acad. Sci.* **116** 1213–8

World Bank 2020 Global economic prospects (available at: www.worldbank.org/en/publication/global-economic-prospects#firstLink01623) (Accessed October 2020)

World Bank 2021 World Bank Group and the 2030 Agenda

World Meteorological Organization 2016 Essential climate variables (available at: https://public.wmo.int/en/programmes/global-climate-observing-system/essential-climate-variables) (Accessed September 2021)

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

Yeh C, Perez A, Driscoll A, Azzari G, Tang Z, Lobell D, Ermon S and Burke M 2020 Using publicly available satellite imagery and deep learning to understand economic well-being in Africa *Nat. Commun.* **11** 1–11