Verma et al. *EPJ Data Science* (2020) 9:8
https://doi.org/10.1140/epjds/s13688-020-00225-y

EPJ Data Science
a SpringerOpen Journal

REGULAR ARTICLE Open Access

A new set of cluster driven composite development indicators

Anshul Verma[1*†], Orazio Angelini[1†] and Tiziana Di Matteo[1,2,3]

*Correspondence: anshul.verma@kcl.ac.uk
[1] Department of Mathematics, King's College London, London, UK
Full list of author information is available at the end of the article
†Equal contributors

Abstract

Composite development indicators used in policy making often subjectively aggregate a restricted set of indicators. We show, using dimensionality reduction techniques, including Principal Component Analysis (PCA) and for the first time information filtering and hierarchical clustering, that these composite indicators miss key information on the relationship between different indicators. In particular, the grouping of indicators via topics is not reflected in the data at a global and local level. We overcome these issues by using the clustering of indicators to build a new set of cluster driven composite development indicators that are objective, data driven, comparable between countries, and retain interpretabilty. We discuss their consequences on informing policy makers about country development, comparing them with the top PageRank indicators as a benchmark. Finally, we demonstrate that our new set of composite development indicators outperforms the benchmark on a dataset reconstruction task.

Keywords: Development economics; Composite indicators; Information filtering; Clustering; World Development Indicators

1 Introduction

Economic indicators are vital in understanding and tracking the macroeconomic state and development of a country [1], informing government policy makers about the health of the economy and also for citizens to evaluate and assess any improvement in their life [2]. However, with the ever expanding number of different indicators and digital records of this data, it becomes difficult to interpret the high dimensional data as a whole, spot overall trends and see how different indicators are related to each other. Often qualitative or obscure factors are used to explain development such as the need to have a good education and healthy citizens.

Additionally, it is not agreed what factors affect development [3–11], and so arbitrarily chosen indicators are often used, ignoring specific information by excluding other indicators. In some cases, a more educated assumption is made by taking only indicators of relevance e.g. those relating to infrastructure. Even in these cases, different classes of indicators are treated separately to each other. Links with other classes of indicators e.g. poverty and infrastructure [12] are disregarded. This is especially relevant when one combines them in some way into composite indicators[13], which aim to describe several development indicators with just one composite version. These can range from more general indicators such as the Human Development Index (HDI) [14], which is used to measure the progress in life expectancy, education and Gross National Income per capita (GNI)[15], to more specific indicators such as the Global Connectivity Index (GCI) [16]. Composite indicators are also often used to summarise the state of a country relating to the specific objective of combining the chosen set of indicators e.g. GCI is used to track the extent of digital infrastructure of a country, whilst HDI is used to track overall human development. In literature they have been used, for example, to relate cancer rates to development [17] or to produce a global rank of a country's competitiveness.

Whilst aggregating indicators into composite ones seems to be a good solution to the problem of summarising information from many different indicators, we propose that the high number of possible ways to combine them calls for the developement of guiding principles on how this should be achieved. Moreover, some indicators are calculated differently for different regions [18], making comparisons based on them much more difficult. By knowing how indicators are inter-related to each other, we will be able to understand in a data-driven, objective way which indicators are most important in characterising a country and how they should be combined to produce economically meaningful composite indicators.

Dimensionality reduction can help here by providing a smaller but faithful version of the relationship between the vast number of available development indicators [19]. This paper proposes to study these relationships in an unbiased way, using these relationships as a basis to propose a new set of composite development indicators. Differently to previous work, we make no subjective restriction on the type of indicators we study, drawing from a large range of scope of indicators to study the relationship between the indicators emerging from the data itself. We test whether we can indeed separate the indicators into different pre defined groups based on the different factors proposed that affect development using PCA (Principal Component Analysis) and Random Matrix Theory (RMT) [20]. We find that the broad topic category they are assigned, e.g. health vs economic vs infrastructure, is not necessarily the best way to aggregate them. We also employ hierarchical clustering algorithms for the first time to analyse the structure of indicators rather than countries, finding that the indicator clusters are a mixture of topics but still retain an economic interpretation. We use these results to overcome traditional problems faced in making composite indicators such as how/what indicators to aggregate to derive a new set of objective, data driven, interpretable and country comparable composite indicators. Leveraging on these composite development indicators, we observe useful observations for policy makers, such as the ability of mobile phone adoption to be able to distinguish between underdeveloped countries. Next, we provide a new application of network filtering to find subsets of highly influential indicators based on PageRank [21]. Finally, we compare the performance of our composite indicators to a random benchmark, a subset of influential indicators and PCA, concluding that our proposed composite indicators outperform the others.

In the context of this problem, dimensionality reduction has been applied in [22], where Principal Component Analysis (PCA) was used on a set of restricted indicators to form infrastructure composite indicator. This indicator was then used to examine the direction of the causal relationship between infrastructure capability and economic growth. The authors of [23] compare the pre-set weights of the HDI to that derived from a PCA. Hierarchical clustering has been applied to analyse clusters of countries such as in [24, 25]. However, in either cases either a restricted set of indicators are used or the focus of the work is on countries, rather than the analysis and development of new indicators themselves. The problem of forming composite indicators using PCA and dimensionality reduction is emphasised in [26], where it has been shown this approach could overlook important information. In particular for PCA, this includes ignoring information from components other than the first and difficult to interpret weights.

Network filtering techniques [27, 28] and their related hierarchical clustering algorithms [29–31] have also proved to be useful when analysing data, with wide ranging applications from finance to biology [30–32]. Network filtering techniques view a similarity matrix as a network, each node being a feature and each link having a weight with the respective non-zero correlation. Within this framework, removing noisy entries in the correlation matrix can be translated into finding a sparse version of the similarity network. These techniques aim to extract the backbone of the structure between generic features by enforcing sparsity in a specific way to the particular technique. The induced sparsity of the network helps make hidden structures more visible. One successful example of this is the Minimum Spanning Tree (MST) [27, 33], which imposes that the correlation matrix is a tree that maximises the total weight of links, and has been applied in a diverse number of fields from electricity networks to taxonomy [32, 33]. A generalisation which includes the possibility of loops is the Planar Maximally Filtered Graph (PMFG) [28, 34], which instead imposes a weaker constraint that the network is planar i.e. it can be embedded on a sphere without any links crossing. Hierarchical algorithms are also highly related and aim to group features with similar properties into clusters that organised in a hierarchical fashion in the form of a dendrogram. An example of this is the Directed Bubble Hierarchical Tree (DBHT) algorithm that is based on the PMFG, having been used for finance [31, 35] and in gene expression data [30]. In particular, the DBHT algorithm has also been shown to outperform other hierarchical clustering algorithms.

This paper is organised as follows. The second section is a description of the dataset and how we amalgamate topics together. In the third section we apply a PCA analysis to our dataset, which we use to show the difference between the structure of the empirical correlation matrix and the preassigned topics. We then find the clustering using the DBHT algorithm in Sect. 4. Developing the clustering results further to form a novel set of composite indicators in Sect. 5, we observe some interesting features of our composite indicators in Sect. 6. For Sect. 7 we apply the PMFG to the empirical correlation matrix in order to derive some influential indicators via PageRank. In Sect. 8 we compare the performance of our composite indicators with a random benchmark and top indicators taken from the PageRank. Finally, we discuss the dynamic stability of our results in Sect. 9 and draw some conclusions in the final section.

2 WDI dataset

The World Development Indicator (WDI) dataset is a vast collection of various yearly development indicators for $C = 218$ countries (where C is the number of countries) and are taken from official, internationally recognised agencies [36]. Note that we have applied the imputation scheme the distribution regularisation procedure detailed in the Supplementary Information Sects. 1.1 (to treat missing data) and in 1.2 (to standardise and normalise the indicators) respectively (see Additional file 1). Note that we have also checked that the amount of missing data does not change the results significantly—see Supplementary Information Sect. 1.1 for more information. We shall use a total of $T = 19$ years, where $t = 1, \ldots, 19$ represents the years from 1998 to 2016 . The number of indicators contained within the dataset is $N = 1574$, and the objectives for collecting these indicators ranges from well known economic data such as Gross Domestic Product (GDP), education data such as the literacy rate and population health such as infant mortality rate. Hence both the large number of indicators and the diverse range of granularity and objectives makes this dataset a perfect candidate in order to study the relationships between different classes of indicators and to infer and derive conclusions that hold globally. Note that we also remove highly correlated indicators with a correlation coefficient of 0.95 since some indicators can be trivially related together e.g. percentage of population that are males and the same but for females, which would bias the results. This reduces the number of indicators N to 1448.

2.1 Amalgamating topics

The indicators are also divided into 94 different topics that include different classes of economic indicators such as Economic Policy and Debts: National Accounts: Growth Rates, which measures growth rates of agriculture, industry, manufacturing and services sectors, and Education: Participation, which measures participation rates across gender, age groups in various levels of education. We show the distribution of all such topics using this classification in Fig. 1(left). We can see that most of the groups of indicators make up a very small fraction of the indicators, which would mean that any averaged statistic across within each group would be subject to significant noise. To counteract this, we aggregate the topics for each classification based on their root objective e.g. Education: Participation and Education: Efficiency are both classes of indicators relating to education and hence we combine these two groups into one group Education, similarly Health: Nutrition and Health: Disease Prevention are combined into Health. Applying this procedure to the entire dataset produces $g = 1, \ldots, G = 12$ different topics for the indicators which we indicate in Fig. 1(right). We can see that each topic has a larger number of indicators, which will increase the statistical reliability of any conclusions drawn from the data.

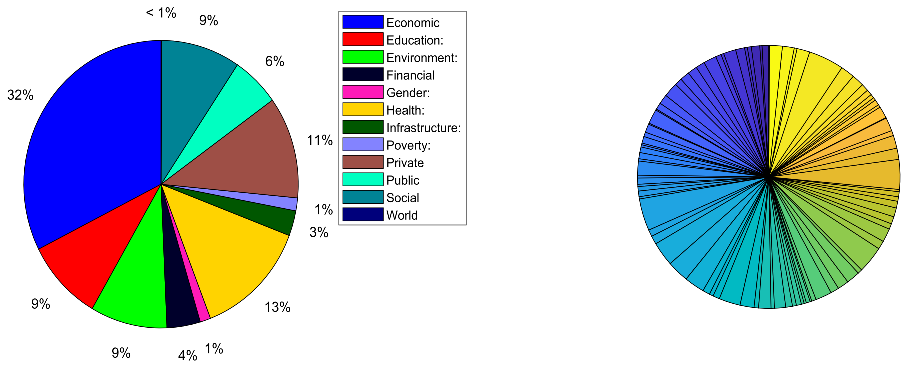


Figure 1 Pie charts for indicator classification. (Left panel) Pie chart of the distribution of the indicators using the classification from the WDI dataset. (Right panel) The same but with the aggregated classification. The legend of the bottom chart indicates the names of the derived aggregated classification of the topics of the indicators

2.2 Data structure

We here start exploring the correlation structure across years as a measure to quantify the relationship between indicators. We aggregate values across years in order to average out correlations that might hold only for specific periods or groups of countries. This also helps to reduce the noise in the correlation matrix, since one-year matrices would be too shallow to reliably obtain correlation estimates.

With this mind, we organise the data matrix **X** as follows. It consists of C matrices of size $T \times N$ matrices stacked vertically, with each cross sectional block representing the data for one specific country c and each column reporting the data for indicator $i = 1, \dots, 1448$. For each cross section, the entries in the first row and i column are the values of the indicator i for $t = 1$ and the last row are the same but for $t = T$. In order to discard spurious correlations in the data, we remove trends by taking the first difference, that is for each block of data we calculate

$$\Delta\mathbf{X}(\tilde{t}, c, i) = \mathbf{X}(t+1, c, i) - \mathbf{X}(t, c, i) \tag{1}$$

with $\mathbf{X}(t, c, i)$ is the value of indicator i for country c at year t. $\mathbf{X}(t+1, c, i)$ is similar but with $t + 1$. $\Delta\mathbf{X}(\tilde{t}, c, i)$ represents the first difference between $\mathbf{X}(t+1, c, i)$ and $\mathbf{X}(t, c, i)$, with $\tilde{t}$ running from $1, \dots, T - 1 = 18$. Every $\Delta\mathbf{X}(\cdot, c, \cdot)$ has $T - 1$ rows and N columns. Stacking each of these vertically forms the $Y = 3924 \times N$ matrix $\Delta\mathbf{X}$, which now contains all the differenced values for all countries and all time steps.

To encode the relationship between the indicators we use the empirical Pearson correlation matrix **E**, which can be calculated from a zero mean, standardised $\Delta\mathbf{X}$ as

$$\mathbf{E} = \frac{1}{C(T-1)} (\Delta\mathbf{X})^{\dagger} \Delta\mathbf{X}, \tag{2}$$

where † represents the transpose. Therefore, we aim to understand the multivariate dependence between development indicators through analysing the main driving factors of the structure of **E**. However, using the raw correlation matrix would be unwise due to its large size (1448 by 1448) and noise present in the system, potentially leaving a certain amount of redundant information in **E**. As mentioned earlier, we can distill the information given in **E** to a smaller version using dimensionality reduction, which should also have the added benefit of making it easier to interpret the structure of **E**.

3 PCA analysis

Within the class of dimensionality reduction methods, PCA is a popular and easy to apply technique used on correlation matrices [37]. This technique has been successfully applied in many diverse areas, ranging from finance [38] to molecular simulation [39]. PCA accomplishes the task of dimensionality reduction by taking a subset of the orthogonal basis for the correlation matrix **E** [37]. The first principal component corresponds to the eigenvector with the highest eigenvalue, providing the direction where the data is maximally spread out i.e. explains the most variance of the system. Each subsequent principal component has a lower eigenvalue and thus explains a lower fraction of the total variation of the system. Therefore, we can reduce the dimensionality of the correlation matrix by taking a subset of principal components, hoping to encode most of the total variance of the data. This subset can be chosen with the help of Random Matrix Theory (RMT) [20], which studies the properties of matrices drawn from a probability distribution. In our specific context of forming composite indicators in a data-driven way, one could then use the chosen subset of components as a basis for composite indicators.

In this section, we apply PCA to the correlation matrix **E** on the dataset of Sect. 2, finding the distribution of its eigenvalues, using results from RMT to help interpret it. We then analyse the contribution of each topic defined in Sect. 2.1 to the eigenvectors corresponding to the principal components.

3.1 Eigenvalue spectrum

As is customary in Random Matrix Theory, we fitted the Marčenko–Pastur (MP) distribution [40] to the eigenvalue distribution of **E** to discern what part of the eigenvalue spectrum is less likely to be a product of finite-sampling noise. We found that MP does not fit our eigenvalue distribution well, which suggests that there is structure in the whole distribution, as opposed to just its right tail. We shuffled the data to destroy all correlations between indicators, and obtained an eigenvalue distribution that fitted the MP near perfectly. These findings suggest that choosing only a subset of the principal components obtained by PCA is likely to discard relevant information. In other words, this is a clue that PCA might be unsuitable to reduce dimensionality on this dataset. For a more detailed discussion of the procedures in this subsection, we refer to the Supplementary Information Sect. 2.

3.2 Eigenvector interpretation

We investigate what the interpretation of the eigenvectors is by calculating the contribution of each of the G topics from Sect. 2.1 that divide the indicators. This will reveal the structure with respect to topics of the principal components so we can see if they are dominated by one specific topic. The analysis will also be particularly relevant for the earlier principal components that are the main contributors to the variance of the system, which will bring to the surface any topics which are more significantly contributing to development.

Specifically, we project the eigenvectors $\mathbf{v}_i$ of **E** onto the G topics which divide the indicators that we defined in Sect. 2 using the projection matrix **P** with entries

$$P_{ig} = \begin{cases} 1/N_g & \text{if } i \text{ is in topic } g, \\ 0 & \text{else,} \end{cases}$$

where N_g is the number of indicators that are part of topic g. From this, for every we can define $\boldsymbol{\rho}_i$, which is G-dim vector with entries $\rho_{g,i}$, and is computed as

$$\boldsymbol{\rho}_i = \gamma_i \mathbf{P}\mathbf{v}_i, \tag{3}$$

where γ_i is the normalisation constant $\sum_{g=1}^{12} \rho_{g,i}$. Each entry of $\boldsymbol{\rho}_i$ gives the contribution of the gth topic to the ith eigenvector. As an example, we plot $\boldsymbol{\rho}_i$ for the top 6 principal components in Fig. 2. In Table 1, we report the one-sided p values of ρ_g for testing against the null hypothesis that the contribution from the topic to the principal component is random using the procedure detailed in Supplementary Information Sect. 3. The bolded values are those below the 5% significance level where we reject the null hypothesis. By

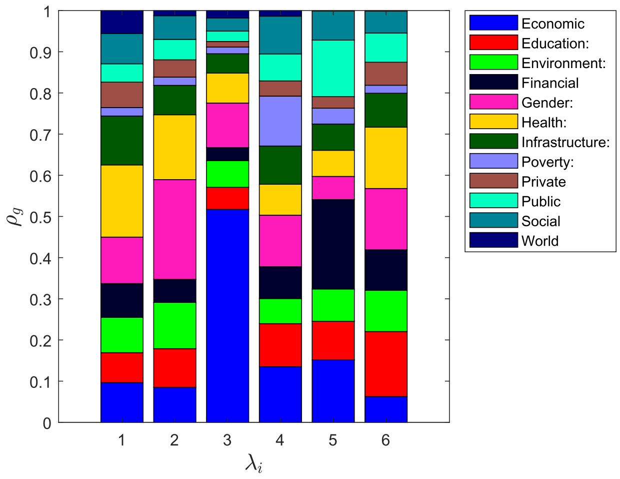


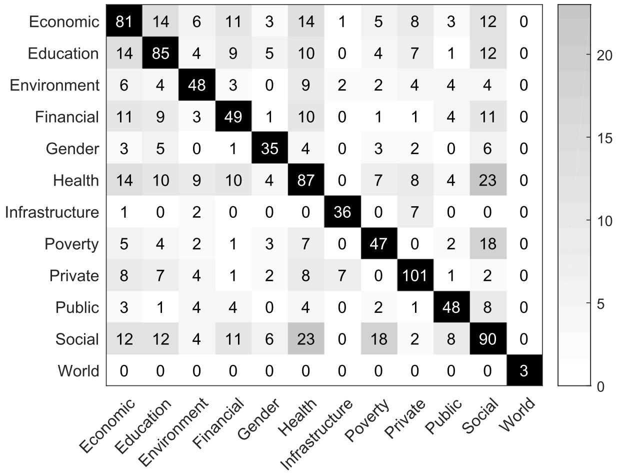
Figure 2 ρ_g for the top 6 principal components. Bar chart of the ρ_g defined in Eq. (3) for the top 6 principal components of **E** using the 12 topics of the indicators in Sect. 2.1. The legend corresponds to these 12 topics

Table 1 One sided p values of the ρ_g defined in Eq. (3) using the 12 topics defined in Sect. 2.1. The values were calculated using the procedure detailed in the Supplementary Material Sect. 3 and the null hypothesis used is that the ρ_g is random. The bolded values are the ones which are below the 5% significance level

	pval					
'Economic'	0.109263	0.503665	**0**	**4.60E-07**	**0**	0.965214
'Education'	0.818697	0.237016	0.987429	0.05904	0.224225	**6.90E-07**
'Environment'	0.447607	**0.01599**	0.932927	0.959271	0.689149	0.110319
'Financial'	0.543933	0.958713	0.999491	0.646416	**0**	0.200756
'Gender'	0.140525	**3.68E-06**	0.175278	0.07279	0.849132	0.016654
'Health'	**0**	**0**	0.838467	0.788524	0.952838	**0**
'Infrastructure'	**0.043397**	0.733465	0.983559	0.312448	0.854746	0.517028
'Poverty'	0.999458	0.999486	0.999633	0.079216	0.98248	0.999533
'Private'	0.958778	0.99881	0.999667	0.999483	0.999664	0.983491
'Public'	0.996577	0.990442	0.999663	0.896429	0.000541	0.812393
'Social'	0.80534	0.97684	0.999635	0.287871	0.853167	0.988461
'World'	0.41984	0.709084	0.656062	0.697268	0.923661	0.909286

looking at Fig. 2 and Table 1, we see that for the first principal component although other topics contribute to the largest eigenvalue, the statistically significant contributions come from the Health and Infrastructure. Similarly for the second principal component the Environment, Health and Gender topics make a statistically significant contribution, and for the third principal component only the Economic related indicators make a significant contribution.

We have also plotted the number of times two topics are simultaneously significant across all principal components in Fig. 3, with a darker grey indicating a higher number of times this occurs. We use a 5% p value with a Bonferroni correction of N, giving the actual p value used to be 3.45×10^{-3}. The black diagonal terms give the number of times a single topic is significant across all principal components using the same p value. If the indicators could be neatly divided into topics then we should see no interaction between them so that Fig. 3 will look almost like a diagonal matrix. In fact, we see that some of the off diagonal elements are quite large relative to the diagonal elements e.g. Education vs



	Economic	Education	Environment	Financial	Gender	Health	Infrastructure	Poverty	Private	Public	Social	World
Economic	81	14	6	11	3	14	1	5	8	3	12	0
Education	14	85	4	9	5	10	0	4	7	1	12	0
Environment	6	4	48	3	0	9	2	2	4	4	4	0
Financial	11	9	3	49	1	10	0	1	1	4	11	0
Gender	3	5	0	1	35	4	0	3	2	0	6	0
Health	14	10	9	10	4	87	0	7	8	4	23	0
Infrastructure	1	0	2	0	0	0	36	0	7	0	0	0
Poverty	5	4	2	1	3	7	0	47	0	2	18	0
Private	8	7	4	1	2	8	7	0	101	1	2	0
Public	3	1	4	4	0	4	0	2	1	48	8	0
Social	12	12	4	11	6	23	0	18	2	8	90	0
World	0	0	0	0	0	0	0	0	0	0	0	3

Figure 3 Number of times two topics are statistically significant in the same principal component. Grey scale map with the off diagonal entries giving the total number of times that the topics labelled by the corresponding row and column are simultaneously statistically significant across all principal components. A p-value of 0.05 with a Bonferroni correction is used so that the new p-value is $0.05/N$ or 3.45×10^{-3}. The diagonal entries are the number of times a single topic is significant at the same p-value across all principal components

Economic and Social vs Health, which indicates that the topics are indeed interacting. We can therefore conclude from this analysis that there is not a clear, single topic that contributes more than others and that in fact statistically significant contributions can come from different topics that combine in different ways.

This has implications for composite indicators aiming to capture some particular aspect of development such as GCI and HDI since it suggests that the inclusion of certain indicators which focus on other aspects of development might improve the quality of the composite indicator. Conversely, some indicators may actually not be representative of the aim of the composite one, which means including it would add no information with respect to the aim of the composite indicator whilst also simultaneously increasing complexity. These problems have also been mentioned in [26] and [41], where it has been shown that potentially important indicators are ignored when forming composite indicators from PCA. Overall, we can conclude that whilst the principal components indicate that the correlations between indicators contains interesting structure, it is difficult to use PCA to form new composite indicators. This means we must turn to other methods to achieve both of these goals.

4 Interpretation of the clustering from the DBHT

This section analyses the relationships between indicators in a data driven way where we make as little assumptions about the structure of the data as possible. In this way, we can develop an interpretation and partition of the indicators which is consistent with the data. In the previous section, we showed that this is not possible with PCA and by dividing indicators based on their a priori topic given in Sect. 2.1.

Hierarchical clustering algorithms [42], which group together data with similar properties in a hierarchical fashion, and their associated network filtering techniques will help in this respect. This is because we can consider information from all indicators. Hierarchical clustering is advantage in our context since it has been shown with many other different kinds of data originating from complex systems that data is organised hierarchically [43, 44]. Therefore it seems natural to apply a hierarchical clustering algorithm to discover this structure. Once we apply the clustering algorithm on the correlation matrix, we have a natural way of accomplishing dimensionality reduction by using one variable to describe each cluster of nodes, with the collection of clusters forming the reduced correlation matrix. The ones associated to network filtering algorithms leverage on the topological properties of the filtered network.

We shall use the PMFG network filtering technique because it is able to retain a higher amount of information about the system than the MST. This is because it preserves a greater number of links of the original network and in fact contains the MST as a subgraph [28]. This is important for us since the MST is a tree and thus contains no loops, whereas the PMFG contains 3 and 4 cliques, and we would like to avoid discarding relevant information about the relationship between indicators. For the PMFG, the associated clustering algorithm is the DBHT algorithm, which takes advantage of the 3 clique structure of the PMFG. There has been objections to the use of cluster analysis due mainly to the distance between points as a metric [45, 46] since it does not measure the similarity between variables. Instead, the DBHT algorithm forms the clustering by measuring forming a distance matrix directly from the similarity matrix, which encodes the relationships between variables. In our case the similarity matrix is **E**, with the entries of the distance matrix D_{ij} commonly defined as [27, 31, 47]

$$D_{ij} = \sqrt{\bigl(2(1 - E_{ij})\bigr)}. \tag{4}$$

D_{ij} therefore measures how far two indicators are in terms of their correlation.

The main advantage of using the DBHT algorithm is that it does not need prior input into the number of clusters, making it preferable over other clustering algorithms so that we can make a-posteriori comparison with less assumptions [30, 31]. This is important for us since we want to uncover the structure of correlations between indicators, making as few assumptions as possible on this same structure. Furthermore, the DBHT algorithm has been shown to retrieve a higher amount of meaningful information[31].

In this section, we investigate whether the indicators can be divided into their topics by applying the DBHT algorithm to **E** in Sect. 4.1. Then by analysing clusters individually, we look for their dominating topics and what their possible interpretation is in Sect. 4.2.

4.1 DBHT results and interpretation

We apply DBHT to **E**. It identifies a total of $K = 102$ clusters which we label $k = 1, \ldots, K$, significantly more than the G preassigned topics, with an average cluster size of 14.2. In Fig. 4 we summarise the clustering labels obtained from DBHT and its topic composition, with the height of the bar representing the number of indicators in each cluster N_k. Each bar is further divided by colours which represent how many indicators belong to that particular topic. Figure 4 shows that cluster sizes are highly heterogeneous—the biggest cluster has 111 indicators versus the smallest with 4.

We can see that some clusters are dominated by certain topics—for example cluster 41 is dominated by economic indicators and in particular indicators related to countries' current account balance and external balance of trade. At the same time however, there

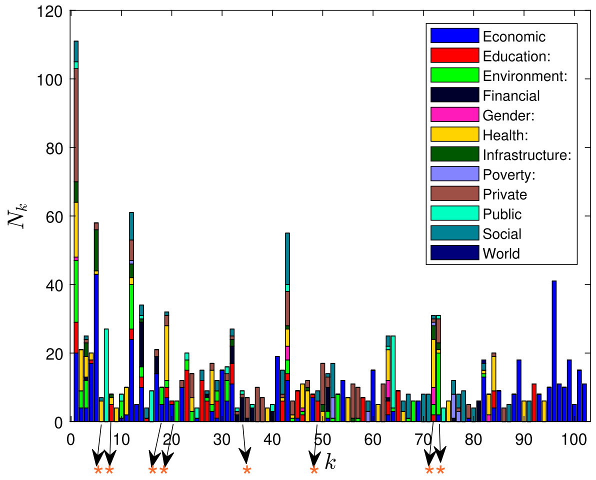


Figure 4 Distribution of topics amongst DBHT clusters. The cluster label *k* versus the number of indicators in each cluster N_k. Each bar is divided into the number of indicators in cluster *k* which belong to each topic, with each colour corresponding to each topic according to the key on the right. The arrows with the orange asterisks point to clusters (starting from the left) 6, 8, 18, 20, 34, 49, 72, 73, which are used in Fig. 5 and Fig. 6

are also some clusters which are instead a mixture of topics but still have an interpretation based on the indicators contained within them such as cluster 72, which contains indicator from disparate topics. A closer inspection reveals that this cluster is made of indicators about access to electricity, railways size, primary and secondary education expenditure, health-related indicator such as HIV incidence and hepatitis immunization, access to sanitation facilities, prevalence of underweight children, number of women who justify a husband's beatings, and the Gini index. All these measurements can be easily used to characterize underdeveloped countries [48–54].

Another interesting fact that we can observe from the data is that the cluster 5 contains very important economic indicators such as GDP per capita, value added contributions of agriculture, industry, manufacturing, services and trade, and also imports/exports of goods and services as a fraction of GDP. This same cluster also contains indicators directly related to measuring the innovation output of a country such as patent, trademark and industrial design applications, suggesting that innovation is an important factor in economic development. We can interpret this by realising that innovation led growth increases productivity through the accumulation of knowledge obtained via education, new products or better processes [11, 55].

Many other interesting clusters are found, such as number 11, which seems to relate to underdevelopement with it contents relating to life expectancy, foreign aid, drinking water availability, fertility rate, and percentage of women married before the age of 18. Cluster 21 puts together CO2 emissions, alternative and nuclear energy, combustible renewables and waste, hydroelectric sources prevalence and power distribution losses. Cluster 101 describes the distress status of a country's debt, including indicators about how much of it has been rescheduled or forgiven.

4.2 Similarity of the DBHT clustering with the topics

Once we have established that each of the clusters has an economic meaning, we quantify how much the clustering outputted by the DBHT in Fig. 4 is similar to the clustering based on topics. Therefore, we will be able to see overall how close the two divisions of topics are in a quantitative way. We do so using the Adjusted Rand Index (ARI), which is 1 if there is a perfect agreement, −1 if there is an anti-agreement and 0 if there is no agreement [56]. It has been successfully used in [31]. Computing the ARI to compare the output of the DBHT algorithm and the topics distribution, we find that this value is 0.0456 i.e. quite close to 0, which corroborates our previous conclusion that overall the clustering of the data is not in general linked to that based on topics.

The analysis can also be made at local level by seeing if any topics have a significant presence in each cluster. In this way, it allows us to also to see locally if more than one topic might be present in each cluster, which is important since whilst on a global level there may not be much similarity . Practically, this is achieved by using the procedure proposed in [57]. Specifically, we test statistically, using a one sided test, the null hypothesis that a cluster k from the DBHT and the gth topic have m common elements is random. Under the null hypothesis, this distribution is hyper geometric. If the null hypothesis is rejected, it means that statistically we say that the gth topic is *overexpressed* in cluster k. We apply this procedure to each of the DBHT clusters and topics using the p value of 8.17×10^{-6} (which is 0.01 with a Bonferroni correction [58] of $1/2KG$), recording the number of overexpressed topics in each cluster. The results of this procedure reveals that whilst a majority of clusters have one or two topics overexpressed, there are a total of 49 clusters which have no overexpressed topics. These particular clusters of indicators still have an economic meaning. For example, cluster 32 contains indicators relating to tertiary education such as pupil to teacher ratio in tertiary education and completed education at a tertiary level, which belong the education topic. However, it also contains indicators such as scientific and technical journal articles, which is classed as relating to infrastructure. These indicators may be linked e.g. because scientific articles are usually always published by authors with at least a tertiary level education. This confirms our conclusions that overall at a system wide and local level, the clustering of the data does not reflect the information given by the topics, suggesting that indicators do not necessarily correlate with other indicators of the same type.

5 Deriving new composite development indicators from DBHT

In the previous section we showed that the distribution of topics amongst the indicators is not an accurate description of the data and may miss key information about the relationship between different classes of indicators. This means that composite indicators based on this premise such as the HDI or the WEF-GCI infrastructure pillar may not be the best way of combining indicators. We want to propose a new set of data driven composite development indicators which can encapsulate this new information based on the results given in Sect. 4.1. In doing so, we would overcome traditional problems faced when forming composite development indicators, mainly on how and which indicators we should aggregate. This section is dedicated to describing a way of using the results in Sect. 4 to derive a novel set of cluster driven composite development indicators.

To define each composite indicator we shall use the set of clusters from DBHT given in Sect. 4.1. It provides a natural way to select the indicators to combine for our composite indicators since each cluster contains indicators which share similar properties, and also has an economic interpretation as highlighted in Sect. 4. Hence, aggregating information for indicators which are members of the same cluster enables us to simply and efficiently summarise the economic information contained within them. In contrast with PCA [26, 41], since we rely on the DBHT clustering, we automatically include information coming from all the indicators since every indicator is a member of some cluster that defines its respective composite version. Condensing the complimentary insight offered by indicators in the same cluster also overcomes the need to make 'educated' assumptions about which indicators are to be combined that other alternative composite indicators often use [14]. In this way, we can more clearly see the overall behaviour of each set of indicators in cluster k by using the corresponding composite indicator value as a proxy. DBHT also is significantly advantageous in this respect since it requires no prior input of the number clusters (and thus number of composite indicators) needed to describe the properties of the data) [30, 31]. Potentially, opposite polarities of indicators in the same cluster could make the interpretation of the value of any composite indicator formed from the same cluster problematic [26]. However, an advantage of forming composite indicators from the DBHT is that the clusters from the sparse correlation network PMFG, which we calculated has only 1% of its non negative entries as negative. Hence, the indicators in the same cluster have same polarity.

5.1 Method used to calculate the composite indicators

Here, we shall define the method used to calculate the new composite development indicators based on the results of Sect. 4.1. In the k composite indicator we want to capture the average behaviour of all indicators in that cluster. Therefore, we aggregate the indicators in cluster k by using the median value across all indicators within this cluster. We can do this because the indicators are standardised and normalised via the procedure in the Supplementary Information Sect. 1.2. This forms composite indicator k, I_k, defined as

$$I_k = \text{median}_{i\in\text{cluster}\,k}\,\mathbf{X}, \tag{5}$$

where the notation $i \in \text{cluster}\,k$ indicates that we only take indicators i that are members of cluster k. An advantage of using the median over the arithmetic mean or even a weighted mean is that the median is more robust to outliers. The median is a valid measure across the different indicators because the entries of $\mathbf{X}$ are also standardised, meaning that their scales are all the same. We highlight that we have chosen to use the median for every k since this provides us with a consistent methodology so that the precise details of how I_k is calculated do not change for every k. This improves some existing methods used in the literature where for example the same indicator may be calculated in different ways for different regions [18] meaning that we can make valid comparisons between indicators. Furthermore, a PCA based approach would require the use of weights, which has been argued to be hard to interpret economically [26, 59]. Instead, the median used does not require weights as an input to form the composite indicators we propose. We use this method to calculate the set of I_k, giving 102 indicators in total and call this set of cluster driven composite indicators (CDCIs).

6 Using the CDCIs to understand country development

One of the main uses of indicators is to track the country development of a country to assess its progress. This is important since it gives an idea of what has been achieved in terms of country development and where to focus policy changes to affect country development positively. One can also use indicators to compare countries either pair wise or globally. On the last point, CDCIs can be useful because the methodology used to compute them is not reliant on subjective, country dependent criteria, which means we can make fair comparisons between different values of the CDCIs for different countries at different times. By comparing the CDCIs with each other for all countries, we can therefore investigate whether they can be used to assess the development of a country. The main CDCIs we shall concentrate on are 6, 8, 18, 20, 34, 49, 82, 73, which are marked by orange asterisks in Fig. 4.

In Fig. 5, we provide some examples of comparisons between different indicators. Overall, we remark that all the plots have a 'hockey-stick' shape. We can specifically see for the plots in the top two panels the vertical leg of the hockey-stick shape is made of developing countries, whilst the horizontal leg, indicating a saturation effect, is made of developed countries. This is interesting since it suggests a country level transition from a group consisting of developing nations to one with developed nations. In fact, this further supports the so called two regime hypothesis [22, 60], where countries below a barrier struggle to develop consistently, which corresponds to nations in the vertical leg of hockey-stick shape. Countries that overcome this barrier have or are experiencing high growth in development, which are represented by the horizontal leg of the hockey-stick shape. This transition can clearly be observed to be consistent across time from the bottom panel comparing I_6 and I_{72} and I_{73} and I_{72}, with the years 1998 and 2016 overlayed.

As a consequence of the consistency of the observed hockey-stick shape, we can make interesting observations regarding the particular pair of CDCIs being plotted. Specifically in the top panel of Fig. 5, we plot I_{34} against I_{49}, where the former represents those who have use mobile and banking services and the latter come from primary school statistics, a key signature of development [61]. We see here that it seems that the vertical leg access to mobile phone technology can, for developing countries, characterise their development. Past a certain point however, the concavity changes, suggesting that access to mobile phones becomes less able to distinguish between countries' development. In this region, we have already remarked that they are mostly developed nations, who have a saturation in mobile phone access due to their higher average income. Likewise, we see in the middle panel, which corresponds to I_8 (secondary school enrollment) and I_{72} (recalling from Sect. 4.1 that this represents underdevelopment), that secondary school enrollment can initially also be used to characterise development. However after a certain level of development, secondary school enrollment saturates in these developed countries, meaning it can no longer be used in this way.

However, not all relationships between certain CDCIs are hockey-stick shaped. Indeed, we can see this from Fig. 6 which plots I_{18} vs I_{72} for 1998 on the left and 2016 on the right in the top panel and the same but for I_{20} vs I_{72} in the bottom panel. For the top panel, I_{18} is a CDCI that represents natural resource abundance, whilst again we recall from Sect. 4.1 that I_{72} corresponds to underdevelopment. We notice from the plots in the top panel that most of the countries with higher abundance of natural resources are underdeveloped countries. This reminds of the so called 'resource curse' [62], where resource-rich nations with inefficient governments are often underdeveloped.

Additionally in the bottom panel of Fig. 6, we plot I_{20} vs I_{72}. I_{20} corresponds to the amount of flow of foreign direct investment (FDI). Interestingly, we can observe an intriguing relationship between underdevelopment and FDI in the plots. All underdevel-

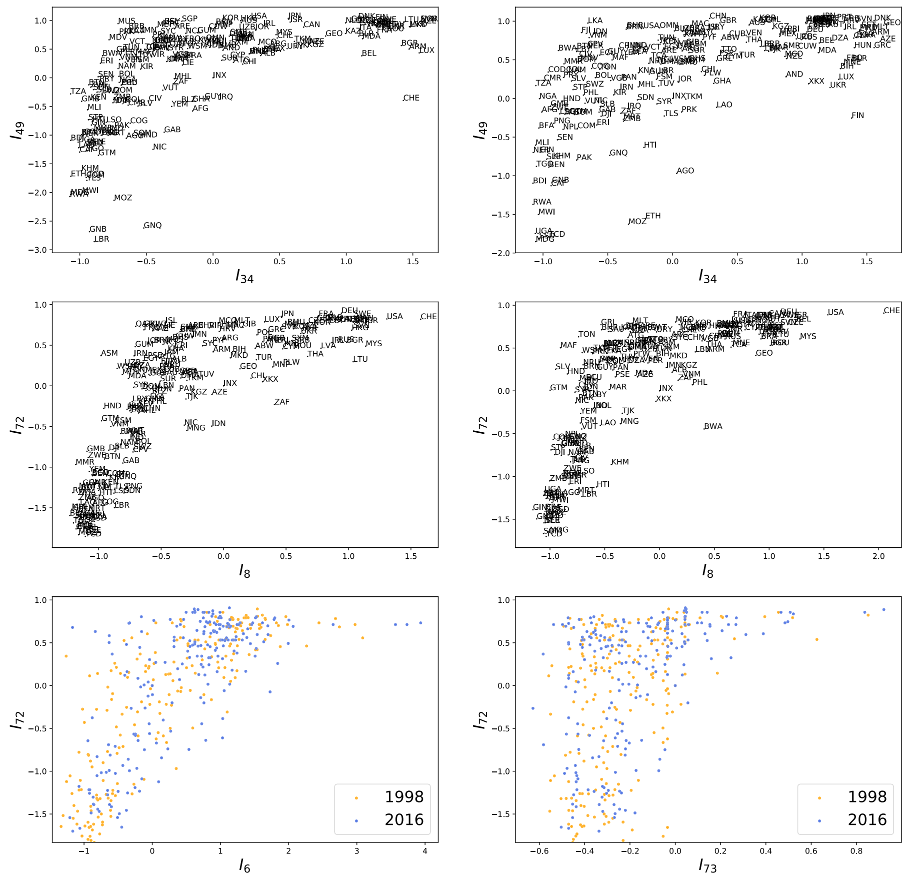


Figure 5 Comparison between different CDCIs with hockey stick shape. (Top panel) I_{34} vs I_{49} for 1998 on the left and 2016 on the right. (Middle panel) The same but with I_8 vs I_{72}. (Bottom panel) On the left we have I_6 vs I_{72} for 1998 in orange and 2016 in blue. On the right is the same but for I_{73} vs I_{72}

oped nations tend to have low FDI, which can be interpreted as being perceived with a low investment potential by foreign investors. However, countries which are not underdeveloped may have both high or low FDI, for example Poland, Kuwait and Uzbekistan are all more highly developed countries that have a low FDI. Attractiveness to foreign investments is not directly correlated to a country's level of developement.

7 Deriving influential indicators by using PMFG

One can imagine that there could be nodes that are very important to the structure of the correlation network than others, implying that these same nodes could be highly influential in the analysis of the development of countries. This would be very interesting for our purposes since they could provide a direct way to form a reliable reduced set of development indicators that would automatically overcome any problems associated with calculating composite versions. More specifically, one could take a subset of the top most influential indicators since these indicators' influence have an aggregate, over-arching influence on all other indicators, and thus the structure of interactions in the system.

In this section we shall apply the PMFG to **E** and identify system wide important indicators. For this purpose, information filtering is useful because it neatly transfers the problem of identifying influential indicators as finding a ranking of important nodes in the

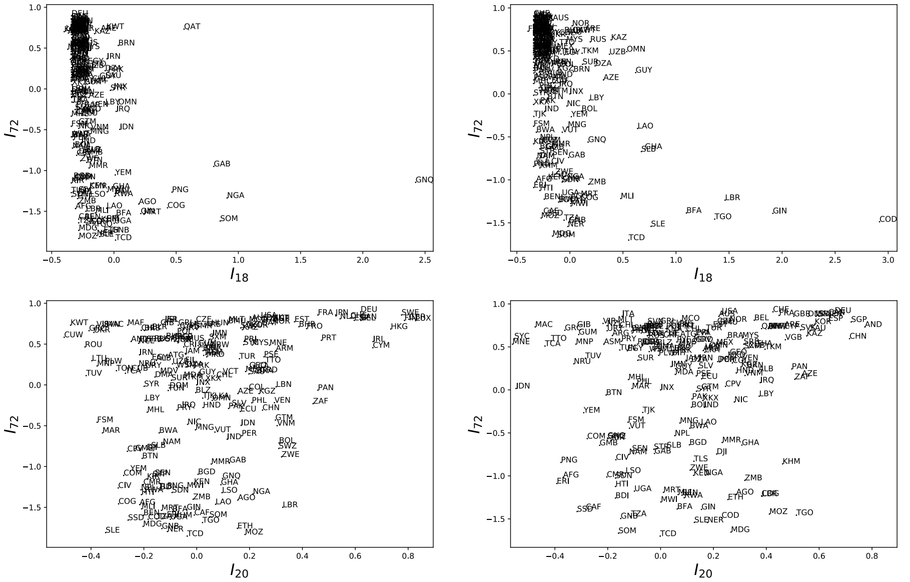


Figure 6 Comparison between particular CDCIs with no hockey-stick shape. (Top panel) I_{18} vs I_{72} for 1998 on the left and 2016 on the right. (Bottom panel) The same, but for I_{20} vs I_{72}

Table 2 The names of the top 9 influential indicators based on PageRank in the second column and their actual PageRank values in the first column

PageRank	Names
0.006764	Mobile cellular subscriptions
0.004666	Share of tariff lines with specific rates, manufactured products (%)
0.003717	Children in employment, wage workers, male (% of male children in employment, ages 7–14)
0.003706	Unemployment, male (% of male labor force) (national estimate)
0.003129	Mobile cellular subscriptions (per 100 people)
0.002939	Central government debt, total (% of GDP)
0.00276	Share of youth not in education, employment or training, female (% of female youth population)
0.002686	Population ages 30–34, female (% of female population)
0.002553	GDP (current US$)

network, for which there exist several so called network centrality measures. We choose to use PageRank [21], which has proven successful in ranking scientists and webpages [21, 63], to identify the most system wide influential indicators that affect the network. In PageRank, we rank nodes of networks on their importance based on the probability of a random walker landing on a particular node [21] with higher values indicating that the node has more importance.

We find the PMFG of **E**. The output network of this visualisation can be seen in the Supplementary Information Sect. 4. Then we apply PageRank to the PMFG of **E**, displaying an example of the top 9 indicators in Table 2.

7.1 Interpretation of the PageRank identified indicators

From Table 2 we observe that there are some indicators which we would expect to be in this ranking: for example GDP measures the value of goods and services an economy produces, and is widely used as a primary development indicator [64]. Central government debt has also been linked to economic development since high levels of debt can drag growth rates down [65]. However, it is interesting to see that mobile cellular subscriptions to be the top ranking indicator, especially considering our comments in Sect. 6 that mobile banking can be used to track the development of countries. This is an interesting result since there are many papers in computational socioeconomics which use mobile data as metric of an average citizen's socioeconomic status due to vast information it can encode [66–70]. In fact, it has been shown for example that mobile data is correlated with household expenditure [66] and poverty [71], and reveal gender inequality [68]. This may be because having a mobile cellular subscription requires a number of milestones in the development of a country e.g. a healthy enough population to make use of them, the education to know how to use them, the relevant infrastructure such as phone masts that can reach all parts of the population.

We have also investigated whether particular topics are overexpressed within the top 102 (chosen because this is the number of clusters identified by the DBHT) PageRank indicators by applying the same hypothesis test used in Sect. 4.2. We find that no topic is overexpressed within this subset of indicators, which again corroborates our conclusion that no single topic is more influential than the other.

8 Performance comparison

If we reduce all of the indicators in the dataset to the composite ones, we have boiled down the structure of the correlations between indicators to more essential constituents. Therefore, when the set of composite indicators are taken together, they should still be a faithful representation of the original **E** since they are main driving factors behind the structure of correlations. We can use this principle to evaluate the performance of the CDCIs against any alternatives. This section is dedicated to comparing the performance of the CDCIs derived in Sect. 5 against some alternatives.

For this purpose we propose, as a first approximation, that each indicator can be written as a linear factor model [72] of composite indicators. The general linear model is

$$\mathbf{X}_i = \sum_{k=1}^{K} \beta_{ik}\tilde{I}_k + \epsilon_i, \tag{6}$$

where $\mathbf{X}_i$ is the ith indicator i.e. the ith column $\mathbf{X}$. $\tilde{I}_k$ is the kth composite indicator of either the CDCIs or the other alternative schemes of composite indicators. β_{ik} is the loading of i for indicator k, which measures the sensitivity of $\mathbf{X}_i$ to changes in $\tilde{I}_k$. Finally, ϵ_i are white noise terms. Equation (6) is an appropriate approximation to use since firstly, we are using the linear correlation matrix, which means it is intimately related to linear factor models. Note we also have that the number of composite indicators in each of the alternatives used in our comparison must be the same as the number of CDCIs K. This is because the size of the indicator set will inevitably affect its ability to describe the correlations, so fair comparison must involve fixing the number of indicators used. We then use elastic net regression (for details see Supplementary Information Sect. 5), which is able to take into consideration the potential correlation between composite indicators, to find β_{ik} and $\beta_{ik'}$ for every i. The performance can then be evaluated on the basis of the error between the linear model and the real indicator values. For this, we define the usual mean squared

Table 3 In the first column, the *ERR* calculated using Eq. (8) for the benchmark using 102 random subsets of indicators, which is repeated 100 times. The second column is the same but instead using 102 of the most influential indicators, assessed via PageRank

Random	PageRank
0.66	0.71

error of the regression as

$$MSE = \sum_{i=1}^{N} \left(\mathbf{X}_i^{(\text{predict})} - \mathbf{X}_i\right)^2, \tag{7}$$

where $\mathbf{X}_i^{(\text{predict})}$ are the predicted values of $\mathbf{X}_i$ using the β_{ik} from the elastic regression. The final metric we use the evaluate the performance of the cluster driven composite indicators is

$$ERR = \frac{MSE_{\text{CDCIs}}}{MSE_{\text{alt}}}, \tag{8}$$

where *ERR* is called the error reduction ratio, MSE_{CDCIs} is the *MSE* calculated in Eq. (7) for the CDCIs. Similarly, MSE_{alt} is the same but for any of the alternative schemes of composite indicators used as a comparison. If *ERR* is below 1 (above 1) then this means that the CDCIs perform better (worse). Also note that of course *ERR* is bounded below by 0.

The choice of alternative schemes of composite indicator, is as follows. We take the top K PageRank indicators that were identified in Sect. 7.1. We choose these particular schemes since they offer the best feasible alternative in forming a basis of composite indicators that have the most influence on correlations system wide. A fourth comparison is also made by randomly selecting 102 indicators from the columns of $\mathbf{X}$ that provides a benchmark of the performance of the other composite indicator schemes since a set of randomly selected indicators should not be able to reliably incorporate anything from the real correlation network.

We carry out the elastic regression and compute *ERR* for all alternative schemes of indicators used. For the random benchmark, we repeat and average the results for 100 different random subsets of indicators. The results are shown in Table 3. We see that in both cases *ERR* is much less than 1, indicating that the CDCIs are able to outperform the random benchmark and the PageRank alternative. We can therefore conclude that the CDCIs are more effective at reducing the dimensionality of the dataset the random benchmark and the PageRank alternative.

9 Dynamical analysis

Since in the analysis so far we have used the static correlation matrix computed over the whole time period, we should also investigate the dynamic stability of the clusters. We start by splitting the whole time period into 16 rolling time window of length 4, with a time shift of 1 year. For each time window $w = 1, \ldots, 16$, we calculate the corresponding correlation matrix $\mathbf{E}^w$ and its DBHT clustering. The similarity between each pair of time windows w and w' is then measured by calculating the ARI between their respective DBHT clusterings. The results are shown in the heat map of Fig. 7(left). We can see that overall the DBHT clusterings display a high similarity with each other, with a median ARI value

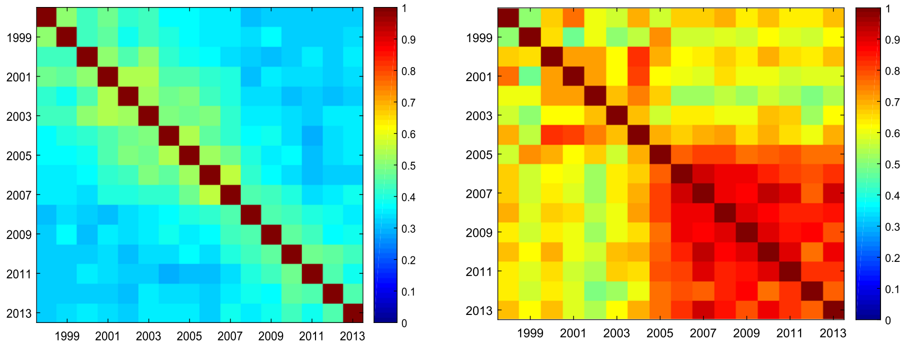


Figure 7 Left panel is a heat map of the ARI computed between the DBHT clusterings of indicators for each pair of 16 rolling time windows of length 4, with a shift of 1 year. Right panel is the heat map of the ARI computed between the DBHT clusterings of the CDCIs, using the same parameters for the rolling window as left panel. The colour legend for both plots is on the right

of 0.376, which is high considering that the static clustering does not reflect the topics as argued through the ARI computed in Sect. 4.2. This also confirms to some extent our choice for the length of the sliding window since we see that the clusters are stable through time.

We also used the same procedure and parameters to investigate the dynamic stability of the relationship between CDCIs (except using the correlation matrix and DBHT clustering between the CDCIs). A heat map of the results can be seen in Fig. 7(right). Again, we see that overall there is a high likeness between the clusterings of the CDCIs in each time window. In fact, the median ARI is even higher at 0.683. Interestingly, we see that starting from the window covering 2005 to 2009, which is the year corresponding to the financial crisis, there is a markedly higher similarity between the clusterings of the CDCIs. This could be explored further.

10 Conclusion

In this paper, we have investigated whether the collection of development indicators given by the WDI database can be divided using their fundamental topic description. Leveraging on PCA and a novel application of information filtering and hierarchical clustering techniques, we showed that the structure of the topics does not mirror the actual structure between the indicators. This suggests that composite development indicators that are aggregated from restricted sets may ignore key information. Instead, we propose a new set of cluster driven composite development indicators that overcomes these problems. They are objective, data driven, interpretable and are able to make valid comparisons between countries. We have used the composite indicators and some highly influential PageRank indicators to give new insights into the development of countries. Some of these may support decisions for policy makers. Lastly, we showed that our proposed composite indicators can outperform schemes of indicators based on a random benchmark and PageRank. We mention that it has been pointed out by [26] that using the correlation matrix to form composite indicators may ignore the presence of causality relationships. We mainly use the CDCIs to group countries to understand how they can be classified in terms of their development so that there is no implicit reliance on there being a different causal relationship between indicators. It would, however, be interesting to develop combining the methodology proposed here with an analysis of the causal relationships in a future work.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1140/epjds/s13688-020-00225-y.

Additional file 1. This file contains all details of calculations which have been referenced in the text as Supplementary Information (PDF 611 kB)

Acknowledgements
We acknowledge the input of Giuseppe Brandi in this article. We thank the two anonymous reviewers, whose contributions have greatly improved the manuscript.

Funding
AV wishes to thank EPSRC for providing funding during his PhD studies under grant EP/M50788X/1. We also wish to thank the ESRC Network Plus project 'Rebuilding macroeconomics'. We acknowledge support from Economic and Physical Science Research Council (EPSRC) grant EP/P031730/1. We are grateful to the NVIDIA corporation for supporting our research in this area with the donation of a GPU.

Availability of data and materials
The WDI dataset analysed during the current study is available at the following link
https://datacatalog.worldbank.org/dataset/world-development-indicators.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
AV, OA and TDM conceived the experiment(s), AV and OA conducted the experiment(s) and AV, OA and TDM authors analysed the results. AV, OA and TDM reviewed the manuscript. All authors read and approved the final manuscript.

Author details
[1]Department of Mathematics, King's College London, London, UK. [2]Department of Computer Science, University College London, London, UK. [3]Complexity Science Hub Vienna, Vienna, Austria.

Publisher's Note

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

Received: 3 December 2019 Accepted: 19 March 2020 Published online: 10 April 2020

References

1. Stock JH, Watson MW (1989) New indexes of coincident and leading economic indicators. NBER Macroecon Annu 4:351–394
2. Mügge D (2016) Studying macroeconomic indicators as powerful ideas. J Eur Public Policy 23(3):410–427
3. Ricardo D (1891) Principles of political economy and taxation. G. Bell, London
4. Leontief W (1956) Factor proportions and the structure of American trade: further theoretical and empirical analysis. Rev Econ Stat 38(4):386–407
5. Bowen HP, Leamer EE, Sveikauskas L (1986) Multicountry, multifactor tests of the factor abundance theory. Working paper 1918, National Bureau of Economic Research
6. Aghion P, Howitt P (1990) A model of growth through creative destruction. Technical report, National Bureau of Economic Research
7. Heckscher EF, Ohlin BG (1991) Heckscher–Ohlin trade theory. MIT Press, Cambridge
8. Kremer M (1993) The O-ring theory of economic development. Q J Econ 108(3):551–575
9. Krueger AB, Lindahl M (2001) Education for growth: why and for whom? J Econ Lit 39(4):1101–1136
10. Egert B, Kozluk TJ, Sutherland D (2009) Infrastructure and growth: empirical evidence. CESifo working paper series
11. Aghion P, Howitt P, Murtin F (2010) The relationship between health and growth: when Lucas meets Nelson–Phelps. Technical report, National Bureau of Economic Research
12. UNDP (1997) Ghana human development report. United Nations Development Programme, Accra
13. Salzman J (2003) Methodological choices encountered in the construction of composite indices of economic and social well-being. Centre for the Study of Living Standards, Ottawa
14. Sagar AD, Najam A (1998) The human development index: a critical review. Ecol Econ 25(3):249–264
15. Todaro MP, Smith SC (2015) Economic development. Pearson, Upper Saddle River
16. Huawei (2018) Global connectivity index 2018
17. Bray F, Jemal A, Grey N, Ferlay J, Forman D (2012) Global cancer transitions according to the Human Development Index (2008–2030): a population-based study. Lancet Oncol 13(8):790–801
18. Huggins R (2003) Creating a UK competitiveness index: regional and local benchmarking. Reg Stud 37(1):89–96
19. Van Der Maaten L, Postma E, Van den Herik J (2009) Dimensionality reduction: a comparative. J Mach Learn Res 10:66–71
20. Bun J, Bouchaud J-P, Potters M (2017) Cleaning large correlation matrices: tools from random matrix theory. Phys Rep 666:1–109
21. Page L, Brin S, Motwani R, Winograd T (1999) The PageRank citation ranking: bringing order to the web. Technical report, Stanford InfoLab

22. Cristelli M, Tacchella A, Cader M (2018) The virtuous interplay of infrastructure development and the complexity of nations. Entropy 20(10):761
23. Lai D (2003) Principal component analysis on human development indicators of China. Soc Indic Res 61(3):319–330
24. Nardo M, Saisana M, Saltelli A, Tarantola S (2005) Tools for composite indicators building. EUR 21682 EN, European Commission, Institute for the Protection and Security of the Citizen, JRC Ispra, Italy
25. Castellacci F (2011) Closing the technology gap? Rev Dev Econ 15(1):180–197
26. Mazziotta M, Pareto A (2019) Use and misuse of PCA for measuring well-being. Soc Indic Res 142(2):451–476
27. Mantegna RN (1999) Hierarchical structure in financial markets. Eur Phys J B, Condens Matter Complex Syst 11(1):193–197
28. Tumminello M, Aste T, Di Matteo T, Mantegna RN (2005) A tool for filtering information in complex systems. Proc Natl Acad Sci USA 102(30):10421–10426
29. Anderberg MR (2014) Cluster analysis for applications. Probability and mathematical statistics: a series of monographs and textbooks, vol 19. Academic Press, Cambridge
30. Song W-M, Di Matteo T, Aste T (2012) Hierarchical information clustering by means of topologically embedded graphs. PLoS ONE 7(3):e31929
31. Musmeci N, Aste T, Di Matteo T (2015) Relation between financial market structure and the real economy: comparison between clustering methods. PLoS ONE 10(3):e0116201
32. Sneath PH (1957) The application of computers to taxonomy. Microbiology 17(1):201–226
33. Graham RL, Hell P (1985) On the history of the minimum spanning tree problem. Ann Hist Comput 7(1):43–57
34. Aste T, Di Matteo T, Hyde ST (2005) Complex networks on hyperbolic surfaces. Phys A, Stat Mech Appl 346(1–2):20–26
35. Musmeci N, Aste T, Di Matteo T (2015) Risk diversification: a study of persistence with a filtered correlation-network approach. J Netw Theory Finance 1(1):77–98
36. WBIEDDD Group (2018) World development indicators. World Bank, Washington
37. Jolliffe I (2002) Principal component analysis. Wiley, Hoboken
38. Plerou V, Gopikrishnan P, Rosenow B, Amaral LAN, Guhr T, Stanley HE (2002) Random matrix approach to cross correlations in financial data. Phys Rev E 65(6):066126
39. Stein SAM, Loccisano AE, Firestine SM, Evanseck JD (2006) Principal components analysis: a review of its application on molecular dynamics data. Annu Rep Comput Chem 2:233–261
40. Marčenko VA, Pastur LA (1967) Distribution of eigenvalues for some sets of random matrices. Sb Math 1(4):457–483
41. Mishra SK (2008) On construction of robust composite indices by linear aggregation. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1147964
42. Bishop CM (2006) Pattern recognition and machine learning. Springer, Berlin
43. Ravasz E, Barabási A-L (2003) Hierarchical organization in complex networks. Phys Rev E 67(2):026112
44. Corominas-Murtra B, Goñi J, Solé RV, Rodríguez-Caso C (2013) On the origins of hierarchy in complex networks. Proc Natl Acad Sci USA 110(33):13316–13321
45. Jain AK, Murty MN, Flynn PJ (1999) Data clustering: a review. ACM Comput Surv 31(3):264–323
46. Wang H, Wang W, Yang J, Yu PS (2002) Clustering by pattern similarity in large data sets. In: Proceedings of the 2002 ACM SIGMOD international conference on management of data, pp 394–405
47. Mantegna RN, Stanley HE (1999) Introduction to econophysics: correlations and complexity in finance. Cambridge University Press, Cambridge
48. Winkler H, Simões AF, La Rovere EL, Alam M, Rahman A, Mwakasonda S (2011) Access and affordability of electricity in developing countries. World Dev 39(6):1037–1050
49. Garcia-Moreno C, Jansen HA, Ellsberg M, Heise L, Watts CH et al (2006) Prevalence of intimate partner violence: findings from the WHO multi-country study on women's health and domestic violence. Lancet 368(9543):1260–1269
50. Smith LC, Haddad LJ (2000) Explaining child malnutrition in developing countries: a cross-country analysis. FCND discussion paper 60, International Food Policy Research Institute
51. Ravallion M (1997) Can high-inequality developing countries escape absolute poverty? Econ Lett 56(1):51–57
52. Bose N, Haque ME, Osborn DR (2007) Public expenditure and economic growth: a disaggregated analysis for developing countries. Manch Sch 75(5):533–556
53. Gupta GR, Parkhurst JO, Ogden JA, Aggleton P, Mahal A (2008) Structural approaches to HIV prevention. Lancet 372(9640):764–775
54. Montgomery MA, Elimelech M (2007) Water and sanitation in developing countries: including health in the equation. Environ Sci Technol 41(1): 17–24
55. Romer PM (1990) Endogenous technological change. J Polit Econ 98(5, Part 2):S71–S102
56. Rand WM (1971) Objective criteria for the evaluation of clustering methods. J Am Stat Assoc 66(336):846–850
57. Tumminello M, Micciche S, Lillo F, Piilo J, Mantegna RN (2011) Statistically validated networks in bipartite complex systems. PLoS ONE 6(3):e17994
58. Feller W (2008) An introduction to probability theory and its applications, vol 2. Wiley, Hoboken
59. Somarriba N, Pena B (2009) Synthetic indicators of quality of life in Europe. Soc Indic Res 94(1):115–133
60. Pugliese E, Chiarotti GL, Zaccaria A, Pietronero L (2017) Complex economies have a lateral escape from the poverty trap. PLoS ONE 12(1):e0168540
61. Keller KR (2006) Investment in primary, secondary, and higher education and the effects on economic growth. Contemp Econ Policy 24(1):18–34
62. Ross ML (1999) The political economy of the resource curse. World Polit 51(2):297–322
63. Liu X, Bollen J, Nelson ML, Van de Sompel H (2005) Co-authorship networks in the digital library research community. Inf Process Manag 41(6):1462–1480
64. Lepenies P (2016) The power of a single number: a political history of GDP. Columbia University Press, New York
65. Checherita-Westphal C, Rother P (2012) The impact of high government debt on economic growth and its channels: an empirical investigation for the euro area. Eur Econ Rev 56(7):1392–1405
66. Blumenstock J, Shen Y, Eagle N (2010) A method for estimating the relationship between phone use and wealth. In: QualMeetsQuant workshop at the 4th international conference on information and communication technologies and development, vol 13, pp 114–125

67. Blumenstock JE, Eagle N (2012) Divided we call: disparities in access and use of mobile phones in Rwanda. Inf Technol Int Dev 8(2):1–16
68. Mehrotra A, Nguyen A, Blumenstock J, Mohan V (2012) Differences in phone use between men and women: quantitative evidence from Rwanda. In: Proceedings of the fifth international conference on information and communication technologies and development, pp 297–306.
69. Gutierrez T, Krings G, Blondel VD (2013) Evaluating socio-economic state of a country analyzing airtime credit and mobile phone datasets. Preprint. arXiv:1309.4496
70. Gao J, Zhang Y-C, Zhou T (2019) Computational socioeconomics. Preprint. arXiv:1905.06166
71. Smith C, Mashhadi A, Capra L (2013) Ubiquitous sensing for mapping poverty in developing countries. Paper submitted to the Orange D4D Challenge
72. Thompson B (2004) Exploratory and confirmatory factor analysis: understanding concepts and applications. American Psychological Association, Washington