

Kraemer, Guido; Reichstein, Markus; Camps-Valls, Gustau; Smits, Jeroen; Mahecha, Miguel D.

Article — Published Version

The Low Dimensionality of Development

Social Indicators Research

Provided in Cooperation with:

Springer Nature

Suggested Citation: Kraemer, Guido; Reichstein, Markus; Camps-Valls, Gustau; Smits, Jeroen; Mahecha, Miguel D. (2020) : The Low Dimensionality of Development, Social Indicators Research, ISSN 1573-0921, Springer Netherlands, Dordrecht, Vol. 150, Iss. 3, pp. 999-1020, <https://doi.org/10.1007/s11205-020-02349-0>

This Version is available at:

<https://hdl.handle.net/10419/288652>

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.


If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.



<https://creativecommons.org/licenses/by/4.0/>



The Low Dimensionality of Development

Guido Kraemer^{1,2,3,4}  · Markus Reichstein¹ · Gustau Camps-Valls³ · Jeroen Smits⁵ · Miguel D. Mahecha^{1,2,4}

Accepted: 20 April 2020 / Published online: 20 May 2020
© The Author(s) 2020

Abstract

The World Bank routinely publishes over 1500 “World Development Indicators” to track the socioeconomic development at the country level. A range of indices has been proposed to interpret this information. For instance, the “Human Development Index” was designed to specifically capture development in terms of life expectancy, education, and standard of living. However, the general question which independent dimensions are essential to capture all aspects of development still remains open. Using a nonlinear dimensionality reduction approach we aim to extract the core dimensions of development in a highly efficient way. We find that more than 90% of variance in the WDIs can be represented by solely five uncorrelated dimensions. The first dimension, explaining 74% of variance, represents the state of education, health, income, infrastructure, trade, population, and pollution. Although this dimension resembles the HDI, it explains much more variance. The second dimension (explaining 10% of variance) differentiates countries by gender ratios, labor market, and energy production patterns. Here, we differentiate societal structures when comparing e.g. countries from the Middle-East to the Post-Soviet area. Our analysis confirms that most countries show rather consistent temporal trends towards wealthier and aging societies. We can also find deviations from the long-term trajectories during warfare, environmental disasters, or fundamental political changes. The data-driven nature of the extracted dimensions complements classical indicator approaches, allowing a broader exploration of global development space. The extracted independent dimensions represent different aspects of development that need to be considered when proposing new metric indices.

Keywords Sustainable development · Sustainability indicators · Dimensionality reduction · PCA · Isomap

Electronic supplementary material The online version of this article (<https://doi.org/10.1007/s11205-020-02349-0>) contains supplementary material, which is available to authorized users.

✉ Guido Kraemer
guido.kraemer@uni-leipzig.de

Extended author information available on the last page of the article

1 Introduction

During the last decades, humanity has achieved on average longer life spans, decreased child mortality, better access to health care and economic growth (UNDP 2019). In emerging countries like China and India many people have escaped extreme poverty (less than 1.90 US\$ per person per day) in the wake of persistent economic growth (UNDP 2016). To measure development, a wide range of variables are routinely made available by the World Bank, describing multiple facets of societal conditions. These “World Development Indicators” (WDIs, revision of 3/5/2018; The World Bank 2018b) have become a key data resource that today contains more than 1500 variables with annual values for most countries of the world.

A widely accepted method for assessing development consists in the construction of indicators (hereafter called “classical” indicators) based on expert knowledge that allow ranking countries by their development status and tracking them over time. A multitude of such classical indicators have been developed over the past few decades (Parris and Kates 2003; Shaker 2018; Ghislandi et al. 2018), focusing on different aspects of development. For instance, the United Nations Development Programme (UNDP) uses a Multidimensional Poverty Index, a Gender Development Index, a Gender Inequality Index, amongst others for reporting on human development (UNDP 2019). The UNDP’s most prominent indicator is the Human Development Index (HDI), which is the geometric mean of indicators describing life expectancy, education, and income (UNDP 2019). However, there are many other efforts to produce relevant indicators, such as the Genuine Progress Index (Kubiszewski et al. 2013), the Global Footprint and Biocapacity indicators (McRae et al. 2016), and the POLITY scores (Marshall and Elzinga-Marshall 2017), to name just a few. These classical approaches are well suited for describing and communicating selected aspects of development, e.g. the HDI has been specifically developed “to shift the focus of development economics from national income accounting to people-centered policies” (UNDP 2018).

An alternative to approach to constructing indices consists in using purely data-driven methods, such as “Principal Component Analysis” (PCA; Pearson 1901) or “Factor Analysis” (FA; Spearman 1904). PCA linearly compresses a set of variables of interest. The resulting principal component or components represent the main dimensions of variability and can then be interpreted as an emerging indicator (OEDC 2008). This approach has been used to create indicators of well-being from sets of co-varying variables (Mazziotta and Pareto 2019). While PCA refers to a well defined method which tries to summarize the variance of an entire dataset, FA refers to a family of methods which assumes a multivariate linear model to explain the influences of a number of latent factors on observed variables. PCA and FA have been used extensively in the social sciences, e.g. to create indicators of well-being (Stanojević and Benčina 2019) or to construct wealth indices (Filmer and Scott 2012; Smits and Steendijk 2015). An advantage of such data-driven methods is that they follow well defined mathematical behaviors and are not subjective, while there is no well established method for the creation of classical indicators (Shaker 2018). A disadvantage of these methods is that they do not consider the polarity of the variables nor allow for expert based weighting (Mazziotta and Pareto 2019). A detailed comparison between classical indicators and data driven indicators can be found in SI Table 1.

The rationale for dimensionality reduction methods like PCA is that often the intrinsic dimension of a dataset is much lower than the number of variables describing it. In climate science, for example, a set of co-varying variables observed over a region in the equatorial

pacific can be compressed into the Multivariate ENSO Index (MEI, Wolter and Timlin 1993, 2011), to describe the state of the El Niño Southern Oscillation (ENSO)—the principal climate mode that determines e.g. food security in many regions of the world. In image vision, the number of main features from a set of images is much less than the number of pixels per image. For example Tenenbaum et al. (2000) shows that pictures taken from the same object at different angles have the viewing angle as the main features of the set of images. These main features are called “intrinsic dimensions” because they are sufficient to describe the essential nature of the entire dataset, the number of such intrinsic dimensions is called the “intrinsic dimensionality” of the dataset (Bennett 1969).

Development is a complex concept though, which is reflected in the large number of variables included in the WDI database. The large number of indicators let us expect substantial redundant information (Shaker 2018; Rickels et al. 2016). This issue has also been discussed in the context of the Sustainable Development Goals (SDGs; The World Bank 2018a). Since their introduction by the United Nations in 2015, the SDGs have become a widely accepted framework to guide policymakers. Today 17 SDGs address the issues of poverty, hunger, health, education, climate change, gender inequality, water, sanitation, energy, urbanization, environment and social justice. To monitor the SDGs, 169 specific targets have been developed which are measured using 232 different indicators included in the WDIs (The World Bank 2018a; United Nations General Assembly 2017a), leading to substantial interactions across and within the targets that need to be analyzed (Costanza et al. 2016). Hence, the question emerges how to extract the key information jointly contained in the WDIs that leads to a succinct, objective, and tangible picture of development.

In this paper, we aim to elucidate the most important dimensions of development contained in the WDI dataset, using a data-driven approach. Specifically, we aim to answer the question, how many independent indicators are necessary to summarize development space and what is their interpretation. We exploit the potential of nonlinear dimensionality reduction to identify dimensions that represent these (typically mutually dependent) variables, while preserving relevant properties of the underlying data. The rationale is that we expect strong interactions between the different WDIs which may not be linear.

Understanding what intrinsic dimensionality our current indicators of development have, could have important implications for policy makers. If the intrinsic dimensionality of development proves to be high, one would indeed need to track many indicators synchronously to understand the interplay of different aspects of development. On the contrary, in the case of a low-dimensional development space, it would be sufficient to track either the emerging dimensions, or the closely related variables to monitor development across countries and time. In fact there is already substantial evidence that supports our hypothesis of a low-dimensional development space. For instance Pradhan et al. (2017) found strong correlations between all SDGs, suggesting that the intrinsic dimensionality of the SDGs is relatively low, but this has not been quantified yet.

This article is divided into five sections. Section 2 presents a data-driven approach to extract nonlinear components from the WDI database, Sect. 3 presents the resulting dimensions, their interpretations, global distributions, trends and trajectories. Section 4 discusses the relation of the indicators produced by the method presented here with previous indicator approaches, and finally Sect. 5 gives some concluding remarks.

2 Data and Methods

2.1 Data

To understand the structure and dimensionality of development we rely on the WDI dataset, which is the primary World Bank collection of development indicators, compiled from officially-recognized international sources. The WDIs comprise a total of 1549 variables with yearly data between 1960 and 2016 for 217 countries. As such, it represents the most current and accurate global development database available (The World Bank 2018b).

Even though the WDI dataset is the most comprehensive set of development indicators available, it contains many missing values. Only for the most developed countries the dataset is (nearly) complete. For many other countries—particularly low and middle income countries—many indicators are partly or completely missing. This is problematic, as for most dimension reduction methods a dataset without missing observations is required. To make our analyses possible, we therefore had to select a subset of indicators, countries and years with few missing observations and to fill in the remaining missing observations using gapfilling techniques (see next section). To avoid arbitrariness of the subset selection, a scoring approach was used (see Sect. 2.2) and the 1000 subsets with the highest scores were selected. These 1000 subsets contained a total of 621 variables, 182 countries and the years ranging from 1990 to 2016. The subsets cover almost all categories of variables. The categories with their respective number of variables in the entire WDI dataset and the subsets are “Economic Policy & Debt” (120 out of 518), “Education” (73 out of 151), “Environment” (74 out of 138), “Financial Sector” (29 out of 54), “Gender” (1 out of 21), “Health” (123 out of 226), “Infrastructure” (19 out of 41), “Poverty” (0 out of 24), “Private Sector & Trade” (103 out of 168), “Public Sector” (31 out of 83), and “Social Protection & Labor” (48 out of 161). Jointly these subsets are representative for the original dataset while avoiding large gaps.

2.2 Gapfilling

The dimensionality reduction approach we have chosen (see Sect. 2.3) relies on a full matrix of distances between the different country–year data points. However, given the large amount of data gaps this global distance matrix cannot be computed directly. In the following, we develop an approach to find subsets of the WDI database which we can gapfill and use for estimating distances among data points.

In order to choose subsets of the WDI database covering a wide range of WDIs, countries, and years, but also having as few missing values as possible, the following method was applied: A series of subset was created from the full WDI dataset using a combination of thresholds for the maximum fraction of missing values for the WDIs, f_v , and countries, f_c , as well as a starting year, y_{start} , and an ending year, y_{end} . We assigned a score to each of the resulting subsets by using a grid search over the parameters, $f_v, f_c \in (0.05, 0.15, \dots, 0.65)$ and $y_{\text{start}}, y_{\text{end}} \in (1960, 1961, \dots, 2017), y_{\text{start}} < y_{\text{end}}$. The size of this parameter space is 80997, each with a different combination of missing value thresholds and starting and ending year combinations. The 1000 subsets with the highest scores were finally chosen to build the global distance matrix. For an overview of the entire method, see Fig. 1.

Each subset was created from the full WDI dataset by choosing consecutive years with starting year, y_{start} , and ending year, y_{end} ; WDIs with a higher missing value fraction, p_v , than the corresponding threshold were dropped ($p_v > f_v$). Then, countries with

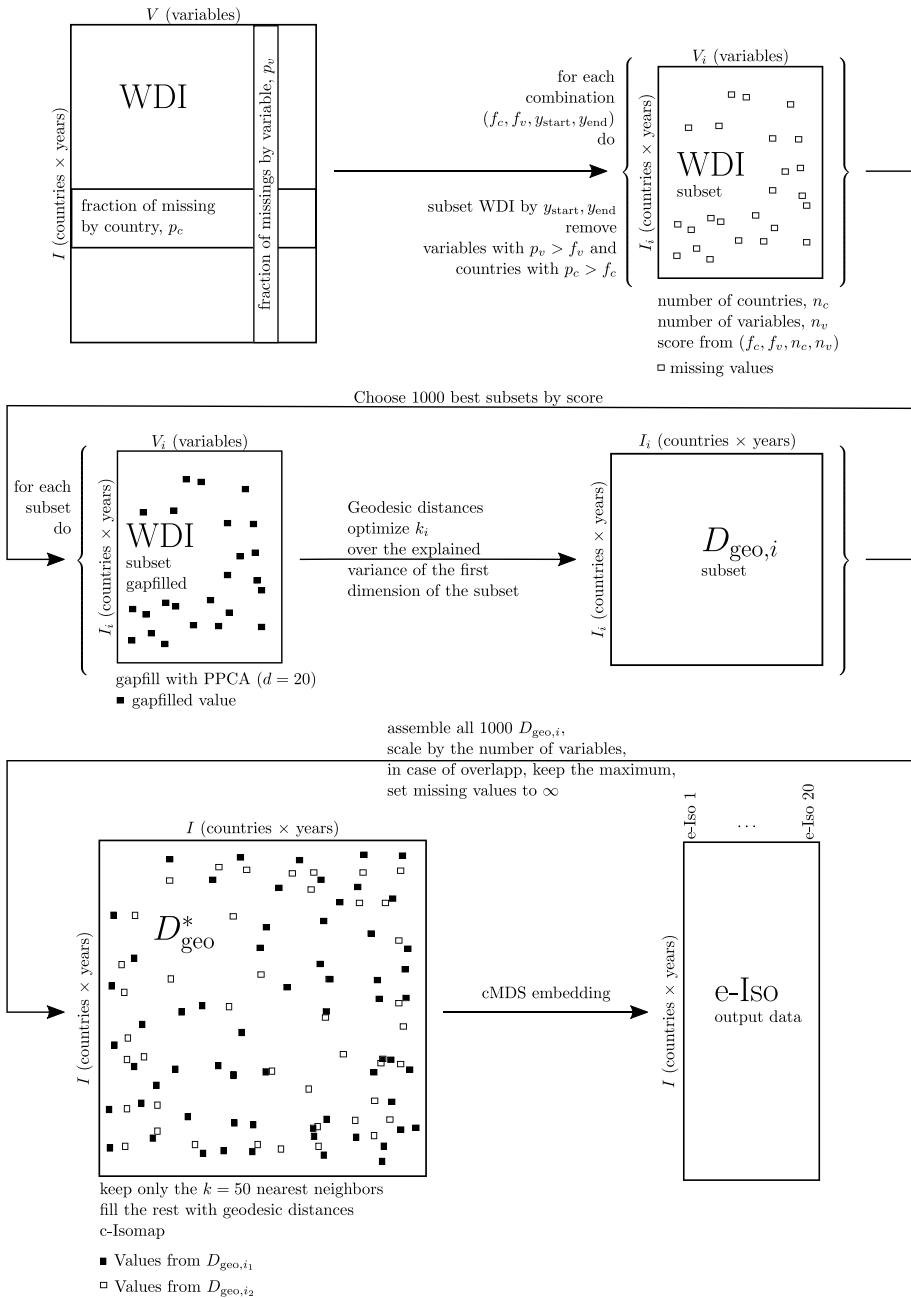


Fig. 1 Schematic presentation of the ensemble Isomap (e-Iso) algorithm, for details see text

higher missing value fractions, p_c , than the corresponding threshold were dropped as well ($p_c > f_c$). The number of remaining countries, n_c , and WDIs, n_v , was recorded and the resulting subsets were filtered to retain more observations (the number of countries times

the number of years) than variables, leaving a total of 77,610 subsets of the WDI for score calculation.

To account for different scales of the parameters, the values had to be rescaled, i.e. we calculated n'_v from n_v by scaling the values from subsets linearly to a minimum of 0 and a maximum of 1, analogously for n'_c , f'_c , and f'_v . The final score was then calculated as

$$\text{score} = \sqrt{n'_v n'_c} - \sqrt{f'_c f'_v}.$$

This score calculates the geometric means of the variables of interest. The geometric mean has the advantage over the arithmetic mean that it is very sensitive to single bad values. As we want to maximize the number of countries and WDIs chosen and have as few missing values possible, the final score is the difference between the geometric means. For further processing, the subsetted WDI data matrices with the 1000 highest scores were selected.

Finally the subsetted WDI data matrices with the 1000 highest scores were selected and a gapfilling procedure using Probabilistic PCA (Stacklies et al. 2007) was performed on the centered and standardized (z -transformed) variables using the leading 20 dimensions.

2.3 Dimensionality Reduction

Dimensionality reduction describes a family of multivariate methods that find alternative representations of data by constructing linear or, in our case, nonlinear combinations of the original variables so that important properties are maintained in as few dimensions as possible. A plethora of algorithms is currently available for dimensionality reduction, both linear and nonlinear (Arenas-Garcia et al. 2013; Van Der Maaten et al. 2009; Kraemer et al. 2018), but PCA is dominating in applied sciences because of ease of use and interpretation.

One method to find an embedding from a known distance matrix is “classical Multi-dimensional Scaling” (CMDS; Torgerson 1952), this method is equivalent to PCA if the distance matrix is computed from the observations using Euclidean distance. CMDS finds coordinates in a reduced Euclidean space of dimension i minimizing

$$\|\tau(D) - \tau(D_i)\|_2,$$

where D is the matrix of Euclidean distances of observations and D_i the matrix of Euclidean distances of the embedded points. $\tau(D) = -\frac{1}{2}HSH$, is the “double centering operator”, with $S = [D^2_{ij}]$, $H = [\delta_{ij} - \frac{1}{n}]$, and $\|X\|_2 = \sqrt{\sum_{ij} X^2_{ij}}$ the L_2 -norm. CMDS and therefore PCA tend to maintain the large scale gradients of the data and cannot cope with nonlinear relations between the covariates.

“Isometric Feature Mapping” (Isomap; Tenenbaum et al. 2000) extends CMDS, but instead of Euclidean distances, it respects geodesic distances, i.e. the distances measured along a manifold of possibly lower dimensionality,

$$\|\tau(D_{\text{geo}}) - \tau(D_i)\|_2.$$

Specifically, Isomap uses geodesic distances, $D_{\text{geo}} = [d_{\text{geo}}(x_i, x_j)]$, which are the distances between two points following a k -nearest neighbor graph of points sampled from the manifold.

Isomap is guaranteed to recover the structure of nonlinear manifolds whose intrinsic geometry is that of a convex region of Euclidean space (Tenenbaum et al. 2000). Isomap

unfolds curved manifold which makes the method more efficient than PCA in reducing the number of necessary dimensions in the presence of nonlinearities.

To construct the geodesic distances, a graph is created by connecting each point to its k nearest neighbors and distances are measured along this graph. If the data samples the manifold well enough, then the distances along the graph will approximate the geodesic distances along the manifold. The value of k will determine the quality of the embedding and has to be tuned.

We applied Isomap on the 1000 previously generated subsets of the WDI database. To find the optimum value k for each subset, k_i , Isomap was calculated first with $k_i = 5$ and the residual variance for the embedding of the first component was calculated (see below). This process was repeated increasing the values of k_i by 5 in each step until there was no decrease in the residual variance for the first component any more (Mahecha et al. 2007). In order to get an intuition of Isomap, we recommend the original publication of the Isomap method (Tenenbaum et al. 2000) which contains an excellent didactic explanation of the method.

2.4 Ensemble PCA and Ensemble Isometric Feature Mapping

An observation consists of a country name and year. To calculate a linear embedding (ensemble PCA) over the union of all countries, years and variables chosen before, we used a Probabilistic PCA ($d = 80$, where d is the number of dimensions used in the probabilistic PCA) to gapfill all the observations and variables occurring in the subsets of the WDI dataset and applied a normal PCA to the gapfilled dataset.

We developed “Ensemble Isometric Feature Mapping” (e-Isomap) to produce the final nonlinear embedding based on the different gapfilled subsets of data. E-Isomap combines $m = 1000$ geodesic distance matrices created from the subsets of the previous step and constructs a global ensemble geodesic distance matrix, D_{geo}^* , from the geodesic distance matrices of the m Isomaps.

Let the total set of observations be $I = \{1, \dots, n\}$ (a country–year combination) and the observed variables $V = \{1, \dots, p\}$ (the WDIs). We first perform one Isomap $i \in \{1, \dots, m\}$ per subset of I and V , I_i and V_i respectively, where $|V_i|$ is the number of variables for Isomap i . The geodesic distance matrix for Isomap i is $D_{geo,i} = (d_{geo,i}(x_j, x_k))_{j,k}$ with $j, k \in I_i$. If a pair of observations (x_j, x_k) does not occur in Isomap i , it is treated as a missing value. First the geodesic distance matrices are scaled element-wise to account for the different number of variables used,

$$d'_{geo,i}(x_j, x_k) = d_{geo,i}(x_j, x_k) \sqrt{\frac{|V|}{|V_i|}},$$

which are then combined into a single geodesic distance matrix D_{geo}^* by using the maximum distance value,

$$d_{geo}^*(x_j, x_k) = \max_i d'_{geo,i}(x_j, x_k).$$

Missing values are ignored if all values are missing for a pair (x_j, x_k) and they are treated as infinite distances. Taking the maximum avoids short-circuiting distances and as long as there are few missing values. This provides an accurate approximation of the internal distances.

Finally the k nearest neighbor graph G is constructed from the distance matrix, and each edge $\{x_i, x_j\}$ is weighted by $\frac{|x_i - x_j|}{\sqrt{M(i)M(j)}}$, where $M(i)$ is the mean distance of x_i to its k nearest neighbors. This last step is called c -Isomap (Silva and Tenenbaum 2003) and it contracts sparsely sampled regions of the manifold and expands densely sampled regions, the c -Isomap step proved to give a more evenly distributed embedding. Finally the geodesic distances are calculated on G and classical scaling is performed to find the final embeddings.

2.5 Quality Measurement of an Embedding and Influence of Variables

The quality for the embedding is estimated by calculating the residual variance (Tenenbaum et al. 2000) computed as

$$\text{residual variance}_i = 1 - r^2(\hat{D}, D_i) = 1 - \text{explained variance}_i,$$

where D_i is the matrix of Euclidean distances of the first i embedded components and \hat{D} is the matrix of Euclidean distances for PCA and the matrix of geodesic distances for Isomap in original space. Note that because D_i and \hat{D} are symmetric, we only use one triangle for the calculation of the residual variance. This notion of explained variance is different from the one usually used for PCA, which is derived from the eigenvalue spectrum, but the measure used here has the advantage that it gives comparable results for arbitrary data such as the HDI and Isomap.

To assess the influence of single variables on the final e-Isomap dimensions, we calculated the distance correlation (dcor, Székely et al. 2007), which is a measure of dependence between variables that takes nonlinearities into account. Due to the strong nonlinearities in the dataset and the embedding method, a simple linear correlation would not have provided sufficient information about the relationships between variables and the embedding dimensions.

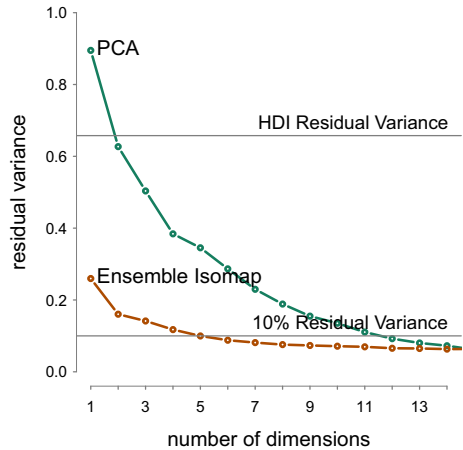
3 Results

3.1 Required Number of Dimensions

Our results suggest that the “development space” described by the WDI data is of low intrinsic dimensionality. Using e-Isomap we needed five dimensions only to explain 90% of the variance of global development (see Fig. 2). The first dimension alone explains 74% of the variance in the WDI data; Dimension 2 explains 9.9% of the variance and dimensions 3–5 explain less than 3% of the variance each. Although the explained variance of dimensions 2–5 seems small compared to that of the first dimension, each of these dimensions still represents a distinct, well defined and highly significant aspect of development, as we will show later. Therefore the raw variances should not be used as the sole measure to discard dimensions.

The finding that such a high compression can be achieved with e-Isomap indicates that the WDIs are highly interdependent and that the underlying processes are highly nonlinear (see Fig. 2). This is also confirmed by an analogous analysis using linear PCA which cannot compress the data with the same efficiency: the first PCA dimension only explains 10% of the variance, and 12 dimensions are required to express more than 90% of the variance. The cumulative explained variances for the first five e-Isomap dimensions are 74%, 84%,

Fig. 2 The residual variance for the first 14 components. The circled lines represent the residual variance of the Ensemble Isomap and the PCA. Isomap is much more efficient in compressing dimensionality of the data requiring only 5 components to describe more than 90% of the variance, while PCA requires 12 components to describe 90% of variance. The upper grey horizontal line represents the residual variance for the HDI (66%) and the lower one the 10% residual variance boundary



86%, 88%, and 90%, which is much more than the respective PCA dimensions (10%, 37%, 50%, 61%, and 65%).

To understand if the HDI can compress the data in the same way, we compute the variance of the HDI in the same way. We find that the HDI captures 34% of the variance (see Fig. 2), which is less than half of the variance captured by the first dimension extracted via nonlinear dimensionality reduction but more than three times the variance explained by the first PCA dimension. If the target is reducing the WDI data to a single dimension, the best performing method is e-Isomap, followed by the HDI, while PCA does not perform this task very well. In other words, the first e-Isomap dimension seems to be a more powerful summary of the WDI data than the HDI.

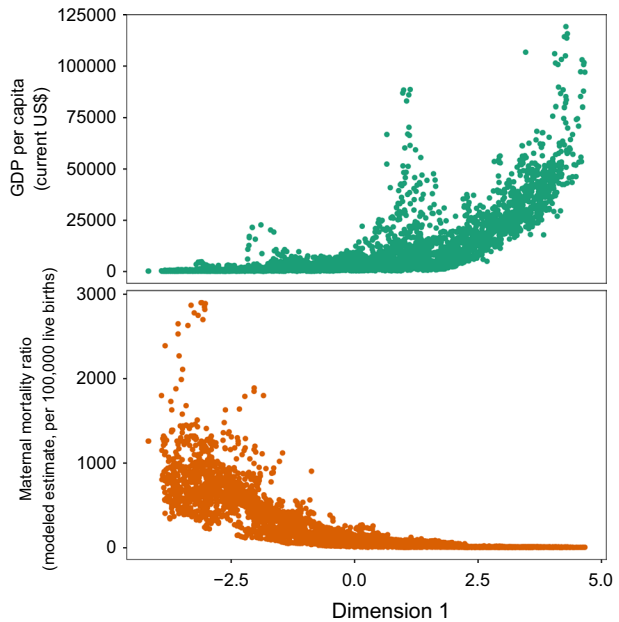
3.2 Intrinsic Dimensions of Development

Our results suggest that the dimensions resulting from the e-Isomap can be indeed interpreted analogously to traditional indicators of development. The main difference from classical indicators is that these dimensions emerge directly from the data. Hence, the interpretation of these indicators has to be achieved *a posteriori*. We also find that the relationship between the WDIs and the dimensions is highly nonlinear (see Fig. 3) requiring the use of nonlinear measurements of correlation. Here we relate the extracted dimensions to the original data using distance correlation. See Fig. 4, for a complete and interactive table in the supporting information.¹

We find that dimension 1 essentially represents progress in education, life expectancy, health, and relates to the population pyramid (see Fig. 4). Additionally, dimension 1 is associated with infrastructure and income-related indicators. Other indicators that strongly correlate with this dimension are related to pollution and primary production and include tariffs and imports as well as trade, the climate impact of GDP (gross domestic product), and development aid received. Because dimension 1 embraces education, health, and life expectancy, it is conceptually similar to the HDI. In fact, dimension 1 has a strong nonlinear correlation with the HDI ($d_{\text{cor}} = 0.93$), and can be

¹ http://bgc-jena.mpg.de/~gkraemer/consolidated_cor_table.

Fig. 3 Illustrating the nonlinear relation between dimension 1 and GDP per capita and maternal mortality rates. Top: There is a positive correlation between GDP per capita and dimension 1. On the positive end of dimension 1 the per capita income increases strongly, while it increases very slowly on the negative side of dimension 1. Bottom: There is a negative correlation between the maternal mortality rate and dimension 1. The maternal mortality rate decreases strongly on the negative end but does not decrease any more on the positive end



interpreted as a measure of development *sensu* HDI, even though it includes much more than the aspects measured by the HDI. We also find that the correlation is much lower for most sub-Saharan countries (Fig. S2).

Dimension 2 (9.9% of the variance) is strongly related to gender ratios in the general population and the labor market, as well as primary energy production and consumption and the fraction of 25–29 year old people. This dimension spans a gradient between the extremes of dimension 1 and former Soviet allied countries on one end, and rich mostly oil exporting nations on the other end (see Fig. 5). On the positive extreme on this axis are countries that have a very high participation of women in the labor market (e.g. Mozambique has the highest participation of women in the labor force with around 55%, similar to countries like Lithuania with a rate of approx 50%) on the negative extreme we can find countries with a very low participation of women in the labor market: Rich countries like the United Arabian Emirates have a female labor force of around 12%, just as poorer countries like Yemen that has a participation rate of women of around 8%, and low death rates. Crude death rates also correlate well with this dimension and do not separate regions, e.g. Latvia in 1994 had a crude death rate of 16.6/1000 people, Denmark in 1993 a crude death rate of 12.1 per 1000 people, while similar crude death rates can be found in undeveloped countries (Democratic Republic of the Congo, 1996, 16.655 death per 1000 people; or Liberia, 2005, 12.128 deaths per 1000 people), on the low extreme we find mostly rich oil exporting nations (e.g. Qatar and the United Arabian Emirates with values around 1.5 deaths per 1000 people).

The third to fifth dimensions explain much less variance but are still important in that they account for variables not found in the first two dimensions: Dimension 3 (1.9% of the variance) is a labor market gradient representing descriptors like ratios of labor force, employment, and unemployment. Dimension 4 (2.4% of the variance) summarizes homicide rates, methane emissions and food exports. Dimension 5 (1.8% of the

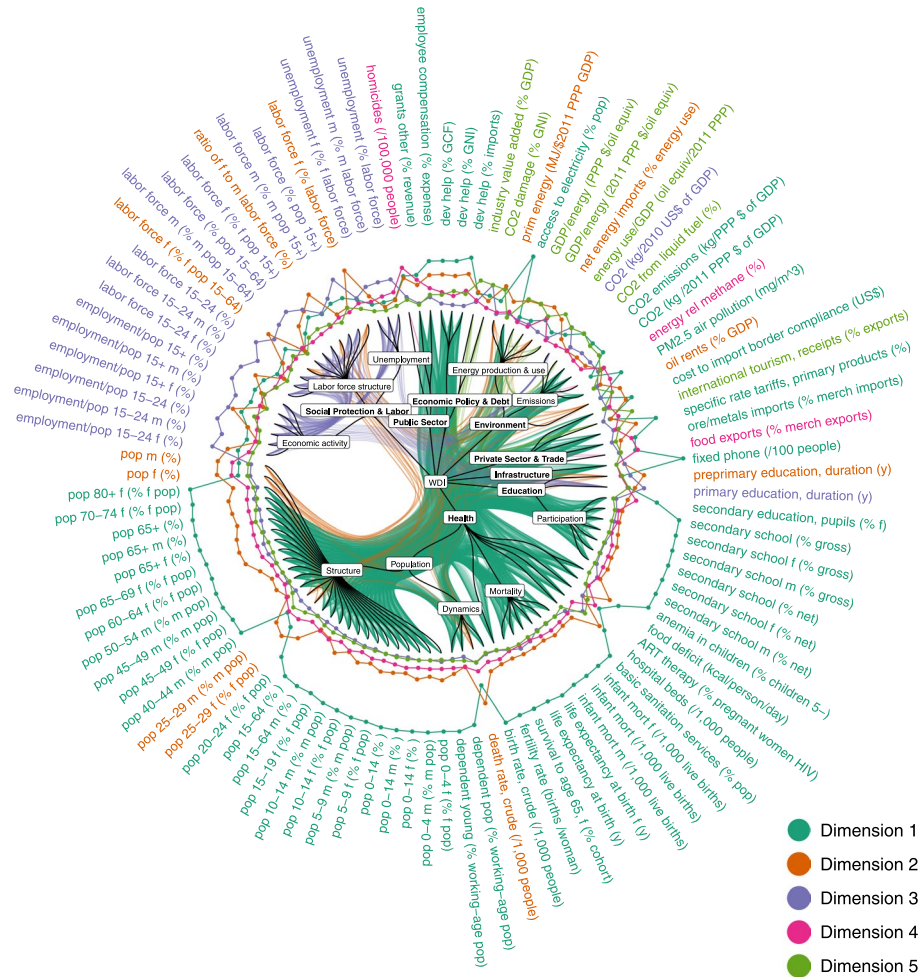


Fig. 4 The importance of single WDIs to the dimensions. Outer circle: indicator names, colored by the maximum distance correlation dimension. Middle circle: distance correlation values with dimensions 1–5. Black center dendrogram: thematic ordering of the indicators by the World Bank. Colored lines in the center circle: connect all indicators that have maximum distance correlation with the same dimension, bundled along the center dendrogram for easier visualization. The figure only shows the WDIs that have the highest distance correlations with dimensions 1–5

variance) represents the CO₂ impact of GDP, tourism and value added to products by industry.

3.3 Global Trends

Development is dynamic. Over time each country moves along a characteristic trajectory in development space. Along the first dimension, clear trends can be observed. Most countries have a positive slope (see Fig. 6). Given that dimension 1 essentially spans a gradient between wealthy and poor countries, this reveals the overall global

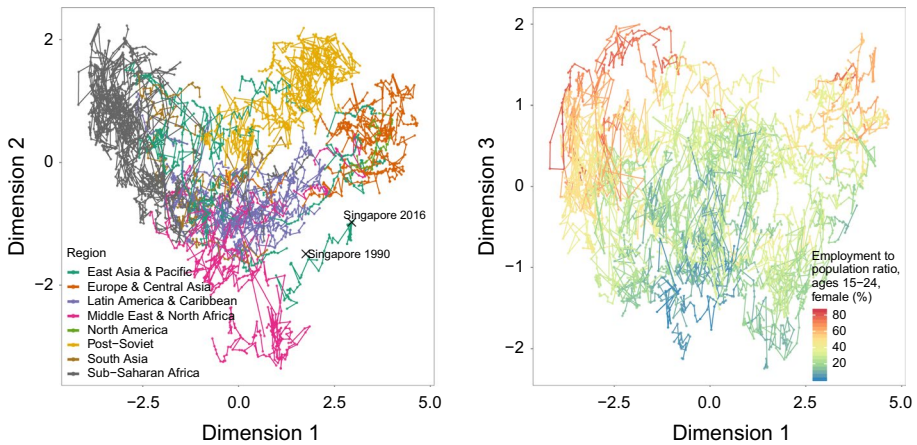


Fig. 5 E-Isomap dimensions 1–3. Left: Dimension 1 and 2, colored by World Bank regions and former East Block and allies in Eastern Europe. Right: Dimensions 1 and 3 colored by Employment to population ratio, ages 15–24, female (%). Dimension 1 (the horizontal axis on both panes) is a general wealth gradient, on the far left side are poor countries, mostly classified as “Sub-Saharan Africa” while on the right side are the developed countries with most Western European countries on the far right. Dimension 2 (vertical axis on the left pane) spans mostly the percentage of female population and labor force participation of women. Dimension 3 (vertical axis on the right pane) spans employment ratios, employment ratios for women and labor force participation of young working age women. There is an interactive online version available (http://bgc-jena.mpg.de/~gkraemer/consolidated_dimred/)

trend towards a wealthier world (Gapminder Foundation 2018). Only a few countries have negative slopes. Comparing the slopes of “Sub-Saharan Africa” with the rest of the world reveals a widening gap in the development gradient *sensu* HDI. Dimensions 2–4 do not show such pronounced overall trends.

Dimension 2 shows positive trends in most of the “Western World” and North Africa and negative trends in most parts of Asia and Sub-Saharan Africa. The positive trends in the “Western World” countries are due to an increased participation of women in the labor market, declining death rates in countries with young populations, and climbing death rates in countries with aging societies. Many developing countries in Sub-Saharan Africa and Asia show negative trends, which seems to be a common interaction between dimensions 1 and 2 on the far negative end of dimension 1.

Dimension 3 shows mostly employment/unemployment ratios, but there are no really strong general trends observable. We note that eastern and western Europe show fundamentally different trends, most of eastern Europe has predominantly negative trends, while in the rest of Europe there are few significant slopes reflecting the increase in unemployment in Eastern Europe. Other notable countries include Peru, Ethiopia, and Azerbaijan, where unemployment rates have strongly decreased; these countries show strong positive trends.

Dimension 4 shows energy-related methane emissions, which have increased in most parts of the northern hemisphere and decreased in most other parts of the world, as well as homicide rates, which have decreased in large parts of the world, but increased in parts of Latin America. The data on homicide rates in large parts of Africa are very sparse.

Dimension 5 shows tourism and the ecological impact of GDP. In general, more GDP is produced per unit of energy. This trend seems to be stronger in the Western World.

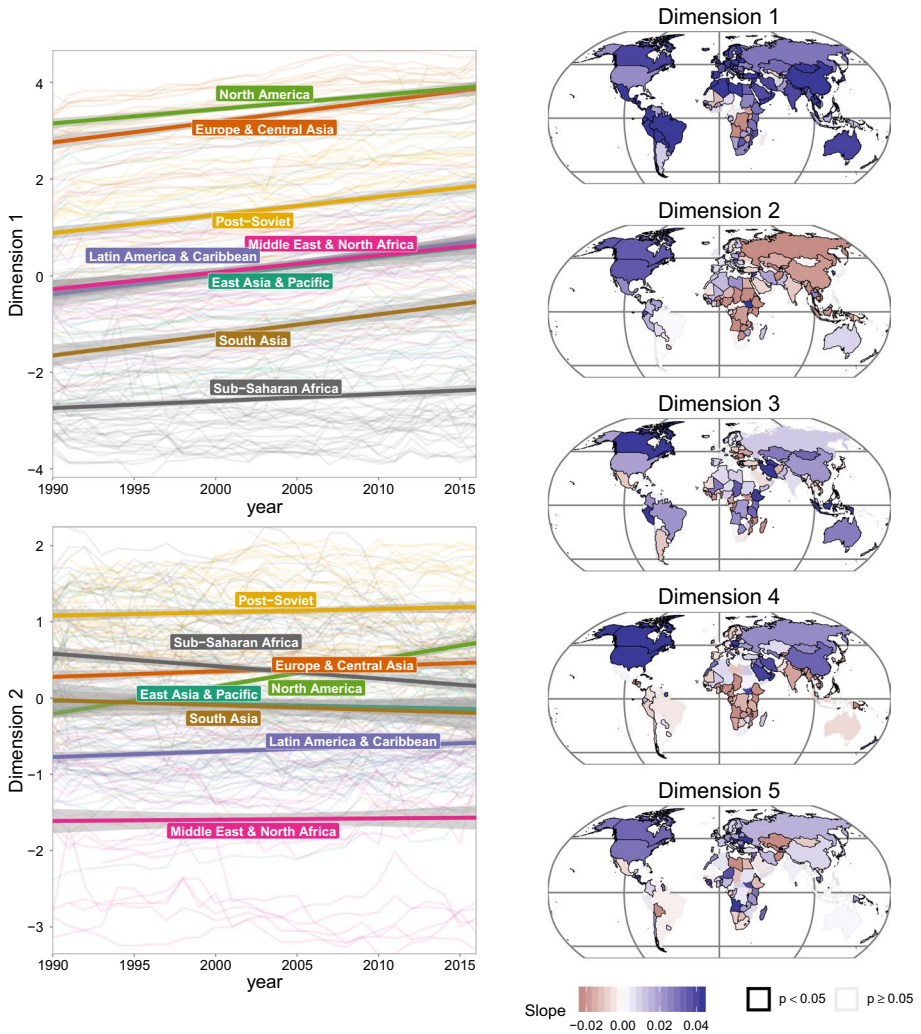


Fig. 6 Trends in the first e-Isomap axis over time. Left: transparent lines are the trajectories of all countries over time (dimensions 1 and 2, other dimension see SI Fig. 1), colored by geographic regions, the straight lines are linear regressions over all data points of a region and the 95% confidence intervals over their coefficients. Right: World maps of slopes over time of dimension 1–5, the color represents the value of Sen’s slope, countries with significant slopes have black borders

3.4 Trajectories

Changes in the direction of trajectories are very likely to be a major disruption of a given development path. Some examples can be found in Fig. 7. For example, the earthquake in Haiti in 2010 coincides with a major disruption in the trajectory. The financial crisis and the onset of austerity measures can be noted from a dent in 2008 in the trajectory of Greece. A few years after massive privatizations in Argentina the trajectory of Argentina changes drastically. Major disruptions in the trajectory of the United States happen during

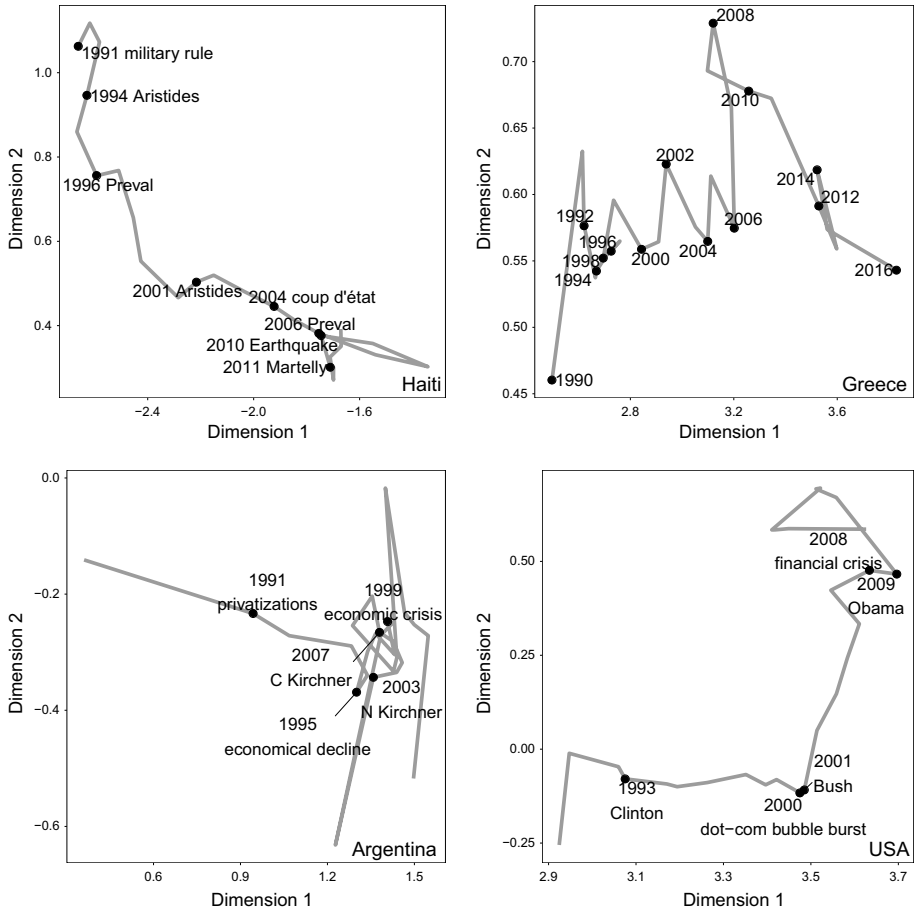


Fig. 7 Example trajectories. Haiti: Large jumps in the trajectory and kinks are preceded by changes in government and natural disasters. Greece: complete change of direction of the trajectory seem to appear during the financial crisis 2007–08. Argentina: The trajectory reflects important changes in economic policies. USA: The trajectory reflects the economic crises and changing presidencies

the burst of the dot-com bubble in 2000–2001 and the financial crisis in 2008. Attribution of changes to the trajectories to only these events can be challenging, and would require a formal causal framework (Pearl et al. 2016; Peters et al. 2017). For instance, in the case of the US, the changes in the trajectory could equally be attributed to changes in the presidency or to politics after 9/11/01. In the case of Argentina, it is not clear if the changes were caused by changes in politics during the Kirchner presidencies, problems that set in later after the privatizations, or a mixture of both, and remain of purely speculative nature.

In the overall view, some countries appear to change their centers of attraction recovered space of human development, e.g. Singapore in 1990 appears to be similar to the rich oil exporting Arab countries, but its trajectory suggests that it is currently gravitating towards most of the wealthy European countries, see Fig. 5. Countries that share similar history also seem to be close in the final dimensions, e.g. former Soviet countries, rich oil exporting nations, western European nations.

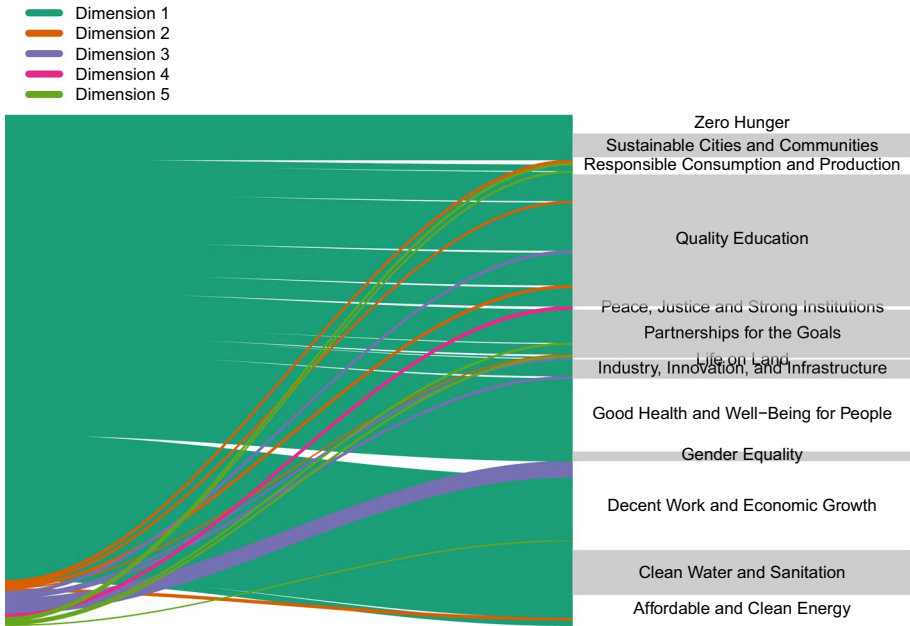


Fig. 8 Showing the importance of the dimensions for the SDGs, color code by dimension (left, unlabelled) are connected to the SDGs (right) through the corresponding WDIs (not shown, see text for details). The thickness of the connection reflects the distance correlation between the WDIs and the dimensions. See SI Fig. 3 for a more detailed version of the figure

3.5 Sustainable Development

To understand the relevance of the emerging dimensions for the different SDGs, we again use distance correlation and the WDIs that the World Bank uses to track the SDGs (United Nations General Assembly 2017b). We consider only the dimension with the maximum distance correlation to each WDI which is used to track an SDG. The results are shown in Fig. 8.

As most goals are poverty related, they load most strongly on the first dimension. The goals “Decent Work and Economic Growth” and “Industry, Innovation, and Infrastructure” also load on dimension 3, as this dimension describes the labor market. Dimension 2 describes educational and energy aspects and is related to “Affordable and Clean Energy” and “Quality Education”. We found a relationship between dimension 4 and the SDG “Peace, Justice and Strong Institutions” due to the homicide rate indicator. Dimension 5 was important to the “Partnership SDG and Responsible Consumption and Production”, due to relatedness of non-renewable energy sources and statistical reporting indicators.

Surprisingly, dimension two does not have any influence on the SDG “Achieve gender equality and empower all women and girls” despite describing aspects of gender equality. The reason for this may be that the SDG “Promote sustained, inclusive and sustainable economic growth, full and productive employment and decent work for all” is described by many of the variables loading on dimension two.

We can also see which SDGs are well represented by the data (the height of the SDG in Fig. 8) and which ones are not. For example, best represented are SDGs that represent

traditional ideas of development, such as “Quality Education”, “Decent Work and Economic Growth”, “Good Health and Well-Being for People”, while environmental SDGs such as “Life on Land”, “Life Below Water”, or “Climate Action” are not well or not at all represented.

4 Discussion

The assessment of development on the basis of a few key indicators has often proven very useful, but has also been controversial. As early as in the 1960s, GDP was recognized to be a very incomplete measure of development (Ram 1982; McGillivray 1991; Göpel 2016). Later, a large number of indicator approaches emerged, each constructed to describe specific aspects of development (Parris and Kates 2003; Shaker 2018). The large number of measured variables and derived indicators that are used today to describe development could suggest that global development is a high dimensional process requiring many indicators to describe it accurately. This perception contrasts with our finding that three quarters of the variability of the development space can be explained by only one dimension, and five dimensions recover 90% of variance. This indicates that the dimensionality of development is much lower than one would expect. Or, to put in other terms, the fact that many properties of development are highly correlated (Ghislandi et al. 2019) also means that one can summarize them efficiently in very few dimensions.

The notion that development is of low-dimensionality, however, does by no means imply that it is a “simple” process. In general it is well-known that low-dimensional spaces can still contain and depict very complex and unpredictable dynamics: Prominent examples are the logistic map (Verhulst 1845, 1847), describing population dynamics in a space of a single dimension, or the Lorenz (1963) attractor in physics, describing hydrodynamic flow in a three dimensional space.

The question whether data-driven indicators as presented here can be an alternative to classical indicators has been widely discussed (Ram 1982; OEDC 2008; Gapminder Foundation 2018). One argument in favour of such an approach is to overcome the lack of objectivity, which is a common criticism of classical indicators (Monni and Spaventa 2013; Göpel 2016). Consequently, PCA is increasingly used for the creation of wealth indicators (Filmer and Pritchett 2001; Smits and Steendijk 2015; Shaker 2018), as well as other approaches to identify suitable variable weights (Seth and McGillivray 2018). In our study we show that the PCA approach is less effective due to the strong nonlinear relations among the covariates present in the dataset.

Nonlinear data dimensionality reduction, however, makes the assessment of the identified dimensions difficult and hard to trace back to the underlying processes. Dimension 1, for example, includes both basic health and wealth variables. Figure 3 illustrates the reason for this. On the negative end of dimension 1, the maternal mortality rate is high and per capita income is low. When moving upwards along this dimension, first maternal mortality rates drop steeply, while the per capita income hardly changes. When moving towards the positive end of dimension 1, maternal mortality cannot decrease much further, as it is already close to zero, but the per capita income starts to increase strongly (Fig. 3). Combining both effects, dimension 1 manages to incorporate wealth as well as mortality related variables into a single (nonlinear) indicator. Each indicator can have a strong influence on a subset of a dimension (e.g. maternal mortality rate on the negative side of dimension 1) and a very low impact on other subsets (e.g. maternal mortality rate on the positive end of

dimension 1). Still, the fact that these factors co-vary in a way that we can represent them in a single dimension can guide the development of novel metric indices.

While dimension 1 allows for a relatively straightforward interpretation, we see in dimension 2 that there are more complex patterns to discuss. We find that Post-Soviet countries, Western European countries and Sub-Saharan African countries all lay on similar high coordinate values in dimension 2 (Fig. 5). Looking at variables that correlate strongly with dimension 2, we find that the participation of women in the labor market can be similar for very different states of dimension 1. We probably also uncover certain socio-cultural divides: most countries classified as “Middle East & North Africa” show a very low participation of women in the labor market while in other parts of the world participation of women in the labor market is much higher and does not depend on the geopolitical region of a country or its development status (see Fig. 5). For example in many European countries 45–50% of the working population is female, the same or even less than in most Sub-Saharan countries. Another variable that is orthogonal to development are crude death rates, where a rich country like Germany can have very similar rates to many countries in central Africa. Death rates in the WDI database are not resolved by age groups, given the aging societies in the developed world and the very young societies in many African countries, the death rates affect mostly older age groups in countries with high values on dimension 1, while it affects many younger age groups in the African countries.

In general, data-driven approaches to index construction can be criticized for not taking the polarity, i.e. the “direction”, into account (Mazziotta and Pareto 2019). This means that it remains subject to a subsequent interpretation whether a high value of a principal component (or non-linearly derived component) is a sign of a positive state in a certain domain or the opposite. The reason is that the underlying eigenvectors can be of arbitrary sign. However, we have shown (in Fig. 4) that an interpretation is possible, and the analysis of trends and trajectories can remedy this issue. Collapsing many aspects of development into a single dimension, which in turn forms the main gradient along which countries move over time, essentially expresses (nonlinear) covariations that should not be studied in isolation. For example, higher employment rates and an increased per capita income often go hand in hand. Here we showed that these connections between the 621 measured variables are so strong that a single dimension suffices to represent 74% of the variance. In this sense, we also see our approach as an opportunity to generate novel hypotheses on development that can guide policy making e.g. towards achieving the SDGs.

A general criticism of machine learning approaches is that underlying data biases are propagated and exacerbated. For instance, if the training data contain biases against minority groups, e.g. gender or race, these groups will systematically be put in a disadvantage by the algorithm (Barocas and Selbst 2016). Latest research tries to detect such biases (Obermeyer et al. 2019) and to avoid them during the training phase (Pérez-Suay et al. 2017). Therefore the implications of every machine learning based analysis have to be seen in the light of the dataset used for training. Here we summarize the WDI database, which represents the efforts of the World Bank to collect information on development at the global scale. The high variance explained by variables representing basic infrastructure, per capita income, and the population pyramid therefore reflects the (historic) emphasis that has been given to these kinds of basic indicators. For instance financial accounting has been ubiquitous, there are large scale efforts to monitor infrastructure and poverty, and census data is globally available.

Our analysis does not reveal an “environmental axis”, a component that is essential to sustainable development (Steffen et al. 2015). We can therefore also read our analysis as a gap analysis and conclude that future versions of the WDI data base should put more

emphasis on environmental data that are now widely available (Mahecha et al. 2020). Another essential component are inequalities (UNDP 2019). While some aspects are recovered by our analysis, such as between country inequalities on dimension one and some aspects of gender inequality on dimension 2, others do not emerge, e.g. income inequalities inside a country.

The best represented SDGs are those related to traditional ideas of development, while “Life on Land”, “Life Below Water” or “Climate Action” are not well or not at all represented. This shows a clear bias towards classical development data, and a lack of environmental data in the WDI data base. The reasons for this lie in the topics that have been emphasized for development historically (Griggs et al. 2013).

An analysis like the present one can be informative for policy making in various ways. It reveals general constraints of the development manifold, i.e. which combinations of WDIs are possible, which trajectories in the development space have been observed and which ones not. In particular, the trajectories can inform policy makers regarding the general present and past position of a country in this space beyond a single metric like the HDI (or our dimension 1). This means that also the less obvious changes, e.g. the changes of post-Soviet countries along dimension 2, can be taken into consideration.

Focusing on these dimensions is not trivial. It allows to target a few orthogonal aspects of developments only, instead of screening hundreds of individual WDIs. Another way how this analysis can guide policies is by seeing the results in the context of the dataset and pointing out weaknesses and underrepresented dimensions in the dataset, such as the environment and within-country inequalities. The key difference between our approach and the classical approaches is that we try to describe development space in its entirety, and hence the extracted components are neutral and agnostic to any societal or political agenda.

In particular the trajectory of single countries can yield essential information on important events for a country. The trajectories analyzed in this paper all showed changes that are obvious to the human eye, such as temporary deviations or changes in speed and directions. We could find connections for all of the observed changes in the trajectories of Fig. 7 with important socioeconomic or environmental events, although we were not able to automatically detect changes in all trajectories due to the different characteristics of each change. Future research is needed, to better understand the anomalies in the extracted trajectories.

In our opinion, a main advantage of data-driven approaches compared to classical indicator approaches is that the number of necessary indicators emerges naturally and the resulting indicators represent orthogonal features. The main disadvantage is the loss of indicators that represent very specific aspects of the data. Obviously, dimensionality reduction can only summarize the available data which also means that data incompleteness, data errors, and reporting biases are inherited—as it is also the case for classical indicators. Still, the proposed approach can help in the planning of adding measures of development and testing their redundancy with respect to the existing indicators, simplifying e.g. reporting of complementary dimensions of development.

A general limitation of the data under scrutiny is their aggregation at the country level. This means that our analyses cannot account for the often large socioeconomic differences and developments *within* a country. Also localized disasters may not influence the trajectory of a large economy as a whole, e.g. a large hurricane causing damage in Florida will only have a very marginal influence on the trajectory of the United States. Today there are efforts to collect data on sub-national levels which would alleviate this problem, see e.g. Smits and Permanyer (2019). However these efforts are relatively recent and there are still not many variables available.

5 Conclusions

In this study we investigated the “World Development Indicators” from 1990 to 2016 using a method of nonlinear dimensionality reduction. Our study led to three key insights. Firstly, the WDI database is of very low intrinsic dimensionality: We found that the WDIs are strongly interconnected, but we also showed that these connections are highly nonlinear. This is the reason why linear indices based on PCA cannot compress the information on human development that efficiently, while our approach only needs five dimensions to represent 90% of the data variance. The first dimension partly resembles the HDI, but also reveals much more differentiated patterns in low-income countries. The subsequent dimensions show orthogonal aspects such as the participation of women in the labor market and complex demographic dynamics. Quantifying such interactions uncovered by this approach can lead to new approaches to quantify different aspects of development. Exploring the meaning of the emerging dimensions allows us to understand which aspects of development are underrepresented in current databases. The second insight is that development as described by the dimensional space, remains to be a highly complex process that involves strong nonlinear interactions. We have elaborated some of these aspects, but a more profound exploration of the five-dimensional development space is still needed. Clearly, our approach can only account for the information in the data and ignore any additional aspects such as environmental issues that are clearly critical for sustainable development. The third insight is that single countries’ trajectories in the low dimensional space show abrupt changes that coincide with major environmental hazards or socioeconomic anomalies. As these changes in the trajectories can be of different nature, automatized detection is non-trivial and may require further causal explorations. Overall, our analysis gives new insights into the general structure of development which is of low dimensionality, but highly nonlinear and interconnected. Future work is needed to understand the observed trajectories in development space in much more detail, as well as to exploit them for achieving the Sustainable Development Goals.

Acknowledgements Open access funding provided by Projekt DEAL. We thank Juliane Mossinger and 3 anonymous reviewers whose comments greatly helped to improve the manuscript. We thank Andrew Durso and Julius Kraemer for proofreading the manuscript. G.K. acknowledges the support of the German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)- FZT 118. G.K., M.R., and M.D.M. thank the European Union for funding the H2020 Project BACI under Grant Agreement No. 640176 and the ESA for support via the “Earth System Data Lab” project. G.C.V. work has been supported by EU under the ERC consolidator Grant SEDAL-647423.

Code availability The code used in this analysis can be found under https://github.com/gdkrmr/low_dimensionality_of_development/ and <https://doi.org/10.5281/zenodo.3706483>. A docker container to reproduce the analysis can be found under <https://dx.doi.org/10.17617/3.3i>.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the

material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References


- Arenas-Garcia, J., Petersen, K. B., Camps-Valls, G., & Hansen, L. K. (2013). Kernel multivariate analysis framework for supervised subspace learning: a tutorial on linear and kernel multivariate methods. *IEEE Signal Processing Magazine*, 30(4), 16–29. <https://doi.org/10.1109/MSP.2013.2250591>.
- Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *SSRN Electronic Journal*,. <https://doi.org/10.2139/ssrn.2477899>.
- Bennett, R. (1969). The intrinsic dimensionality of signal collections. *IEEE Transactions on Information Theory*, 15(5), 517–525. <https://doi.org/10.1109/TIT.1969.1054365>.
- Costanza, R., Fioramonti, L., & Kubiszewski, I. (2016). The UN sustainable development goals and the dynamics of well-being. *Frontiers in Ecology and the Environment*, 14(2), 59–59. <https://doi.org/10.1002/fee.1231>.
- Filmer, D., & Pritchett, L. H. (2001). Estimating wealth effects without expenditure data-or tears: an application to educational enrollments in states of India. *Demography*, 38(1), 115–132. <https://doi.org/10.1353/dem.2001.0003>.
- Filmer, D., & Scott, K. (2012). Assessing asset indices. *Demography*, 49(1), 359–392. <https://doi.org/10.1007/s13524-011-0077-5>.
- Gapminder Foundation (2018) *Gapminder: Unveiling the beauty of statistics for a fact based world view*. Retrieved May 17, 2020, from <https://www.gapminder.org/>.
- Ghislandi, S., Sanderson, W. C., & Scherbov, S. (2018). A simple measure of human development: The human life indicator. *Population and Development Review*,. <https://doi.org/10.1111/padr.12205>.
- Ghislandi, S., Sanderson, W. C., & Scherbov, S. (2019). A simple measure of human development: The human life indicator. *Population and Development Review*, 45(1), 219.
- Göpel, M. (2016). *The Great Mindshift, The Anthropocene: Politik–Economics–Society–Science* (Vol. 2). Cham: Springer. <https://doi.org/10.1007/978-3-319-43766-8>.
- Griggs, D., Stafford-Smith, M., Gaffney, O., Rockström, J., Öhman, M. C., Shyamsundar, P., et al. (2013). Sustainable development goals for people and planet. *Nature*, 495(7441), 305–307. <https://doi.org/10.1038/495305a>.
- Kraemer, G., Reichstein, M., & Mahecha, M. D. (2018). dimRed and coRanking—Unifying dimensionality reduction in R. *The R Journal*, 10(1), 342–358.
- Kubiszewski, I., Costanza, R., Franco, C., Lawn, P., Talberth, J., Jackson, T., et al. (2013). Beyond GDP: Measuring and achieving global genuine progress. *Ecological Economics*, 93, 57–68. <https://doi.org/10.1016/j.ecolecon.2013.04.019>.
- Lorenz, E. N. (1963). Deterministic nonperiodic flow. *Journal of the Atmospheric Sciences*, 20(2), 130–141. [https://doi.org/10.1175/1520-0469\(1963\)020<0130:dnf>2.0.co;2](https://doi.org/10.1175/1520-0469(1963)020<0130:dnf>2.0.co;2).
- Mahecha, M. D., Martínez, A., Lischeid, G., & Beck, E. (2007). Nonlinear dimensionality reduction: alternative ordination approaches for extracting and visualizing biodiversity patterns in tropical montane forest vegetation data. *Ecological Informatics*, 2(2), 138–149.
- Mahecha, M. D., Gans, F., Brandt, G., Christiansen, R., Cornell, S. E., Fomferra, N., et al. (2020). Earth system data cubes unravel global multivariate dynamics. *Earth System Dynamics*, 11(1), 201–234. <https://doi.org/10.5194/esd-11-201-2020>.
- Marshall, M. G., & Elzinga-Marshall, G. (2017). *Global Report 2017, conflict, governance, and state fragility. Center for Systemic Peace*. Retrieved May 17, 2020, from <http://www.systemicpeace.org/vlibrary/GlobalReport2017.pdf>.
- Mazziotta, M., & Pareto, A. (2019). Use and misuse of PCA for measuring well-being. *Social Indicators Research*, 142(2), 451–476. <https://doi.org/10.1007/s11205-018-1933-0>.
- McGillivray, M. (1991). The human development index: Yet another redundant composite development indicator? *World Development*, 19(10), 1461–1468. [https://doi.org/10.1016/0305-750X\(91\)90088-Y](https://doi.org/10.1016/0305-750X(91)90088-Y).
- McRae, L., Freeman, R., Marconi, V., & Canadian Electronic Library (Firm) (2016) Living planet report 2016: Risk and resilience in a new era. WWF, oCLC: 1001121301
- Monni, S., & Spaventa, A. (2013). Beyond GDP and HDI: Shifting the focus from paradigms to politics. *Development*, 56(2), 227–231. <https://doi.org/10.1057/dev.2013.30>.

- Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464), 447–453. <https://doi.org/10.1126/science.aax2342>.
- OECD. (2008). Handbook on constructing composite indicators: Methodology and user guide. OECD, Paris, oCLC: ocn244969711
- Parris, T. M., & Kates, R. W. (2003). Characterizing and measuring sustainable development. *Annual Review of Environment and Resources*, 28(1), 559–586. <https://doi.org/10.1146/annurev.energ.y.28.050302.105551>.
- Pearl, J., Glymour, M., & Jewell, N. P. (2016). *Causal inference in statistics: A primer*. Hoboken: Wiley.
- Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2(6), 559–572.
- Pérez-Suay, A., Laparra, V., Mateo-García, G., Muñoz-Marí, J., Gómez-Chova, L., & Camps-Valls, G. (2017). Fair kernel learning. In M. Ceci, J. Hollmén, L. Todorovski, C. Vens, & S. Džeroski (Eds.), *Lecture notes in computer science, machine learning and knowledge discovery in databases* (pp. 339–355). New York: Springer. https://doi.org/10.1007/978-3-319-71249-9_21.
- Peters, J., Janzing, D., & Schölkopf, B. (2017). *Elements of causal inference: Foundations and learning algorithms*. New York: MIT press.
- Pradhan, P., Costa, L., Rybski, D., Lucht, W., & Kropp, J. P. (2017). A systematic study of sustainable development goal (SDG) interactions: A systematic study of SDG interactions. *Earth's Future*, 5(11), 1169–1179. <https://doi.org/10.1002/2017EF000632>.
- Ram, R. (1982). Composite indices of physical quality of life, basic needs fulfilment, and income: A 'principal component' representation. *Journal of Development Economics*, 11(2), 227–247.
- Rickels, W., Dovern, J., Hoffmann, J., Quaas, M. F., Schmidt, J. O., & Visbeck, M. (2016). Indicators for monitoring sustainable development goals: An application to oceanic development in the European Union. *Earth's Future*, 4(5), 252–267. <https://doi.org/10.1002/2016EF000353>.
- Seth, S., & McGillivray, M. (2018). Composite indices, alternative weights, and comparison robustness. *Social Choice and Welfare*, 51(4), 657–679.
- Shaker, R. R. (2018). A mega-index for the Americas and its underlying sustainable development correlations. *Ecological Indicators*, 89, 466–479. <https://doi.org/10.1016/j.ecolind.2018.01.050>.
- Silva, V. D., & Tenenbaum, J. B. (2003). Global versus local methods in nonlinear dimensionality reduction. In S. Becker, S. Thrun, & K. Obermayer (Eds.), *Advances in neural information processing systems* (pp. 721–728). New York: MIT Press.
- Smits, J., & Permanyer, I. (2019). The subnational human development database. *Scientific Data*, 6, 190038. <https://doi.org/10.1038/sdata.2019.38>.
- Smits, J., & Steendijk, R. (2015). The international wealth index (IWI). *Social Indicators Research*, 122(1), 65–85. <https://doi.org/10.1007/s11205-014-0683-x>.
- Spearman, C. (1904). "General intelligence," objectively determined and measured. *The American Journal of Psychology*, 15(2), 201–292. <https://doi.org/10.2307/1412107>.
- Stacklies, W., Redestig, H., Scholz, M., Walther, D., & Selbig, J. (2007). pcaMethods—A bioconductor package providing PCA methods for incomplete data. *Bioinformatics*, 23(9), 1164–1167. <https://doi.org/10.1093/bioinformatics/btm069>.
- Stanojević, A., & Benčina, J. (2019). The construction of an integrated and transparent index of wellbeing. *Social Indicators Research*, 143(3), 995–1015. <https://doi.org/10.1007/s11205-018-2016-y>.
- Steffen, W., Richardson, K., Rockström, J., Cornell, S. E., Fetzer, I., Bennett, E. M., et al. (2015). Planetary boundaries: Guiding human development on a changing planet. *Science*. <https://doi.org/10.1126/science.1259855>.
- Székely, G. J., Rizzo, M. L., & Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35(6), 2769–2794. <https://doi.org/10.1214/009053607000000505>.
- Tenenbaum, J. B., Silva, Vd, & Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500), 2319–2323. <https://doi.org/10.1126/science.290.5500.2319>.
- The World Bank.(2018a). *Sustainable development goals (SDG) data catalog*. Retrieved May 3, 2018, from <https://datacatalog.worldbank.org/dataset/sustainable-development-goals>.
- The World Bank. (2018b). *World development indicators (WDI) data catalog*. Retrieved May 3, 2018, from <https://datacatalog.worldbank.org/dataset/world-development-indicators>.
- Torgerson, W. S. (1952). Multidimensional scaling: I theory and method. *Psychometrika*, 17(4), 401–419. <https://doi.org/10.1007/BF02288916>.
- UNDP. (2016). *Human development report. Human development for everyone, United Nations Development Programme*. New York, NY: Human Development Reports.

- UNDP. (2018). Human Development Reports|United Nations Development Programme. <http://hdr.undp.org/>
- UNDP. (2019). Human Development Report 2019 Beyond income, beyond averages, beyond today: Inequalities in human development in the 21st century. United Nations Development Programme, New York, NY, USA: Human Development Reports.
- United Nations General Assembly. (2017a). *Global indicator framework for the Sustainable Development Goals and targets of the 2030 Agenda for Sustainable Development*. <https://doi.org/10.1891/9780826190123.0013>
- United Nations General Assembly. (2017b). *Work of the Statistical Commission pertaining to the 2030 Agenda for Sustainable Development*. http://ggim.un.org/meetings/2017-4th_Mtg_IAEG-SDG-NY/documents/A_RES_71_313.pdf
- Van Der Maaten, L., Postma, E., & Van den Herik, J. (2009). Dimensionality reduction: A comparative review. *Journal of Machine Learning Research*, 10, 66–71.
- Verhulst, P. (1845). Recherches mathématiques sur la loi d'accroissement de la population. *Nouveaux mémoires de l'Académie Royale des Sciences et Belles-Lettres de Bruxelles*, 18, 14–54.
- Verhulst, P. (1847). Deuxième mémoire sur la loi d'accroissement de la population. *Mémoires de l'Académie Royale des Sciences, des Lettres et des Beaux-Arts de Belgique*, 20, 1–32.
- Wolter, K., & Timlin, M. (1993). Monitoring ENSO in COADS with a Seasonally Adjusted Principal Component Index. NOAA/NMC/CAC, NSSL, Oklahoma Clim. Survey, CIMMS and the School of Meteor., University of Oklahoma, Norman, OK
- Wolter, K., & Timlin, M. S. (2011). El Niño/Southern Oscillation behaviour since 1871 as diagnosed in an extended multivariate ENSO index (MEI.ext). *International Journal of Climatology*, 31(7), 1074–1087. <https://doi.org/10.1002/joc.2336>.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Affiliations

Guido Kraemer^{1,2,3,4}  · Markus Reichstein¹ · Gustau Camps-Valls³ · Jeroen Smits⁵ · Miguel D. Mahecha^{1,2,4}

Gustau Camps-Valls
gustau.camps@uv.es
<http://isp.uv.es>

Miguel D. Mahecha
miguel.mahecha@uni-leipzig.de

¹ Max Planck Institute for Biogeochemistry, 07745 Jena, Germany

² German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, 04103 Leipzig, Germany

³ Image Processing Lab, Universitat de València, 46980 Paterna, València, Spain

⁴ Remote Sensing Centre for Earth System Research, Leipzig University, 04103 Leipzig, Germany

⁵ Global Data Lab, Institute for Management Research, Radboud University, PO. Box 9108, 6500 HK Nijmegen, The Netherlands