THE WORLD BANK

ESG GAPS RESEARCH

List of Tables

List of Figures

Introduction

This document is the continuation of the paper *Options for Improving Use of ESG data for Sovereign Bond Analysis* (World Bank 2018). Following interaction with investors in sovereign bonds that use ESG indicators in their country analyses and risk/return profiles of sovereign securities, the World Bank Group (WBG) presented a set of options for improving accessibility, quality (e.g. timeliness and regularity of publication, geographic coverage) and transparency of Emerging Markets data, in particular for ESG data. This paper aims to continue the analysis, better understand underlying data production and management issues that affect availability and provide recommendations for improving the accessibility, quality and coverage of ESG indicators.

Team

This document was written and developed by Tim Herzog, R.Andres Castaneda Aguilar, and Tony Fujs. Andrei Ilas and Hiroko Maeda contributed substantial research support.

Replicability and license

This document is fully replicable. All the text files, codes, underlying data, and dashboards can be found in its GitHub repository worldbank/ESG_gaps_research.

1
Background

This document builds upon research first presented in the discussion paper *Options for Improving Use of ESG for Sovereign Bond Analysis* (World Bank, 2018).

Interviews with ESG data providers found that most obtain at least some and often a substantial amount of data from World Bank databases. 137 indicators were specifically identified from these interviews, of which 127 could be mapped to active databases in the World Bank's data API, enabling the authors to perform a rapid assessment of data coverage and gaps, as included in the 2018 discussion paper. The 2018 paper found that "data coverage is a significant issue in WBG data used for ESG." Looking at most recent available values (MRVs) by indicator and country, the paper found that just 41 ESG indicators (out of 127) had a value from 2017 or later (1 year old) for at least 50% of countries; 74 ESG indicators had a value from 2015 or later (3 years old) for at least 50% of countries. However, while the 2018 paper suggested a set of options for improving the availability and usefulness of ESG data, it stopped short of further investigating the reasons that might give rise to gaps in data coverage or suggesting specific strategies to address them.

The objective of this report is to pick up where the previous paper left off, and to better understand the circumstances that explain gaps in data coverage. The hope is that with a better understanding of why gaps occur and the significance of various explanatory factors, effective steps can be established to eliminate or mitigate gaps, and better understand which kinds of gaps are most relevant for ESG analysis.

The study set of ESG indicators in this report is different than the one in the 2018 report. Whereas the 2018 report excluded indicators used by providers that the Bank no longer actively maintains, this report includes those since they are relevant to the analysis. Additionally, this report includes indicators from various products introduced since the 2018 report, including the World Bank's own curated ESG dataset. Conversely, we decided to remove a subset of indicators used

by a single provider because the strong similarities among them (e.g., very similar trade or debt measures) were resulting in double-counting that could potentially skew the findings. Accordingly, this report is based on a body of 134 indicators, compared to 137 in the 2018 report.

The other major difference between the indicators in the two reports is that many of them have been updated since the 2018 report was completed. Many statistical indicators have been updated several times. Accordingly, if the 2018 analysis were re-run using the indicators from this analysis, the findings would likely be quite different, and different yet again if the analysis were run a year hence. One of the goals of this analysis is to provide a framework for thinking about data availability and coverage that is reasonably independent of the data curation cycles for the indicators under study.

1.1 What is a "Data Gap"?

The term "data gap" is somewhat ambiguous, so we should start by discussing what kinds of gaps can exist in datasets. For instance, data could be unavailable for a number of relevant economies, or there could be gaps in the time series over a relevant time period. There could also be gaps in metadata and other documentation. Data could also simply be unavailable or undefined for important concepts (such as "resilience"), necessitating the use of data proxies. While all of these are potentially relevant, the most important gaps in the context of ESG likely involve the most recently available values compared to the current time period, since ESG analysis concerns investment decisions being made today and in the near future. Accordingly, this paper defines a "data gap" as a significant difference between the current calendar year and the most recent available value(s) (MRVs) for the indicators and economies under study. Gaps in metadata or in time periods before the MRV are not a primary focus of this analysis.

1.2 How This Paper is Structured

This paper applies three separate approaches to better understand coverage gaps in ESG indicators:

1. **Coverage Analysis.** This approach provides a more detailed and visual picture of temporal gaps in the ESG indicator set, both historically and by MRV.
2. **Explanation Framework Analysis.** This approach sets out a set of reasons why data gaps might occur as a framework for classifying ESG indicators, and looking at what approaches might be used to mitigate coverage issues.

3. **Variance Analysis.** This approach looks at the temporal variance of ESG indicators to better understand the impacts of missing data for analysis. It may be possible to impute missing observations for indicators with low variance, mitigating the impact of data gaps.

The paper then concludes with a discussion section and set of recommendations based on the analysis and findings of each of these sections.

1.3 About the Data Used in This Report

The indicator database used in this report consists of 134 indicators extracted from the World Development Indicators and other World Bank Databases in October, 2019.

Table 1.1 provides a summary of the 134 indicators analyzed in this report disaggregated by pillar and origin. 44 indicators are environmental indicators, 66 are social indicators, and 24 are governance indicators. The World Bank is the primary source of 36 indicators, whereas the UN system is primary source of 66 indicators, and other organizations are the source for 32 indicators.

Table 1.1: Indicators by origin and sector

Origin	Env	Soc	Gov	Total
WBG	10	12	14	36
UN System	9	49	8	66
Other orgs	25	5	2	32
Total	44	66	24	134

Unless otherwise noted, the study period is limited to 2000-2018 since collection of 2019 data was still in its early stages at the time of compilation. 4 indicators include only projections data for the year 2050, and thus have been excluded from analysis unless otherwise noted. Another 15 indicators have been dropped or deprecated and, except as noted in the chapter on "Explanation Framework Analysis," have also been excluded, leaving 115 indicators as the primary focus of analysis.

2
Coverage Analysis

While the 2018 paper looked at coverage gaps primarily in terms of most recent available values (MRVs), in this analysis we wanted to develop a more detailed approach to identify different types of coverage gaps over a broader time span. Accordingly, we developed the heat map shown in Figure 2.1. In this chart, discrete indicators are arranged along the Y axis while time is plotted on the X axis for the 2000-2018 period. Colors indicate the number of observations (i.e., countries) for the corresponding indicator and year. Darker colors in the purple part of the spectrum indicate relatively low-density coverage, while lighter colors in the yellow part of the spectrum represent high-density coverage, up to the maximum of 217 countries. Blank areas indicate no data for that particular indicator and year.

Several patterns emerge from a visual assessment of Figure 2.1, which are not necessarily mutually exclusive, nor is a visual assessment the only approach to identifying indicator clusters. Appendix 2 includes several alternatives to and in-depth definitions of the patterns discussed in this section.

2.1 Consistently low coverage

One group of indicators is characterized by consistently low country coverage over the 2000-2018 time period. In this case we've defined "low coverage" has having values for no more than 100 countries in any given year. These indicators generally appear as steady and consistently dark horizontal lines towards the bottom of Figure 2.1.

It's important to note, however, that while total country coverage may be low for these indicators in any given year, the country composition often varies from year to year for reasons discussed in the next section. For example, "Poverty Headcount Ratio" is available for no more than 59 countries in any given year, but includes values for 135 countries across all years. By comparison, "Incidence of Malaria" is available for 99 countries in nearly all years with very little variation

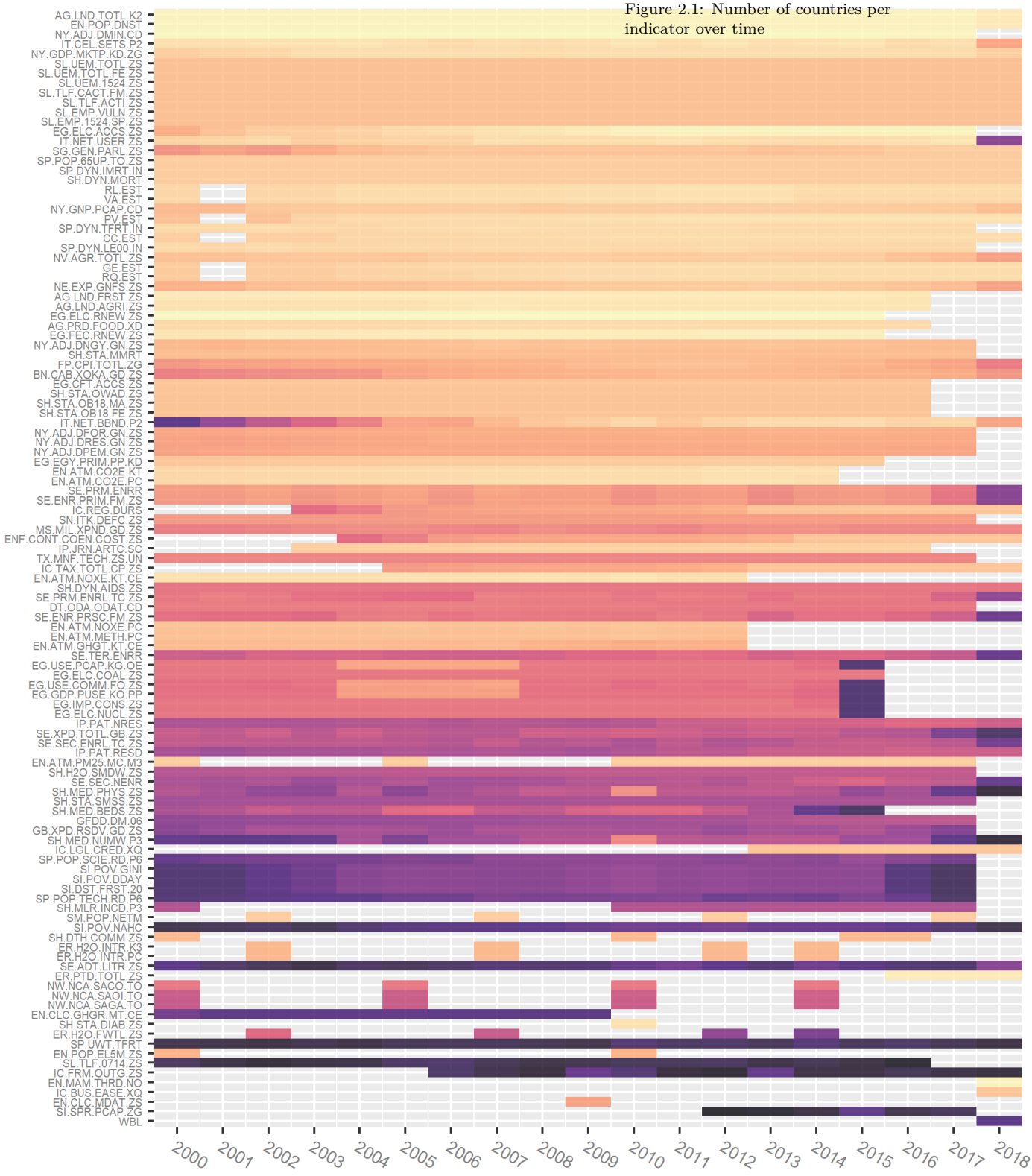

Figure 2.1: Number of countries per indicator over time

Table 2.1: Indicators with consistently low coverage

Indicator	Code	Max. Countries
Children in employment, total (% of children ages 7-14)	SL.TLF.0714.ZS	31
Unmet need for contraception (% of married women ages 15-49)	SP.UWT.TFRT	37
Retirement Age	WBL	42
Annualized average growth rate in per capita real survey mean consumption or income, total population (%)	SI.SPR.PCAP.ZG	46
Value lost due to electrical outages (% of sales for affected firms)	IC.FRM.OUTG.ZS	51
GHG net emissions/removals by LUCF (Mt of CO2 equivalent)	EN.CLC.GHGR.MT.CE	58
Poverty headcount ratio at national poverty lines (% of population)	SI.POV.NAHC	59
Technicians in R\&D (per million people)	SP.POP.TECH.RD.P6	69
Literacy rate, adult total (% of people ages 15 and above)	SE.ADT.LITR.ZS	71
Income share held by lowest 20%	SI.DST.FRST.20	84
Poverty headcount ratio at $1.90 a day (2011 PPP) (% of population)	SI.POV.DDAY	84
GINI index (World Bank estimate)	SI.POV.GINI	84
Researchers in R\&D (per million people)	SP.POP.SCIE.RD.P6	84
People using safely managed sanitation services (% of population)	SH.STA.SMSS.ZS	94
Research and development expenditure (% of GDP)	GB.XPD.RSDV.GD.Z	99
Incidence of malaria (per 1,000 population at risk)	SH.MLR.INCD.P3	99

in any given year. These patterns may be important for ESG analysis if it is possible to extrapolate or impute missing values from prior years; indicators whose countries vary from year-to-year (and thus have larger coverage in the aggregate) may benefit to a greater degree.

2.2 Moderate to high coverage

By contrast, most indicators include values for at least 100 countries in at least one year. 99 indicators have single-year coverage of at least 100 countries, 78 indicators cover at least 150 countries, and 27 indicators cover at least 200 countries. In Figure 2.1 these indicators range from magenta to light yellow in the middle to upper sections of the heat map.

12 indicators in this cluster include values for 2018 or later for at least 90% of countries. These indicators appear at the top-most section of Figure 2.1 and correspond to the "perfect case" classification in the next chapter.

Among the remaining indicators, year-to-year composition of coverage can vary in a manner similar to those in the "consistently low coverage" group for methodological reasons, as discussed in the next section.

2.3 Measurable improvement

A handful of indicators demonstrate significant, measureable improvement in country coverage over time. We define "measurable improvement" by regressing country coverage over time for each indicator (gap-filling for years where coverage is missing entirely). Indicators with a coefficient greater than 1 are shown in Table 2.2. In Figure 2.1 these appear as indicators that are colored dark purple or magenta on the left side of their coverage with increasingly light colors on the right side.

These and similar indicators may warrant further study to better understand the factors behind the increases in country coverage. For instance, if country coverage improved as a result of better methodologies, increased production capacity, or broader demand, they may provide a model for improving country coverage for indicators that need it.

2.4 High coverage and sudden decline

A significant group of indicators has consistent coverage through most of the time period, but with declining coverage or no coverage in recent years. In Figure 2.1 these tend to appear as "truncated" series

Table 2.2: Select indicators that improve over time

Indicator	Code
Current account balance (% of GDP)	BN.CAB.XOKA.GD.ZS
Access to electricity (% of population)	EG.ELC.ACCS.ZS
Enforcing contracts: Cost (% of claim)	ENF.CONT.COEN.COST.ZS
Outstanding international public debt securities to GDP (%)	GFDD.DM.06
Time required to start a business (days)	IC.REG.DURS
Total tax and contribution rate (% of profit)	IC.TAX.TOTL.CP.ZS
Patent applications, nonresidents	IP.PAT.NRES
Patent applications, residents	IP.PAT.RESD
Fixed broadband subscriptions (per 100 people)	IT.NET.BBND.P2
Literacy rate, adult total (% of people ages 15 and above)	SE.ADT.LITR.ZS
Proportion of seats held by women in national parliaments (%)	SG.GEN.PARL.ZS
Annualized average growth rate in per capita real survey mean consumption or income, total population (%)	SI.SPR.PCAP.ZG

with no coloring for large portions of the right side of the chart. Table 2.3 summarizes indicators by the the year of their most recent available value (MRV). As shown, over 50% of ESG indicators in the study period have no values for the most recent study year, and 13% of indicators have no values for the most recent four years or more.

Table 2.3: Indicators by Most Recent Available Year

Year of MRV	# Indicators
2018+	55
2017	25
2016	10
2015	10
<2015	15

As noted previously, MRV years are a significant factor in ESG data use, as older data is less relevant to investment decisions being made today and in the near future. Many important factors could explain the wide variance in MRV years, and this is the focus of the next chapter.

2.5 Intermittent coverage

A handful of indicators are only available for periodic years with no values available for intermediate years. These appear in Figure 2.1 as as intermittent series resembling "dashed" lines, the majority (but not all) of which are environmental indicators. Appendix 2 provides a technical description of indicators in this category.

While not the primary focus of this analysis, there are a handful of factors that could explain the coverage characteristics of this group. The most obvious explanation is that indicators may simply not be designed as time-series data. This is the most likely explanation for Retirement Age and Threatened Mammal Species, which are available for only a single year. In other cases, there may not be resources to collect data on an annual basis, even if doing so would be useful. Other indicators may measure environmental or social phenomena that change gradually so that annual data collection would not be efficient. This last possibility is material to ESG data use because it implies that older data may still be relevant if properly understood.

Table 2.4: Indicators with intermittent coverage

Indicator	Code
Droughts, floods, extreme temperatures (% of population, average 1990-2009)	EN.CLC.MDAT.ZS
Mammal species, threatened	EN.MAM.THRD.NO
Population living in areas where elevation is below 5 meters (% of total population)	EN.POP.EL5M.ZS
Annual freshwater withdrawals, total (% of internal resources)	ER.H2O.FWTL.ZS
Renewable internal freshwater resources, total (billion cubic meters)	ER.H2O.INTR.K3
Renewable internal freshwater resources per capita (cubic meters)	ER.H2O.INTR.PC
Natural capital, subsoil assets: coal (constant 2014 US$)	NW.NCA.SACO.TO
Natural capital, subsoil assets: gas (constant 2014 US$)	NW.NCA.SAGA.TO
Natural capital, subsoil assets: oil (constant 2014 US$)	NW.NCA.SAOI.TO
Cause of death, by communicable diseases and maternal, prenatal and nutrition conditions (% of total)	SH.DTH.COMM.ZS
Diabetes prevalence (% of population ages 20 to 79)	SH.STA.DIAB.ZS
Net migration	SM.POP.NETM
Retirement Age	WBL