King's College London
Department of Mathematics
Submission Cover Sheet for Coursework

The following cover sheet must be completed and submitted with any dissertation, project or essay submitted as a part of formal assessment for degree within the Mathematics Department.

You are not required to write your name on your work

Candidate Number:	AE12467
Module Code:	7CCMFM17
Title of Project:	

Declaration

By submitting this assignment I agree to the following statements:

I have read and understand the King's College London Academic Honesty and Integrity Statement that I signed upon entry to this programme of study.

I declare that the content of this submission is my own work.

I understand that plagiarism is a serious examination offence, an allegation of which can result in action being taken under the College's Misconduct regulations.

Your work may be used as an example of good practice for future students to refer to. If chosen, your work will be made available either via KEATS or by paper copy. Your work will remain anonymous; neither the specific mark nor any individual feedback will be shared. Participation is entirely optional and will not affect your mark. If you consent to your submission being used in this way please tick the box.	

Office Use Only: Date/Time Stamp	Agreed Mark

Exploring the CDCI Approach for Composite Development Indicators

K23110623

July 8, 2024

Abstract

Approaches to creating composite development indicators are vast and full of criticism [1]. Cluster Driven Composite development Indicators (CDCIs) give a new approach to composite development indicators to resolving the subjective problem of applying supervised clustering methods to the clustering problem of development indicators [2]. CDCIs are data-driven, non-weighted and can be used to track development for different countries which makes them a good candidate for composite development indicators. This paper will explain the methodology and analysis of CDCIs presented in [2] as well as compare its methodology with others in the literature. We discuss the trade-offs made by the approach of CDCIs to be data-driven, non-weighted and robust to outliers. Explaining also how the DBHT algorithm used in CDCIs excels in dimensionality reduction [2] whilst being an unsupervised clustering technique which many of the techniques used in the field are not [1]. Additionally, as part of a personal contribution, further research into overlapping clustering is suggested. With the argument that in the case of creating composite development indicators, it may not be beneficial to assume uniqueness of clustering, highlighting areas in the literature where relaxing this condition has been shown to be helpful and proposing where this condition could be relaxed in the DBHT algorithm.

July 8, 2024
King's College London
M.Sc. in Financial Mathematics-Econophysics module FM17

1 Introduction

Development indicators aim to measure a specific area of a country's development [3]. Development indicators are used by many, such as economists or policymakers for a number of reasons [2] consisting of, but are not limited to: describing trends, development planning and target setting, finding interdependent development relationships and diagnosing particular development situations [3].

Creating a development indicator for specified developmental domains poses varying levels of difficulty, mainly dependent on whether the targeted domain is measurable or non-measurable [1]. In instances where development in certain areas is non-measurable such as well-being [1], corresponding indicators often rely on indirect metrics to gauge progress [3]. It is a common belief that "no single yardstick exists to measure development just as no single set of objectives can describe adequately the diversity of development conditions in the world" [4]. Thus many are left wondering if by combining multiple indirect development metrics we may better represent a non-measurable target domain. Composite development indicators are one such solution to this problem [5], aiming to represent and summarise new areas of development by aggregating existing development indicators together.

The development of composite indicators in essence comes down to dimensionality reduction to identify which indicators to aggregate together and choosing an aggregation method that best suits the indicators in question [1]. There is no undisputed method for solving either of these problems which has led to a large number of proposed methods for creating development indicators in the literature. However many of the current methods proposed suffer from high levels of subjectivity [1]. Common areas for which subjectivity occurs include reducing dimension based on non-data driven techniques [1], using data-driven non-parametric methods to reduce dimension [2] and using weighted methods of aggregation [1].

In this paper we will review the paper "A new set of cluster driven composite development indicators" [2] by Anshil Verma, Irazio Angelini and Tiziana Di Matteo, who propose a data-driven clustering method of aggregating development indicators that aims to answer some of the problems with subjectivity.

The remaining sections of this paper will cover the following: Section 2 will give a brief summary of the paper [2] highlighting core results. Section 3 aims to explain further the methodology used to create and analyse the proposed Cluster Driven Composite development Indicators (CDCIs) as well as provide an interpretation for each analysis. Section 4 compares the CDCI methodology with other techniques in the literature outlining the trade-offs of each, whilst Section 5 seeks to extend the CDCIs for further research. Finally Section 6 will conclude this paper by offering an overview of the points discussed throughout.

2 Summary

"A new set of cluster-driven composite development indicators" gives readers insight into a new method of creating composite development indicators, that is data-driven and unsupervised [2].

The paper [2] proves that aggregating indicators based on their fundamental categories such as infrastructure, economic and environment (see Table 1 for the

full list) is far from the best method of dimension reduction. In fact, it is shown that the hidden structure between development indicators has clustering containing development indicators from different fundamental categories. Therefore subjectively choosing indicators to aggregate based on these categories should not be done to avoid missing key information [2].

A natural progression is to suggest a data-driven method of dimensionality reduction of which Principle Component Analysis (PCA) is a common tool [6]. One notable drawback of this method is that PCA does not reduce the dimension of the data unless a subset of the principle components is taken. Choosing a number of principle comments to represent the data is subjective by nature and therefore leaves room for criticism[6]. Fortunately, it is also found in this paper, by fitting a Marcendko-Pasture (MP) distribution to the eigenvalue distribution obtained in the PCA and to shuffled data, that choosing only a subset of the principle components is likely to disregard relevant information [2].

The authors of [2] introduce Cluster Driven Composite development Indicators (CDCI) as the groundwork for a new set of development indicators derived from the DBHT algorithm. The DBHT is an information filtering technique based on graph theory that benefits from being an effective clustering tool for finding hidden structures in networks that are data-driven, unsupervised and capable of forming hierarchical clustering [7]. Local analysis of the DBHT clustering results shows a level of interoperability in each cluster. In other words, we can somewhat guess the aspect of development that indicators in these clusters share [2].

The CDCIs are formed using a non-weighted median to aggregate the development indicators found in each DBHT cluster [2]. This aggregation technique is less sensitive to anomalies, can be applied consistently across indicators of different sizes and is not inherently subjective like many other aggregation methods [2]. The CDCIs are then analysed using a pairwise comparison, performance comparison and dynamic analysis.

Dynamic analysis reveals that the clusters formed in the DBHT have high similarity with each other when computed over a rolling time window [2]. An important robustness result, since comparing CDCI for different points in time would not be viable otherwise. The performance of CDCIs is measured by attempting to reconstruct the information presented in the original correlation matrix from which the CDCIs are derived. Other sets of composite indicators can also be used to try and reconstruct the same information to provide a comparison. A random and PageRank are used as comparison benchmarks where it is found that the CDCIs are a better dimension reduction method for the data [2]. CDCIs can also be used to track development between countries and over different periods [2]. This fact is highlighted by the pairwise comparison where insightful information on the relationship between two areas of development can be extracted.

3 Method

3.1 Data

As with any analysis, one starts with the data. The author in [2] begin their exploration with a reduced version of the World Development Indicator (WDI)

dataset, where no pairwise development indicator in the dataset has a Pearson correlation coefficient greater than 0.95. This dataset, denoted as $\mathbf{X}$, contains $N = 1448$ different development indicator for $C = 218$ countries over a $T = 19$ year period spanning from 1998 to 2016 [2].

The dataset is re-structured with the goal of removing spurious correlations [2]. The first difference is therefore taken to remove any trend-related correlations by applying

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

to X, where $\mathbf{X}(t, c, i)$ denotes the value of the indicator i at time t for country c [2]. The blocks of data $\Delta\mathbf{X}(\tilde{t}, c, i)$ are then stacked together to create the matrix ΔX that contains all the difference indicator values for all counties c and time steps $\tilde{t}$. Another benefit of using the matrix $\Delta\mathbf{X}$ for analysis rather than its predecessor $\mathbf{X}(t)$ is that $\Delta\mathbf{X}$ is more dense and consequently less sensitive to noise intern reducing spurious correlations related to randomness [2].

Finally the Pearson correlation matrix E is calculated [2] for $\Delta\mathbf{X}$ using

$$E = \frac{1}{C(T-1)}(\Delta\mathbf{X})^\top \Delta\mathbf{X}$$

The entries in matrix E represent the correlations between a pair of development indicators for the N indicators in $\Delta\mathbf{X}$.

Additionally, each of the N development indicators are also sorted into 12 different categories [2] which will be referred to as the set of fundamental categories, as shown in Table 1.

Economic	Environment	Gender	Infrastructure	Private	Social
Education	Financial	Health	Poverty	Public	World

Table 1: Table of Fundamental Categories for Development Indicator

3.2 Principle Component Analysis (PCA)

When aggregating development indicators, a common approach in the literature is to gravitate towards grouping indicators based on categories similar to that found in Table 1 [2]. One way of checking the validity of this method of aggregation is to perform a Principle Component Analysis (PCA) [2]. Since the PCA conducted in [2] does not directly contribute to the development of CDCIs and serves more as a motivation, we will briefly touch upon the main point of this analysis.

PCA is a method of dimensionality reduction [8]. Explained simply, PCA transforms the data E onto a new set of basis also known as the principle components. The principle components must maximise the variability of the data along each of the components whilst remaining orthogonal to one another and centred at the mean of the data. The principle components are ranked from 1 to N with the first explaining the most variability and the last explaining the least. Users of PCA choose a subset of the principle components i.e. the top m principle components to explain the data doing so reduces the dimension of the original dataset [8].

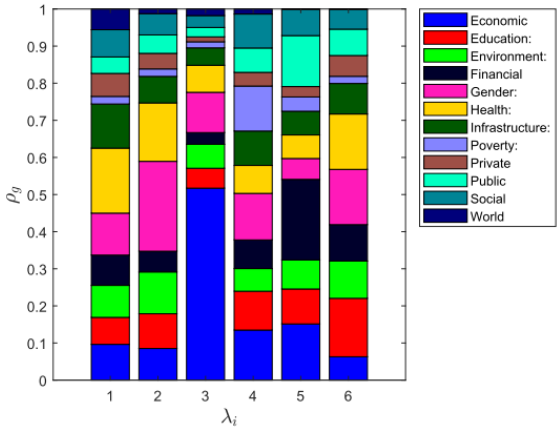

Figure 1: Bar chart displaying the ρ_g values, where ρ_g is defined as the contribution of the gth topics to the ith eigenvector in the PCA, for the top 6 principal components λ_i of E [2]. These topic are the 12 fundamental categories as displayed in the legend [2]. Image Sourced from [2].

It is possible after performing a PCA to check the composition of each principle components in terms of the topic classifications [2] as shown in Figure 1. The main observation from Figure 1 is that none of the top 6 principal components are represented by a single fundamental category and instead are comprised of multiple categories [2]. This result suggests that a composite development indicator based on topics alone may not fully capture all the relevant information [2]. In other words, composite development indicators need to be aggregated not by topic classifiers but rather by some other dependency measures that partition the data better.

Following the results of the PCA and by recalling again the points made in Section 1, we can see a clear need for a method of combining development indicators that does not fall victim to the problems that most common composite indicators fail to address. The rest of [2] aims to answer this calling by means of using DBHT to reduce the dimension of the correlation matrix E and identify a new set of dependency measures (clusters) that are consistent with the data and avoid the need for educated guesses and human interaction. These clusters can then be used to form composite development indicators that are free of the constraints of typical composite development indicators. The proposed Directed Bubble Hierarchical Tree (DBHT) provides an unsupervised method of dimensionality reduction that is data-driven [2].

3.3 Planar Maximum Filter Graph (PMFG)

To explain DBHT we must first describe the Planar Maximum Filter Graph (PMFG) for which the DBHT is derived in [2]. The PMFG is a network filtering technique that can be applied to a similarity matrix in order to strip away irrelevant information while also maximise the quantity of information retained in the resulting graph [9]. The PMFG is similar to the Minimal Spanning Tree (MST) and in fact, contains the MST but unlike the MST does not remove as

much information because of its planar structure.

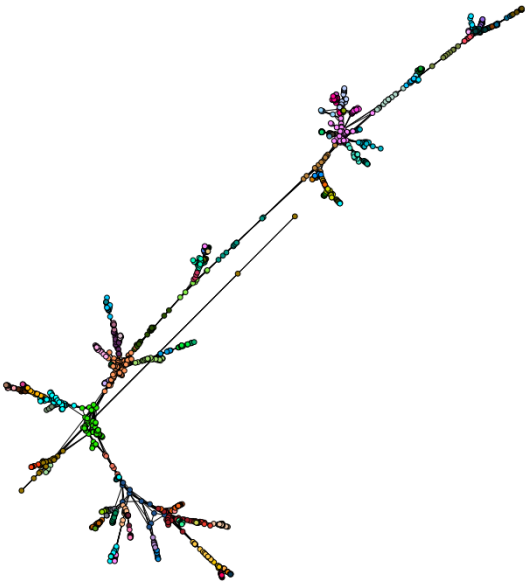

Figure 2: The PMFG $G(v,\, e,\, w,\, d)$ of E. Each node v_i is coloured depending on its cluster assignment according to DBHT algorithm. Image sourced from the supplementary documentation of [2].

The PMFG starts with a similarity matrix, E, and a genus g set to zero ($g = 0$) as done in [2] for which the filter graph will be embedded onto. The similarity matrix E is then paired with a distance matrix D defined as

$$D_{i,\,j} = \sqrt{(2(1 - E_{i,\,j})}$$

where $D_{i,\,j}$ measures the distance between two indicators in terms of there correlation [2]. Elements in the lower diagonal of D excluding the diagonal are then ranked from lowest down to highest to form a list L of distances between pairwise development indicators. Each node (vertex v_i) in the PMFG will represent a specific indicator value and the connections between the nodes (edge $e_{i,j}$) will hold information regarding the correlation $w_{i,j} = E_{i,j}$ and distance $d_{i,j} = D_{i,j}$ between the two nodes v_i and v_j [9].

The PMFG is then constructed using the following algorithm, start at the top of the list L and plot the edge $e_{i,j}$ between the two respective nodes v_i and v_j for which the distance $d_{i,j}$ are measured for, if and only if the resulting graph remains a planner graph [9] i.e. no edges on the graph cross one another. Work down the list disregarding any edges that would make the graph non-planner [9]. Once all edges have be plotted we have created a PMFG for E (see Figure 2) which we denote as $G(v,\, e,\, w,\, d)$.

3.4 Directed Bubble Hierarchical Tree (DBHT)

The Directed Bubble Hierarchical Tree (DBHT) algorithm is applied to previously formed PMFG $G(v,\, e,\, w,\, d)$ to identify data dependent clustering. The

DBHT takes advantage of the topological 3-clique property of the PMFG [7]. In graph theory a cycle is defined as a path that connects a series of vertices such that the start and end vertices of the path are the same [10]. The simplest cycle is a 3-clique which contains only three vertices [7]. Cycles can either be separating or non-separating cycles in a Planar graph [11], one can consider a separating 3-clique as a cycle that can be removed from the original planar graph without breaking its shape. As described a planner graph can be divided into two disconnected sub-planar graphs by a separating 3-clique which we denote as the exterior G_p^{ex} and interior G_p^{ex} sub-planar graphs [7]. The exterior and interior sub-planar graphs are consequently assigned an edge to connect them. One repeatedly splits $G(v,\, e,\, w,\, d)$ as well as the there after sub-planner graphs G_n by there separating 3-cliques until the resulting series of sub-planar graphs known as bubbles b_i are comprised of only non-separating 3-clique [7]. The Graph created by this process is known as a undirected bubble tree H.

An edge direction is then established between connected bubbles by comparing the sum over the weights of the edges in the PMFG connecting the 3-clique k_p with the interior G_p^{in} and exterior G_p^{ex} sub-graphs [7]. Using both

$$W_p^{in} = \sum_{v \in k_p,\, u \in G_p^{in}} A_G(v,u) \qquad\qquad W_p^{ex} = \sum_{v \in k_p,\, u \in G_p^{ex}} A_G(v,u)$$

where $A_G(v,u) = w_{u,v}$, to compare W_p^{in} and W_p^{ex} directing the edge towards the bubble with the highest value creating a directed bubble tree $\vec{H}$ [7]. Note there are three types of bubbles in our graph: Converging bubbles, where all connected edges are directed into the bubble, diverging bubbles where all connected edges are directed away from the bubble and passage bubbles that have both edges directed in and out of the bubble [11]. Once a directed bubble tree is formed we can identify non-discrete clusters $(\alpha, \beta, ...)$ by following the path of bubbles from a diverging bubble down to a converging bubble $(b_\alpha, b_\beta, ...)$ [7].

The next step in the DBHT algorithm is to form discrete clusters by allocating each vertex v_i to a unique cluster [11]. For the vertices that are contained in more than one converging bubble e.g. b_α and b_β, we consider the strength of attachment

$$\chi(v,\, b_\alpha) = \frac{\sum_{u \in V(b_\alpha)} A_G(v,\, u)}{3(|V(b_\alpha)| - 2)}$$

where $|V(b_\alpha)|$ represents the number of vertices in the bubble b_α [7]. Assigning each vertex to the bubble with the largest strength such that each converging bubble now has a new unique set of vertices $(V^0(\alpha), V^0(\beta), ...)$. For remaining vertices contained in more than one non-discrete cluster $(\alpha, \beta, ...)$ we consider the minimum mean average shortest path distance

$$\bar{L}(v,\, \alpha) = \text{mean}\{l(u,\, v) \,|\, u \in V^0(\alpha) \cap v \in V(\vec{h_\alpha})\}$$

with respect to all other converging bubbles, where $\vec{h}_\alpha$ is the path (sub-tree) defining the cluster α and $l(v,u)$ is the the smallest sum of distances $d_{r,s}$ over any path between v and u [7]. Once each vertices has been assigned to a unique cluster we have achieved discrete clustering.

Finally to introduce the complete hierarchical structure of the DBHT we must now look at how the clusters are internally structured and how different

clusters cluster together [7]. We use the tailored linkage procedure to achieve intra-bubble hierarchy, intra-cluster hierarchy and inter-cluster hierarchy.

Imagine that the vertices in a cluster, say α, were assigned again to there respective bubbles again b_i in the sub-tree $\vec{h}_\alpha$ where the vertices in the converging bubbles have been assigned to the set $V^0(\alpha)$ [7]. Then by using the strength of attachment measure to we can achieve a discrete assignment of each vertex to a bubble in $\vec{h}_\alpha$ resulting in each bubble having a unique set of vertices unique to α which we call $V^\alpha(b_i)$ and define as Intra-bubble hierarchy. $(V^\alpha(b_i), V^\beta(b_i), ...)$ can be used to perform the complete linkage procedure [7].

Intra-cluster hierarchy is created by performing a complete linkage procedure between bubbles in $(\vec{H}_\alpha, \vec{H}_\beta, ...)$ using the distance matrix

$$d^I_\alpha(b_i,\, b_j) = \max\{l(u,\, v) \,|\, u \in V^\alpha(b_i) \cap v \in V^\alpha(b_j)\}$$

[7]. Similarly inter-cluster hierarchy is achieved using a complete linkage procedure but between clusters $(\alpha, \beta, ...)$ by the distance matrix

$$d^{II}(\alpha, \beta) = \max\{l(u,\, v) \,|\, u \in V(\alpha) \cap v \in V(\beta)\}$$

[7]. The result of the complete linkage procedure will create a hierarchical dendrogram that can also be used for analysis

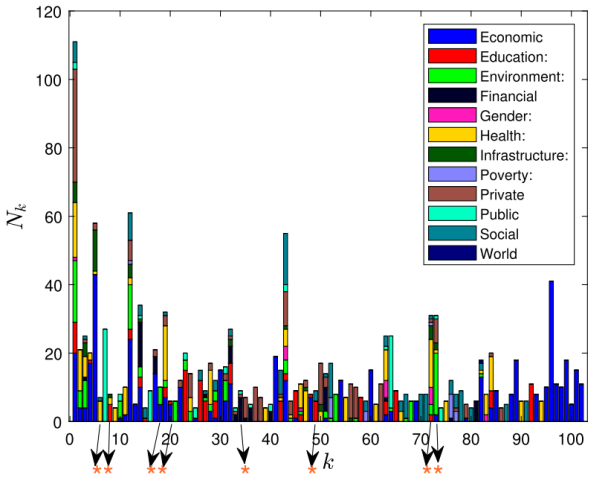

Figure 3: A bar chart depcting the composition of DBHT cluster with respect to the fundamental categories displayed on the legend. Each bar correspond to a cluster cluster k with its height N_k depicting the number of indicators found in cluster each cluster [2]. Clusters 6, 8, 18, 20, 34, 49, 72, and 73, highlighted by arrows with orange asterisks are used in the CDCI analysis [2]. Image sourced from [2]

The clustering results of the DBHT algorithm found a total of $k = 102$ clusters with varying sizes, with the largest clustering containing 111 indicators and the smallest having only 4 [2]. When overlaying the fundamental categories with the clustering results of the DBHT as shown in Figure 3, we see similar results to that of the PCA. In fact, analysis using the Adjusted Rand Index indicates that 48 out of the 102 clusters have no single dominates category

within in the cluster. This again suggests that the fundamental categories do not reduce the dimension of our data well [2].

Local analysis of each cluster shows us that although most clusters contain indicators from different categories there still remains a level of interpretability found in each of the clusters [2]. Some clusters contain two or more groups of indicators such as cluster 5 that contains very important economic indicators as well as innovation indicators implying innovation could be important to economic development or vice versa [2]. One also finds overarching themes in other clusters such as, economic measures particularly related balance of current account and external trade (cluster 41), classifiers of underdeveloped countries (cluster 72), important economic measures (cluster 5), energy (cluster 21) and distress status of a countries debt (cluster 101) [2]. These are promising result when looking to create composite indicators using the clusters found in DBHT [2] as its gives us a flavour of what development composite indicators based on these clusters could measure.

3.5 Cluster Driven Composite Development Indicators (CDCI)

The authors in [2] consequently propose to aggregate the DBHT clusters by means of using the no-weighted median to create the Cluster Driven Composite development Indciators (CDCIs).

The median is chosen in [2] due to its robustness to outliers and consistency when applied to different clusters. Additionally the median is a non-parametric method of aggregating indicators therefore keeping the proposed CDCI's free of the need for human interaction [2]. Note since the development indicators are already standardised and normalised thus we do not need to apply any of these procedures before taking the median [2].

$$I_k = \text{median}\{(X_i)_{i \in \text{cluster k}}\}$$

The CDCI created from clusters 6, 8, 18, 20, 34, 49, 82 and 73 will be used for further analysis and can be seen on Figure 3 as the clusters marked with an asterisk [2].

3.5.1 Dynamic Analysis

A dynamical analysis of the CDCI's can be performed to check the stability of the cluster over time. To test this we compute E^w and then the respective DBHT for a 4 year window for a rolling window of 16 years [2]. Similarity of the DBHT can be crosswise compared for each of the 4 year window using the Adjusted Rand Index (ARI) [2]. The results of the dynamic analysis can be found in Figure 4 with both heat maps boasting high similarity between different years. This means that the clusters over time have high stability and thus comparing CDCIs at different time points can be done without worry of a vastly change network structure [2].

3.5.2 Performance Comparison

The performance comparison of CDCIs stems from the following argument, if the cluster produced from the DBHT have effectively reduced the dimensions

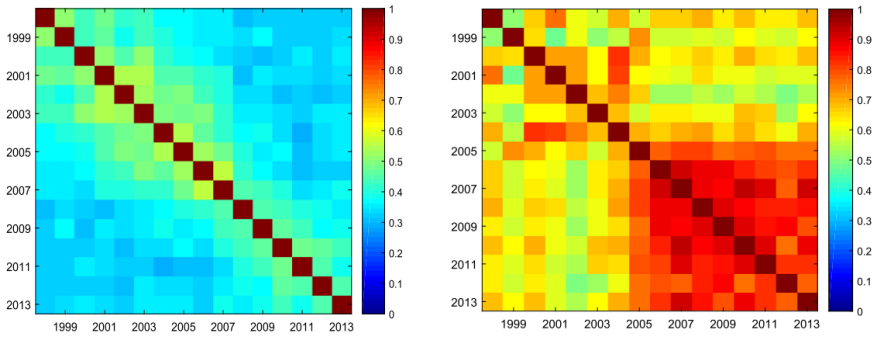

Figure 4: The heatmap on the right displays the ARI computed between the DBHT clustering of the CDCIs, utilising identical parameters for the rolling window as the left panel. Conversely, the left heatmap represents the Adjusted Rand Index (ARI) calculated between the DBHT clustering of indicators for each pair of 16 rolling time windows, each with a length of 4 and a shift of 1 year. See the colour legend located on the right of each panel with 1 indicating complete similarity and 0 depicting no similarity. Sourced from [2]

of the original data set E then it should be possible to reconstruct the E using the the CDCIs [2]. A comparison between CDCIs and other benchmark sets of composite indicators by comparing how well each set of composite indicators can recreate E [2].

Let us assume that each indicator X_i can be represented as a linear factor model of composite indicators,

$$X_i = \sum_{k-1}^{K} \gamma_{ik} \tilde{I}_k + \epsilon_i$$

where $\tilde{I}_k$ is the k^{th} composite indicator of either the CDCI's or alternative schemes of composite indicators used as bench marks [2]. Then an elastic regression can then be fitted and the error of the regression MSE measured

$$MSE = \sum_{i=1}^{N} (X_i^{\text{predicted}} - X_i)^2$$

where $X_i^{\text{predicted}}$ is the predicted value for the indicator X_i using the elastic regression of the composite indicators [2]. Following this, the error reduction ratio ERR can be computed to measure the performance of the CDCI which we denote as

$$ERR = \frac{MSE_{\text{CDCIs}}}{MSE_{\text{Alternative}}}$$

The results of this comparison as seen in Table 2 show that the CDCI are the most effective method of reducing the dimentionality of E compared to the random benchmark and a page rank alternative due to the ERR value being less than 1 [2]. Note PageRank is a network centrality measure than is calculated from the PMFG to identify the most system wide influential indicators in the network [2].

Random	PageRank
0.66	0.71

Table 2: The column labeled random displays the ERR computed for the random benchmark, employing 102 random subsets of indicators, repeated 100 times [2]. The column labelled PageRank presents the same calculation, but with 102 of the most influential indicators identified through PageRank assessment [2]. Results sourced from [2]

3.5.3 Pairwise Comparison

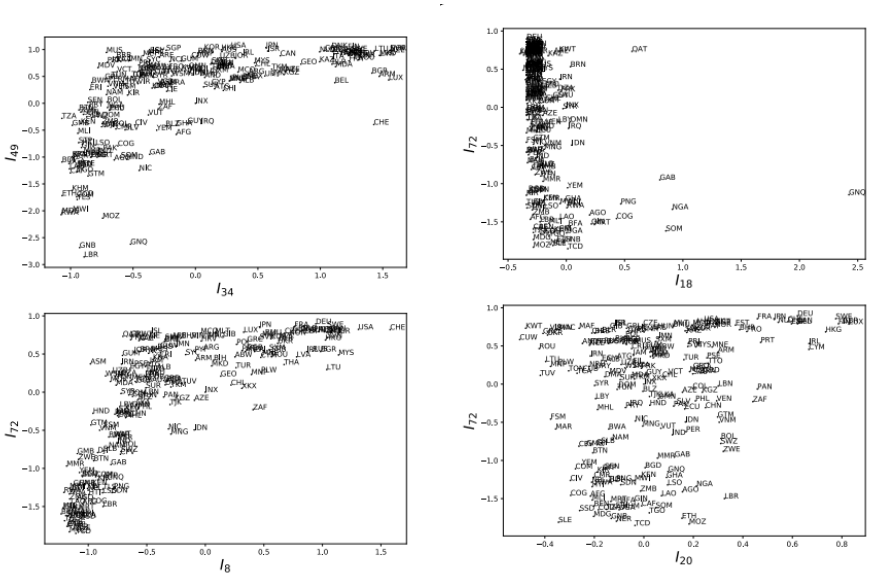

Figure 5: Pairwise comparison of different CDCIs for 1998, each point represents the CDCI values for specific country. Sourced from [2].

One ways CDCI can be used to track development is through pairwise comparison [2]. By plotting two CDCIs against each other we can observe the performance of all contries over different points in time or globally at a single point in time. One notable characteristic displayed in some pairwise plots for different counties and years is a hockey-stick shape graph [2], see Figure 5. When present the hockey-stick shape shows a clear partition of underdeveloped and developed counties indicated by the horizontal and vertical leg of the hockey-stick shape respectively. This shape also supports the two regime hypotheses [2], stating that countries below a barrier suffer from consistent development but once overcoming this barrier will experience high levels of growth. An insightful finding for policy makers as it may suggest areas of development for which a county should focus its efforts towards improving [2]. Some examples of this are I_{34} Vs I_{49} (mobile and banking services Vs Primary school statistics) which has the interpretation that access to mobile phone technology may help for developing countries help improve primary school statistics [2]. Additionally, I_8 Vs I_{72} (Secondary school enrolment Vs Underdeveloped countries) show how

secondary school enrolment can be used to indicate a counties development [2].

Conversely graphs that do not show a hock-stick shape tell a different story, see Figure 5. The plot I_{18} Vs I_{72} (natural resource abundance Vs underdevelopment) shows the so called resource curse were countries with high levels of natural resource abundance cannot make use of its resources due to inefficient governments, a characteristics of underdevelopment [2]. I_{20} Vs I_{72} (flow of foreign investment Vs underdevelopment) has a different type shape implying that there is no relationship between the flow of foreign investment and a countries underdevelopment [2].

4 Discussion

The development of composite indicators is a game of trade-off with each method sacrificing some kind of benefit in the pursuit of another [1]. In this section we we shall discuss the CDCI methodology and compare it to other methods in the literature outlining some of the trade-offs made for each.

Choosing which indicators to combine can be categorised into two approaches: top-down and bottom-up. The top-down approach identifies a target area of development for which the composite indicator aims to represent and then chooses appropriate indicators based on this. Composite Indicators based on a top-down philosophy such as the Human Development Index (HDI) [12] and Global Competitiveness Index (GCI) [13], are normally formed from a best educated guess. The drawbacks of this approach are obvious. There is no telling if the proposed individual indicators actually contribute to the target area of development. Additionally, the resulting composite indicators are completely objective and depend solely on the opinions of their creator [1]. This is why methods such as the ones used for HDI and GCI are highly criticised in the literature [1], [13] . To avoid these problems one can implement a bottom-up approach starting with the data and perform some kind of dimensionality reduction to find groupings of indicators based on some kind of similarity metric [6]. CDCI uses this approach and although shown to effectively group development indicators [2] it is not void of having problems. The trade-off of the bottom-up approach comes in the form of interoperability of the data-driven grouping where the task of identifying perfectly the area of development is near impossible [1]. Take clustering formed from a Pearson correlation metric for example such as DBHT, although we can identify that that clustered indicators are linearly related in some form we cannot tell what the cause for this dependency is [14]. Although previously mentioned at the end of Section 3.4 that the clusters formed from the DBHT algorithm are somewhat interoperable it is important to point out that this is still a best educated guess and any conclusion drawn should be done with caution.

Another problem with commonly used bottom-up approaches such as PCA and K-mean clustering [6], is that many of these methods are supervised in some form or another. In the case of PCA the user has to choose the number of components used to reduce the dimension of the data [8] and similarly K-mean clustering requires you to input a desired clustering number before starting [6]. This is where DBHT and other unsupervised clustering methods shine, being totally dependent on the data rather than human intervention they provide an objective way of forming clusters [2].

Once an appropriate grouping of indicators has been formed a common approach is to apply weighting to each of the indicators in a cluster [6]. The logic behind weighting is that it can be assumed that each indicator is essentially capturing a certain level of the desired development and thus representing the indicator in proportion to this level would result in an indicator that doesn't contain irrelevant information [1]. Thus not weighting indicators (i.e. not applying a coefficient to the indicators value before aggregation) means there is no diffracting between important and less important indicators in the aggregation process. The trade-off for weighting is inconsistency [1]. Whether the weighting system is data-driven like correlation based weighting or not like the budget allocation process (BAP), we encounter this problem. BAP weights indicators by asking a group of experts to allocate points depending on a spending budget to indicators which could consequently be different depending on the group of individuals allocating the points [1]. Correlation based weighting weights indicators by their respective correlation but again could be subjected to inconsistency depending on what indicators are being used since indicators with a non-linear dependency could be discounted as unimportant despite having a relationship with the area of development in question [1]. Alternative methods to weighting indicators is no-weighting and equal weighting. Both approaches attempt the fix the inconsistency weighting indicators but ignore the added value previously mentioned. CDCIs takes a no weighting approach to forming composite indicators [1].

Aggregating non-weighted and weighted indicators is the final issue encountered on the path to a composite development indicator [6], with the mean, median, additive are just such some of the simplest methods for aggregating [6], [2]. In the case of an additive method, we find that this method is highly susceptible to anomalies and may encourage countries to focus on a single area of development to increase a composite indicator score [6]. Similarly, the mean is sensitive to extremes, yielding unrealistic results when a dataset contains few very high or very low values The median however is not susceptible to extreme values and therefore is the better aggregation method in this respect. As we move away from simple methods of aggregation to more complex methods we may see better representative statistics but loses constancy when comparing composite indicators.

5 Extension

While the DBHT strives to be free of assumptions, its algorithms inherently imposes uniqueness when assigning vertices (indicators) to clusters. One could argue that in the case of creating a composite development indicator, it may not be advantageous to assume that a development indicator belongs to only one cluster. In this section, I discuss this argument further and suggest a similar but alternative path for future research.

The problem with mutually exclusive clustering is that some indicators may have similar contributions to more than one cluster. Therefore by ignoring this fact, the CDCIs proposed in [2] may in fact be a weaker measure of development, due to the missing information that these multi-cluster dependent indicators could provide. One can argue that many development indicators are affected by different areas of development such as mobile cellular subscriptions which

require a healthy population to use mobiles (health), education to use mobiles (education) and the relevant infrastructure to provide data (infrastructure) [2]. Therefore it would not be a wild proposition for indicators like mobile cellular subscriptions to belong to more than one fundamental category especially when found to have the highest centrality measure in terms of PageRank for the PMFG of E [2]. Despite already establishing that the clusters formed by the DBHT algorithm are better suited to classifying areas of development for which these composite indicators attempt to measure [2]. It still remains, like for the fundamental categories, that these indicators could be members of more than one grouping. Without ample research into this area, it would be crude to discount that removing this assumption in the DBHT may yield promising results.

Non-mutually exclusive clustering has been demonstrated to work well in feature extraction process of complex systems where items in the system have multiple features [15]. Some such problems that have been shown to benefit from overlapping clustering are emotion detection when listening to music [16], video classification based on genres [17], text clustering [18] and social network analysis [19]. In the case of community detection in complex networks, for which finding communitys of development indicators is a subset of, overlapping clustering techniques based on graph theory have been shown to be useful, see [18], [19], [20], [21] and [22].

One such example of graph-based non-mutually exclusive clustering is the cluster-overlap Newman Girvan algorithm (CONGA) presented in [23] which acts as an extension of the Girvan and Newman algorithm (GN algorithm). Like with many of the cluster overlapping algorithms in the literature [15], CONGA seeks to introduce overlapping clustering into an existing clustering algorithm by weakening the unique clustering condition of a clustering algorithm [23]. This is done in CONGA by implementing an extra step in the GN algorithm that allows for splitting vertices by considering the vertex betweenness and split betweenness, permitting vertices to exist in multiple clusters. In [23] CONGA is tested against its predecessor GN for a series of synthetic and real networks, considering networks with disjoint communities and networks with non-disjoint communities. The results of this analysis show that for disjoint community networks, CONGA has similar if not worse results than GN but for non-disjoint community networks CONGA is the superior clustering algorithm compared to GN [23]. This is just one example how of relaxing the condition of unique clustering may be beneficial when the network in question has non-disjoint communities.

Similar to CONGA [23] and other overlapping clustering algorithms [15], i propose that further research is taken to explore relaxing the uniqueness condition of the DBHT algorithm for clustering development indicators. Since the DBHT algorithm subjects uniqueness when considering the strength of attachment for vertices in converging bubbles and mean average shortest path for all other vertices. The notion of vertex assignment to clusters based on the largest strength of attachment or minimum mean average shortest path ignores the condition when there these values are similar (close). One could instead look to introduce a weaker notion of assignment that factors in this condition or replace these measurements with measurements that allow for multiple cluster assignment. A notable caveat to this approach is that most measures of similarity often introduce a cut-off point to distinguish between similar and non-similar. This

would thus relinquish the perk of having a parameter-free clustering method unless a non-parametric measure of closeness is introduced. One possible way to overcome this issue, although computationally taxing, would be to choose a similarity cut-off point that solves some kind of optimally condition.

6 Conclusion

In this paper we have explored the article "A new set of cluster driven composite development indicators" [2] which sets out to create a new set of composite development indicators that are objective and data-driven. We have summarised the main results found in this article outlining that grouping development indicators through means of fundamental categories and PCA does not fully capture all the relevant information within their respective grouping [2]. DBHT is shown to be a valid method of clustering the development indicators and is therefore proposed to be used in a new set of Cluster Driven Composite development Indicators (CDCI) [2]. We explain the methodology behind the creation of CDCIs, in particular explaining the Planar maximum filter graph (PMFG), Directed Bubble Hierarchical Tree (DBHT) algorithm and the non-weighted median method of aggregation. The results of the analysis of CDCIs conducted in [2] are explored finding that: through dynamic analysis, it can be shown that the DBHT clustering has high similarity over different rolling time windows suggesting robustness when comparing CDCIs over different periods of time. Through performance comparison, CDCIS are shown to be better at explaining the original data set E than random and PageRank benchmarks when compared in a performance comparison [2]. Through pairwise compassion, CDCIS can be used to track the development of countries through a pairwise analysis [2]. We then compared the approach for CDCI with others in the literature, noting that using data-driven clustering algorithms sacrifices interpretability, requiring educated guesses to define each cluster's true measure of development. Conversely, the drawback of employing a non-weighted aggregation approach is the inability to differentiate between highly significant development indicators and less relevant ones during aggregation. However, one notable advantage of CDCIs found compared to other composite development indicators is that the DBHT algorithm is unsupervised and thus free from the influence of its creators. Finally, we introduced the domain of overlapping clusters and how and how this may benefit composite development indicators. Suggesting that one path for further research may be to look into relaxing the unique clustering condition of the DBHT algorithm and instead produce composite develop indicators based on clusters that are not mutually exclusive.

References

[1] Salvatore Greco, Alessio Ishizaka, Menelaos Tasiou, and Gianpiero Torrisi. On the methodological framework of composite indices: A review of the issues of weighting, aggregation, and robustness. *Social Indicators Research*, 141(1):61–94, Jan 2018.

[2] Anshul Verma, Orazio Angelini, and Tiziana Di Matteo. A new set of cluster driven composite development indicators. *EPJ Data Science*, 9(1), Apr 2020.

[3] Nancy Baster. Development indicators: An introduction. *Measuring Development*, page 1–20, Oct 2018.

[4] R. K. Wilson and C. S. Woods. *Patterns of world economic development*. Longman, 1982.

[5] Frederik Booysen. An overview and evaluation of composite indices of development. *Social Indicators Research*, 59(2):115–151, 2002.

[6] Michela Nardo, Michaela Saisana, Andrea Saltelli, and Stefano Tarantola. Tools for composite indicators building. In *JRC*, 2005.

[7] Won-Min Song, T. Di Matteo, and Tomaso Aste. Hierarchical information clustering by means of topologically embedded graphs. *PLoS ONE*, 7(3), Mar 2012.

[8] Ian T. Jolliffe and Jorge Cadima. Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202, Apr 2016.

[9] M. Tumminello, T. Aste, T. Di Matteo, and R. N. Mantegna. A tool for filtering information in complex systems. *Proceedings of the National Academy of Sciences*, 102(30):10421–10426, Jul 2005.

[10] Darij Grinberg. An introduction to graph theory, 2023.

[11] Suman Saha, Junbin Gao, and Richard Gerlach. A survey of the application of graph-based approaches in stock market analysis and prediction. *International Journal of Data Science and Analytics*, 14(1):1–15, Jan 2022.

[12] Jeni Klugman, Francisco Rodríguez, and Hyung-Jin Choi. The hdi 2010: New controversies, old critiques. *The Journal of Economic Inequality*, 9(2):249–288, May 2011.

[13] María-Dolores Benítez-Márquez, Eva M. Sánchez-Teba, and Isabel Coronado-Maldonado. An alternative index to the global competitiveness index. *PLOS ONE*, 17(3), Mar 2022.

[14] Priya Ranganathan and Rakesh Aggarwal. Common pitfalls in statistical analysis: The use of correlation techniques. *Perspectives in Clinical Research*, 7(4):187, 2016.

[15] Chiheb-Eddine Ben N'Cir, Guillaume Cleuziou, and Nadia Essoussi. Overview of overlapping partitional clustering methods. *Partitional Clustering Algorithms*, page 245–275, Oct 2014.

[16] Alicja Wieczorkowska, Piotr Synak, and Zbigniew W. Raś. Multi-label classification of emotions in music. *Advances in Soft Computing*, page 307–315, 2006.

[17] Cees G. Snoek, Marcel Worring, Jan C. van Gemert, Jan-Mark Geusebroek, and Arnold W. Smeulders. The challenge problem for automated detection of 101 semantic concepts in multimedia. *Proceedings of the 14th ACM international conference on Multimedia*, Oct 2006.

[18] Airel Pérez-Suárez, José F. Martínez-Trinidad, Jesús A. Carrasco-Ochoa, and José E. Medina-Pagola. Oclustr: A new graph-based algorithm for overlapping clustering. *Neurocomputing*, 121:234–247, Dec 2013.

[19] George B. Davis and Kathleen M. Carley. Clearing the fog: Fuzzy, overlapping groups for social networks. *Social Networks*, 30(3):201–212, Jul 2008.

[20] Michael R. Fellows, Jiong Guo, Christian Komusiewicz, Rolf Niedermeier, and Johannes Uhlmann. Graph-based data clustering with overlaps. *Discrete Optimization*, 8(1):2–17, Feb 2011.

[21] Sanjeev Arora, Rong Ge, Sushant Sachdeva, and Grant Schoenebeck. Finding overlapping communities in social networks. *Proceedings of the 13th ACM Conference on Electronic Commerce*, Jun 2012.

[22] Michele Coscia, Giulio Rossetti, Fosca Giannotti, and Dino Pedreschi. Uncovering hierarchical and overlapping communities with a local-first approach. *ACM Transactions on Knowledge Discovery from Data*, 9(1):1–27, Aug 2014.

[23] Steve Gregory. An algorithm to find overlapping community structure in networks. *Knowledge Discovery in Databases: PKDD 2007*, page 91–102, 2007.