
Pilot 2006 Environmental Performance Index

Yale Center for Environmental Law & Policy
Yale University

Center for International Earth Science Information Network (CIESIN)
Columbia University

In collaboration with

World Economic Forum
Geneva, Switzerland

Joint Research Centre of the European Commission
Ispra, Italy

Pilot 2006 Environmental Performance Index

Appendix F: Methodology & Measurement Challenges

Appendix F: Methodology & Measurement Challenges

The Pilot 2006 EPI introduces a policy-relevant framework for environmental performance assessment. The framework depends on the data it contains. While the methodology of the EPI is quite straightforward (as explained in Chapter 2), calculation of the EPI scores and rankings requires considerable numerical alignment and statistical processing. The purpose of this Appendix is to provide a detailed description of the steps included in calculating the EPI and of the statistical techniques and methods used. We offer this detail because we believe that transparency is an essential foundation for good analysis and policymaking.

The issues addressed in the following sections mirror those commonly encountered in the computation of composite indices: indicator and country selection, missing data treatment, standardization, aggregation and weighting methodologies, as well as performance testing (OECD, 2003).

F.1. Country Selection Criteria

While the data metrics for the 16 indicators contain information for as many countries as possible, the EPI contains only those countries with complete data coverage across all indicators and policy categories. There are two exceptions to this rule. First, data availability for two indicators—Overfishing and the Timber Harvest Rate—depends on a country's geographical location. Therefore, landlocked countries without data for the overfishing indicator and countries with no natural or planted forests are included in the EPI if they are not missing any other data. The second exception applies to two indicators found in the Environmental Health policy category: access to improved drinking water and access to sanitation. The very high correlation between these indicators permits us,

in the event that one of the data points is missing, to use the available data point as a proxy for the missing one. A further discussion on the treatment of missing data is given in the next section.

F.2. Missing Data

Data gaps remain a very serious obstacle to a more refined EPI and to data-driven policymaking more generally. Persistent data gaps or incomparability of data across countries means that several important policy challenges cannot presently be addressed. And many countries, particularly in the developing world, lack data on a number of critical indicators.

For example, air quality indicators based on ground-monitoring are simply not available for many developing countries and are further limited by weak data comparability even in developed countries. Pollutants such as lead, ultra-fine particulate matter (PM_{2.5}), tropospheric ozone, and volatile organic compounds (VOCs) do not have sufficient ground observations available and are not updated on a sufficiently frequent basis to permit robust performance metrics. Although satellite-based observation of air pollutants is advancing rapidly and provides more reliable estimates to fill in the gaps, availability and use of these technologies is still constrained. The result of these data gaps and inconsistencies is that only measures of ground-level ozone and particulates are included in the Pilot 2006 EPI to represent air pollution. These inadequacies point to the need for increased national and international focus on this data situation, specifically with regard to better air quality measures.

Missing data are a major source of uncertainty in index construction. Although statistical methods exist for imputing missing data, they are not free of assumptions regarding the causes for the missing values. In addition, application of these methods requires knowledge and careful consideration of the strengths and weaknesses of various techniques in light of the available data. To continue the air pollution example, such data are highly dependent on spatial and temporal conditions, which complicate the development of imputation models that are applicable to different regions and countries.

Because of the lack of robust, well-tested imputation models, missing data are not imputed in the Pilot 2006 EPI, with the exception of the Drinking Water and Adequate Sanitation indicators. These two measures were found to correlate so strongly with each other that one parameter can justifiably be used to estimate missing values in the other. In the future, as data quality improves and time series data becomes available, further investigation will address the use of imputation models to increase the geographical coverage of the EPI. But the essence of the EPI – as a gauge of actual environmental results – requires particular confidence that any numbers imputed reflect on-the-ground circumstances and outcomes.

Because of the limited data quality, the coverage of countries for the non-imputed indices is necessarily smaller than if missing data had been imputed. The EPI's stringent data requirements mean that the data presented and the analysis derived from them are free of the uncertainty that arises when missing data are imputed. In addition, the authors believe that at present, performance analysis benefits more from a conservative approach to data availability than from the application of sophisticated but untestable missing data imputation models.

As the understanding of the drivers of superior environmental performance grows over time, it is anticipated that statistical modeling of missing data may become more appropriate in the context of performance measurement.

Another important requirement of performance measurement is the ability to measure all relevant environmental policy areas. Several additional theoretically important environmental indicators were identified for inclusion in an ideal performance index, however, these could not be included due to the practical limitations noted above. Most importantly, data are often not measured widely enough or with a sufficient degree of methodological consistency to be useful within the context of a broad analysis. Exposure-effect indicators for many important environmental toxins belong in this category. To compensate for this information gap, proxy indicators that link exposure and outcomes are used, including increased exposure to toxins and increased mortality.

An additional challenge arises from the difficulty of determining clear sustainability targets for some of the indicators. For example, setting targets for mortality rates due to environmental factors requires far-reaching assumptions about a range of health and socio-economic parameters. The specification of targets is discussed in Chapter 2 of the main report.

The urgent need to improve the availability and quality of policy-relevant environmental indicators cannot be overemphasized. Effective environmental policy requires dependable and timely data, not only to identify problems, but also to monitor implementation of response measures, and to follow-up on their effectiveness. Time-series data is also crucial in this regard, allowing for cause-effect analyses and the illumination of best practices with respect to pressing environmental problems.

F.3. Calculation of the EPI and Policy Category Sub-Indices

Indicator Transformation for Cross-Country Comparisons

The raw data for each of the 16 indicators requires standardization to permit cross-country comparisons and to ensure that no indicator dominates the aggregated EPI and policy indices. The other main objective of standardization is to convey information about a country's environmental performance in an easy-to-understand and meaningful way. Thus, we used a proximity-to-target approach that evaluates how close a country is to a desirable performance target for each of the 16 indicators. The choice of the targets is based on sustainability criteria and expert judgments, and in some cases, such as CO₂, had to be based on pragmatic realities rather than ideal goals.

To calculate proximity-to-target values, each indicator is first converted to point in the same direction so that higher values correspond to better performance. Then, the observed values are winsorized at the lowest fifth percentile. Winsorization means that all values falling below the fifth percentile are set to the value corresponding to the fifth percentile. The logic for this approach is to prevent a few extremely low values from skewing the indicator's distribution and exerting an unacceptably high influence on the EPI.

Similarly, countries exceeding the specified target for an indicator are not given additional credit but rather have their value set to the target. This form of "target winsorization" is done to reduce the ability of countries to use above-target performance in one area to make up for poor performance on other indicators. Since the majority of targets also reflect sustainability criteria, overachievement is not desirable with respect to the efficient

deployment of a country's resources. In some cases, moreover, above-target results may be a function of data anomalies or reporting errors.

Following the winsorization of the upper and lower tails of the indicators, proximity to target is calculated as the difference between the observed value and the target divided by the range between the worst observed value and the target. Calibration of the results to the interval zero to 100 then allows interpretation of a country's performance as the shortfall from achieving the target expressed in percent. For example, a country's score of 80 for the Drinking Water indicator means that it is 20% short of meeting the target; in this case 20% of the population do not have access to drinking water.

Since the standardization only transforms the indicator data to fall into the interval zero to 100 but does not alter the spread, i.e., the range of values covered in this interval, the indicators contribute differently to the aggregated policy and EPI scores. We are, therefore, testing an alternative transformation methodology, which aims to stabilize the variation in the data prior to standardization. The Box-Cox family of transformations is designed to estimate the transformation parameter that moves the data distribution closest to normality. The by-product of transformation to a more normal distribution is variance stabilization since the variance does not depend on the expected value. Once complete, this approach will be made available on the EPI website at www.yale.edu/epi.

Data Aggregation and Weighting

Aggregation is always a potential area of methodological controversy in the field of composite index construction. The choice of the two broad objectives, the six policy categories, and the 16 indicators, as well as the EPI's aggregation methodology, are grounded on:

extensive consultations with indicator experts, scientists, and public policymakers from national and international organizations; analyses of existing performance measurement initiatives (most notably the Millennium Development Goals); and detailed literature reviews.

Composite indices are aggregations of sets of variables for the purpose of meaningfully condensing large amounts of information. Various aggregation methods exist and the choice of an appropriate method depends on the purpose of the composite indicator as well as the nature of the subject being measured.

Appropriate choice of the components of composite indices and their weights is an important part of the aggregation process.

For the EPI, we decided to combine a statistical method with a policy-oriented expert judgment approach for deriving the composition of the EPI indicators and their respective weights. Principal component analysis (PCA) was carried out on the proximity-to-target data to identify

which indicators form natural dimensions of environmental performance and how much each indicator contributes to its component.

The results of the PCA were astonishingly clear and appealing from an environmental policy perspective. Of the six dimensions with eigenvalues larger than one (see Box F1 for a description of the concept underlying PCA), three major dimensions emerged: (1) Environmental Health, which represents the impacts of environmental degradation on human health and well-being and contains the Urban Particulates, Indoor Air Pollution, Drinking Water, and Child Mortality indicators, (2) Sustainable Energy, encompassing the indicators measuring Energy Efficiency, Renewable Energy, and CO₂ per GDP, and (3) Biodiversity and Habitat, covering the indicators Water Consumption, Timber Harvest Rate, Wilderness Protection, and Ecoregion Protection.

Box F1: Principal Component Analysis

Principal component analysis (PCA) is a statistical method for dimension reduction through identification of patterns inherent in a multivariate model. It is a useful tool to investigate the relationships between the 16 indicators in the EPI. PCA summarizes a p -dimensional dataset into a smaller number, q , of dimensions while preserving the variation in the data to the maximum extent possible. The q new dimensions are constructed such that:

1. They are linear combinations of the original variables.
2. They are independent of each other.
3. Each dimension captures a successively smaller amount of the total variation in the data.

The objective is to capture those features in the data that help better understand an issue of interest or to discover interesting new patterns among the relationships between variables.

The p original variables are combined into q linear combinations, which form the new principal components of the system. A standardized linear combination Z_1 of a data vector, $X_1=(X_{11}, X_{12}, \dots, X_{1p})$, of length p is defined as:

$$Z_1=w_1^t X_1, \text{ where the sum of the squares of the weights, } w_i, \text{ is 1.}$$

PCA chooses the weights by determining the linear combination of all p variables in the transformed dataset that maximizes the variance of the data. That is, the vector w of weights is calculated such that the squared difference of the new variable values and their respective means is maximized in relation to the total variance of the untransformed data.

The results for w_1 determine the first principal component. The second principal component with weights w_2 is then obtained analogously by maximizing the variance orthogonal to the direction of the first component, and so forth. Orthogonality of the principal components means that they are statistically independent so that any changes in one component do not impact the others. This is sometimes a desirable feature of composite indicators.

The consecutive process of maximizing residual variance implies that at every step less variance is remaining. Once it falls below a specified threshold, the procedure is halted and no more additional principal components are calculated. Several criteria exist to determine the threshold value. One method considers the eigenvalues of the data matrix. The eigenvalue, λ , is the value that solves the equation:

$$X_{corr} a = \lambda a,$$

where X_{corr} is the ($p \times p$) correlation matrix calculated from the data for n countries and p variables and a is a vector in $\mathbb{R}^p \neq 0$.

Values of λ less than 1 indicate that there is no gain to be expected from adding the principal component to the set of selected components. The first $(j-1)$ components are sufficient to summarize the data.

Each principal component provides a set of factor loadings of the indicators, which correspond to their importance for the component, i.e., the higher the loading of an indicator, the more useful it is for explaining variation in the direction of the principal component. Indicators with similarly large loadings on the same principal component can be interpreted as being related along the direction of this component. The loadings from the principal component analysis can also be treated as inherent weights of the indicators for the aggregation process.

The fourth through sixth components explained less variation (i.e., structure) in the data and hence were more ambiguous in their interpretation as policy areas. For this reason, we chose to combine the first three principal components with three policy categories formed by expert judgment. These latter components are titled Water Resources, Air Quality, and Productive Natural Resources.

The Water Resources category consists of indicators for nitrogen loading and over-subscription of water resources. The Air Quality category is comprised of measures for ground-level ozone and particulates, and the Productive Natural Resources category evaluates timber harvesting rate, negative agricultural subsidies, and the extent of overfishing. For landlocked countries the Overfishing indicator is waived. We also note that three indicators contribute to

two policy categories, respectively. In each case, the indicator is a distinct contributor to both human health and ecological vitality. The Urban Particulates measure is important to Environmental Health and Air Quality. Water Consumption affects both Water Resources and Biodiversity and Habitat categories, while the Timber Harvest Rate indicator contributes to Biodiversity and Productive Natural Resources.

For each country in the EPI, the six policy categories are, therefore, calculated as the weighted averages of their constituent indicators. Environmental Health, Sustainable Energy, and Biodiversity and Habitat use PCA derived weights. Water Resources, Air Quality, and Productive Natural Resources use equal weights. The weights from the PCA are given in Table F1.

Table F1: PCA Derived Weights of the EPI Indicators.

Policy Category	Indicator	PCA-derived weight
Environmental Health	Urban Particulates	0.539401
	Indoor Air Pollution	0.900439
	Drinking Water	0.905929
	Adequate Sanitation	0.908663
	Child Mortality	0.888496
Sustainable Energy	Energy Efficiency	0.804238
	Renewable Energy	0.192102
	CO ₂ per GDP	0.868776
Biodiversity and Habitat	Water Consumption	0.154027
	Timber Harvest Rate	0.355348
	Wilderness Protection	0.920753
	Ecoregion Protection	0.905158

The Pilot Environmental Performance Index is then calculated as the weighted average of the six policy categories. The weighting of the categories mirrors the distinct policy sectors and responsibilities within government allocated to human health and ecological integrity. The overarching importance of an intact and healthy environment for human health and well-being is reflected in the higher weight of 50% given to this category. The remaining policy categories are each weighted at 10%, so that the final EPI is calculated as:

$$EPI = 0.5 \times \text{Environmental Health} + 0.1 \times (\text{Air Quality} + \text{Water Resources} + \text{Productive Natural Resources} + \text{Biodiversity and Habitat} + \text{Sustainable Energy}).$$

F.4. Data Quality and Coverage

The EPI should be seen as a *pilot* index because a number of serious data gaps and methodological questions remain open. Data gaps relate to both the lack of available information on important environmental policy issues and serious shortcomings in the quality, geographical coverage, or timeliness of the available data.

For example, to measure environmental health policy outcomes, we would ideally like to use indicators measuring the exposure-effect relationships of major environmental toxins such as lead and mercury. Many important environmental health indicators are, however, available only for very few countries or at limited sub-national or regional levels. A major initiative in this context is a project under the guidance of the World Health Organization to estimate the Global Burden of Disease, including environmental diseases.¹⁷ Due to serious data gaps and methodological issues, these estimates

are published for the WHO's regional areas only. Hopefully, continued efforts will make it possible to report country-level data in the future.

The need to incorporate economic policy decisions into environmental performance measurement is exemplified through the issue of governmental subsidies. Perverse subsidies in agriculture, fisheries, and energy sectors have been shown to have negative impacts on resource use and management practices. But data on the amount of subsidies and especially on their impacts are extremely difficult to obtain. The EPI contains an improved agricultural subsidy measure that builds on the variable used in the 2005 Environmental Sustainability Index (Esty, Levy et al., 2005).

Biodiversity and habitat protection have recently received greater attention with a focus on developing new and better indicators. Wetland protection, for example, is an important aspect of biodiversity protection. Yet, it is not routinely measured on a grand scale. The issue of land degradation, which affects many countries worldwide, is so complex that scientists and experts at the Food and Agricultural Organization of the United Nations have not yet been able to harmonize existing methodologies to the extent necessary to obtain routine, high-quality global assessments of the extent and severity of anthropogenic land degradation.¹⁸

Another noteworthy issue affecting national performance measurement is that not every indicator is equally applicable or relevant for each country. For example, the EPI includes a measure of timber harvesting. Not every country has forests, however, making this indicator less valuable to these countries. The index does not consider timber harvesting for countries without

¹⁷ WHO Burden of Disease Project. More information is available at <http://www.who.int/healthinfo/bodproject/en/index.html>

¹⁸ We considered, for example, inclusion of the GLASOD land degradation assessment but refrained because the data are outdated and not comparable enough to permit cross-country performance assessments.

natural or planted forests. Equally relevant in this context is consideration of how environmental pollution and resource use affect countries at different stages of economic development.

The cluster analysis and presentation of EPI results for various “country peer groups” highlights that different EPI indicators are of high importance to various country groupings. While this is an important issue for weighting the indicators, it also demonstrates that indicator selection for a global index is a difficult task.

While our search for additional and better data is ongoing, this Pilot EPI contains 16 indicators for 133 countries, which we believe reflect the most important and best available measures to track and assess environmental performance. Aside from policy relevance, only datasets with sufficient coverage, data “freshness”, and methodological consistency were chosen.

F.5. Cluster Analysis

Cluster analysis refers to a rich suite of statistical classification methods used to determine similarities (or dissimilarities) of objects in large datasets. We use this technique to identify groupings of relevant peer countries. Within each peer group, countries have a better basis for benchmarking their environmental performance because the group members are similar with respect to the data used to classify them and the differences across the groups are maximized.

Cluster analysis helps to advance this process by grouping beyond the level of development alone. In doing so, it enables countries to identify others who are similarly situated – thus providing a good starting point in the search for best practices. In this context, the question of interest in carrying out a cluster analysis of the EPI is whether there are similarities among countries in their environmental performance at

the aggregate EPI level and with respect to the EPI indicators and policy categories.

Cluster Analysis Techniques

There is no best method for cluster analysis and the results of cluster analyses are subject to interpretation. Therefore, we applied two different algorithms. Specifically, we explored the data structure using a non-parametric, distance-based agglomerative clustering algorithm known as Ward’s method.

A feature of agglomerative clustering is that it starts with as many individual clusters as there are countries. It then successively combines countries that are most similar to each other with respect to a quantitative similarity measure until all countries are joined in a single cluster.

The similarity measure decreases during this process, while the within-cluster dissimilarity increases as more and more countries are added. The trade-off lies therefore in choosing a similarity measure, or “pruning value,” that yields both a relatively small number of clusters and a high level of similarity. We determine that six clusters yield a reasonable division between the countries.

After determining the number of country clusters, we use the k means clustering method developed by Hartigan and Wong (Hartigan and Wong, 1979) to determine cluster membership. K means is a non-hierarchical method that requires that the number of clusters, k , be specified upfront (hence the preliminary use of Ward’s method) and then iteratively finds the disjoint partition of the objects into k homogeneous groups such that the sum of squares within the clusters is minimized.

The algorithm converges in fewer than 10 iterations for the 16 proximity-to-target indicators.

The differences between the six country groupings at the indicator level can also be illustrated by a plot of the respective cluster centers (see Figure F1).

Indicators that are particularly influential in determining the differences between the groups have large deviations in the cluster centers.

These indicators are:

- Regional Ozone (OZONE),
- Indoor Air Pollution (INDOOR),
- Water Consumption (OVRSUB),
- Energy Efficiency (ENEFF),
- CO2 Emissions per GDP (CO2GDP),
- Drinking Water (WATSUP),
- Adequate Sanitation (ACSAT), and
- Ecoregion Protection (PACOV).

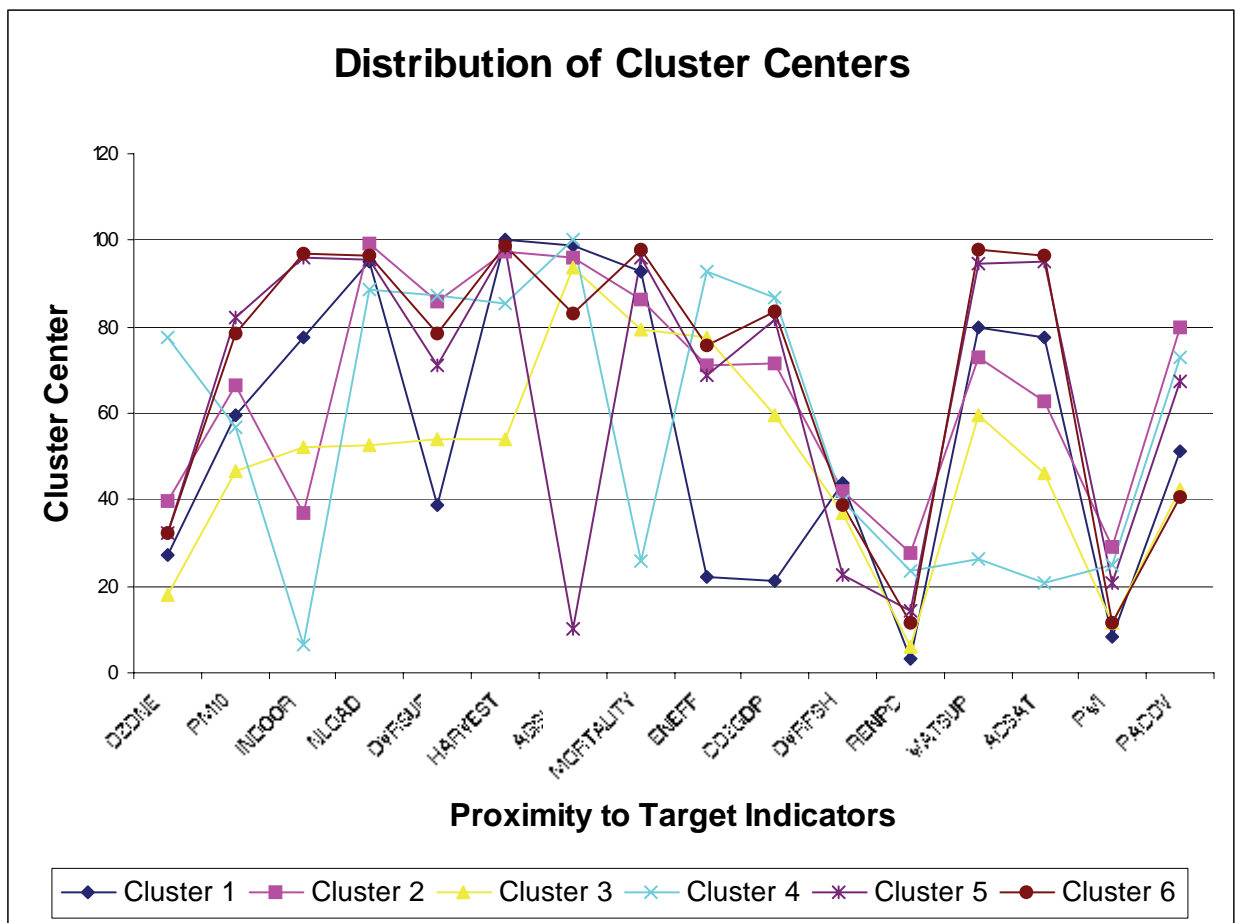


Figure F1: Distribution of Cluster Centers for the Six Country Peer Groups and Proximity-to-Target Indicators

