A robust machine learning approach to SDG data segmentation

MWITONDI, Kassim S. <http://orcid.org/0000-0003-1134-547X>, MUNYAKAZI, Isaac and GATSHENI, Barnabas N.

Available from Sheffield Hallam University Research Archive (SHURA) at:

https://shura.shu.ac.uk/27584/

This document is the Published Version [VoR]

Citation:

MWITONDI, Kassim S., MUNYAKAZI, Isaac and GATSHENI, Barnabas N. (2020). A robust machine learning approach to SDG data segmentation. Journal of Big Data, 7 (1), p. 97. [Article]

Copyright and re-use policy

See http://shura.shu.ac.uk/information.html

Sheffield Hallam University Research Archive
http://shura.shu.ac.uk

Journal of Big Data

RESEARCH Open Access

A robust machine learning approach to SDG data segmentation

Kassim S. Mwitondi[1*], Isaac Munyakazi[2] and Barnabas N. Gatsheni[3]

*Correspondence:
k.mwitondi@shu.ac.uk
[1] College of Business, Technology and Engineering, Sheffield Hallam University, Sheffield, United Kingdom
Full list of author information is available at the end of the article

Abstract

In the light of the recent technological advances in computing and data explosion, the complex interactions of the Sustainable Development Goals (SDG) present both a challenge and an opportunity to researchers and decision makers across fields and sectors. The deep and wide socio-economic, cultural and technological variations across the globe entail a unified understanding of the SDG project. The complexity of SDGs interactions and the dynamics through their indicators align naturally to technical and application specifics that require interdisciplinary solutions. We present a consilient approach to expounding triggers of SDG indicators. Illustrated through data segmentation, it is designed to unify our understanding of the complex overlap of the SDGs by utilising data from different sources. The paper treats each SDG as a Big Data source node, with the potential to contribute towards a unified understanding of applications across the SDG spectrum. Data for five SDGs was extracted from the United Nations SDG indicators data repository and used to model spatio-temporal variations in search of robust and consilient scientific solutions. Based on a number of pre-determined assumptions on socio-economic and geo-political variations, the data is subjected to sequential analyses, exploring distributional behaviour, component extraction and clustering. All three methods exhibit pronounced variations across samples, with initial distributional and data segmentation patterns isolating South Africa from the remaining five countries. Data randomness is dealt with via a specially developed algorithm for sampling, measuring and assessing, based on repeated samples of different sizes. Results exhibit consistent variations across samples, based on socio-economic, cultural and geo-political variations entailing a unified understanding, across disciplines and sectors. The findings highlight novel paths towards attaining informative patterns for a unified understanding of the triggers of SDG indicators and open new paths to interdisciplinary research.

Keywords: Big Data, Consilience, Data randomness, Data Science, Development Science Framework, K-Means, Principal Component Analysis, Sample-Model-Assess, Sustainable Development Goals, Unsupervised Modelling

Introduction

The 17 Sustainable Development Goals (SDGs) signed up by 193 United Nations member states in 2015, as the blueprint for achieving a better and more sustainable future for mankind and planet earth span across various aspects of life [1]. Each goal is defined with measurable aims for improving our quality of life to be achieved by 2030 [2]. Since

Springer Open

© The Author(s) 2020. This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

then, governments, institutions, businesses and individual researchers across the world, have increasingly paid attention to the SDGs, mainly for national development strategies, technical and business improvements as well as theoretical and practical aspects of their implementation. The complex interactions of the SDGs, the magnitude and dynamics of inherent data attributes and the deep and wide socio-economic and cultural variations across the globe are both challenges and opportunities to the SDG project. In the light of the recent technological advances in computing power and explosions in data generation, this paper treats each SDG as a source of Big Data [3–5]. Across sectors and nations, Big Data challenges and opportunities manifest in technical and application forms. Technically, they are pathways towards addressing issues ranging from data infrastructure, governance, sharing, modelling and security and from an application perspective, they potentially lead to influential policies and improving decision making at institutional, national, regional and global levels. In particular, Big Data challenges and opportunities present potential knowledge for unlocking our understanding of the mutual impact—positive and negative, resulting from our interaction with our environment [6].

Indicators for the 17 SDGs pool together a wide range of issues—hunger, poverty, inequality, health, species facing extinction, land degradation, gender inequality, gaps in education quality, productivity and technological achievements. These issues span across sectors and regions and our sustainability requires an adaptive understanding of their triggers. It is in that context that we view them as highly voluminous, volatile and dynamic data attributes, the behaviour and variations of which we need to track and understand in a unified and interdisciplinary manner. A unified interdisciplinary understanding of the challenges we face hinges on the relationship between knowledge extraction from data and development, which is well-documented [7–9]. The United Nations has a series of publications relating to the relevance of Big Data to SDGs [10–12]-but none of these have specifically focused on Big Data modelling.

Innovations in data acquisition, storage, dissemination and modelling have taken different forms at different levels, most notably visual pictures of what the world is like [13, 14], while the Millenium Institute (https://www.millennium-institute.org/isdg) has developed tools for simulating patterns based on alterations of some key SDG metrics. All these tools provide enhanced visualisation and are capable of generating an infinitely large number of patterns, depending on the choices or perturbations made. However, they can be viewed as enhanced descriptive statistics generators and often simulated patterns are based on pre-determined assumptions, parameters, environment etc, which vary invariably in a spatio-temporal context.

A recent research work-Development Science Framework (DSF) for Big Data modelling of SDGs [15, 16], combines data streaming from external factors like Government policies, cross-border legislations, technical and socio-economic and cultural factors with data directly attributable to the SDG indicators. In the form of highly voluminous and dynamic data, the SDG indicators are inevitably associated with spatio-temporal and other forms of variation. It is in this context that this paper seeks to highlight paths for expounding triggers of SDGs indicators. Based on the original ideas in [15, 16], the approach is *consilient* in that it adopts an interdisciplinary approach to unifying the underlying principles, concepts and reasoning for a comprehensive SDG modelling. One of its major strengths is that it is designed to add a predictive power to existing tools for SDG data visualisation. The approach is adaptive to the well-documented root-cause analysis [17] and an automated observation mapping. It is modelled on existing knowledge systems and cross-sectoral governance arrangements [18], to extract huge chunks of data from selected SDGs for identifying and modelling triggers of indicators across SDGs. We shall be making some key assumptions-notably on regional homegeneity and heterogeneity, allowing data simulations based on one country's real data to be used as real data proxies. The paper is organised as follows. Section 1 presents the study background-including the study motivation, objectives and research question. The methodology-data sources and implementation strategy are in Section 2, followed by analyses and discussions in Section 3 and concluding remarks in Section 4.

Motivation

The motivation for this work derives from years of interdisciplinary work relating to modelling high-dimensional data. In particular, the complexity of SDGs interactions and dynamics renders itself readily to the problem of data randomness [19, 20]. Identifying triggers of the indicators amounts to uncovering *what works* in different sectors and countries which, given the spatio-temporal variations can be challenging. This work looks at variations in SDG data across sectors from the Southern, Eastern and Western parts of the African continent.

Research question and objectives

To uncover triggers of the indicators, we adopt a general pragmatic approach to examine similarities and dissimilarities among data attributes that could lead to uncovering potentially useful information in the attributes. Although this work is inspired by the narrative of SDG Big Data Modelling [15, 16], it does not carry out Big Data modelling, in the strictest sense of the word. Instead, it provides a pathway for a consilient approach to complex SDG data modelling via the research question *How can interdisciplinary research revolutionise knowledge extraction from SDG data?* It aims to demonstrate the complexity of answering the foregoing question through the following six objectives.

1. To exhibit the impact of data randomness and variations through data visualisation.
2. To promote interdisciplinary activities for problem identification and attainment of agenda 2030.
3. To highlight and support interdisciplinary paths for a unified understanding of global phenomena.

Methodology

This paper adopts the concept of Development Science Framework-DSF [15, 16], the main idea of which is to view each SDG as a Big Data node and the UN SDG data repository as a multi-disciplinary data fabric. The DSF consists of two layers-the inner and outer shells, via which it associates data streams and variations with internal and external factors. The former relates to variations within the actual data attributes while the latter relates to factors such as infrastructure, legislations and other socio-economic and geo-political variations which directly impinge on data modelling. This section outlines the mechanics of the framework, based on those considerations.

Data sources

The main data source for this work is the United Nations SDG data repository [2] which holds the full list of targets and indicators for all the 17 SDGs from 2000 to 2018. The full description of the targets and indicators is provided by the United Nations [21]. This work focuses only on structured data from five of the 17 SDGs-i.e., # 1, 2, 3, 4 and 9, using hundreds of SDG indicators in six African countries–Botswana, Cameroon, Ghana, Kenya, Rwanda & South Africa. Indicators for each of these SDGs fall within specific targets as illustrated in Table 1.

Selection of the variables was guided by the research question. For instance, it was reasonable to assert that the level and quality of education in a country would impinge on the level of innovation and productivity, hence the attained level of manufacturing and Research and Development (R&D). The data attributes were cleaned and reformatted to fit in with the modelling strategy. Labelling of the data could be carried out in various ways—by country, by indicator, by region, etc. This work focuses on country variations, implying that performance within countries provides potentially useful information on triggers of indicators variations and that indicator variables are predictors of geographical locations. The implementation strategy adopted in this study is outlined below.

Implementation strategy

This work applies two commonly used techniques—i.e., Principal Component Analysis (PCA) and data clustering-a technique used to group data objects according to their homogeneity. For the latter we use the K-Means technique [22, 23]. Both methods use scaled matrix of sampled features to reduce the data dimensionality.

Table 1 Selected indicators, associated SDGs and number of cases used in the study

Indicator	Variable name	SDG
Empl. pop. below intern. poverty line	EBI (EB)	1: End poverty in all forms
Prevalence of undernourishment %	UNDERNOUR (UN)	2: End hunger; food security
Infant mortality rate per 1K live births	INFMORTPER1K (IN)	3: Healthy lives for all
Maternal mortality ratio	MATMORTRATIO (MA)	3: Healthy lives for all
Deaths from non-comm. diseases	NONCOMMDEATHS (NO)	3: Healthy lives for all
Road traffic deaths per 100K people	ROADACCIDEATHS100K	3: Healthy lives for all
Participation rate in organized learning	EARLYORGLEARN	4: Inclusive quality education
Prop. of teachers with min. pedag. train.	TRAINEDTEACHERSMINI (TR)	4: Inclusive quality education
Minimum proficiency in mathematics	MINIPROMATHS (MI)	4: Inclusive quality education
Minimum proficiency in reading	MINIPROREAD	4: Inclusive quality education
CO2 fuel emissions (millions of tonnes)	COEMISSFUEL	9: Infrastruct. & Innovation
Kg. of CO2 per unit of GDP in USD	COEMISSGDP	9: Infrastruct. & Innovation
Kg. of CO2 per manufact. unit in USD	COEMSSPUMAV	9: Infrastruct. & Innovation
Total official flows for infrastructure	INFRASFLOW	9: Infrastruct. & Innovation
Manufact. value added (GDP prop.)	MANUFGDP (MG)	9: Infrastruct. & Innovation
ManufacT. value added per capita	MANUFPCTA (MP)	9: Infrastruct. & Innovation
Prop. of med. & high-tech value added	MEDHIGHTECHI (ME)	9: Infrastruct. & Innovation
Mobile net. coverage (pop. proportion)	MOBCOVERAGE (MO)	9: Infrastruct. & Innovation

Dimensional reduction using PCA

Principal component analysis (PCA) seeks to transforms a number of correlated variables into a smaller number of uncorrelated variables, called principal components. It uses the correlation among variables to develop a small set of components, which empirically summarise the correlations among them. Its main goal is to reduce data dimensionality—i.e., reduce the number of variables while retaining most of the original variability in it.

Principal components are extracted in succession, with the first component accounting for as much of the variability in the data as possible and each succeeding component accounting for as much of the remaining variability as possible. More specifically, PCA is concerned with explaining the variance-covariance structure of a high dimensional random vector through a few linear combinations of the original component variables. The indicators in Table 1 can be formulated in a generic form as in Eq. 1.

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & x_{13} \ldots x_{1n} \\ x_{21} & x_{22} & x_{13} \ldots x_{2n} \\ x_{31} & x_{32} & x_{33} \ldots x_{3n} \\ \ldots & \ldots & \ldots\ldots\ldots \\ x_{p1} & x_{p2} & x_{p3} \ldots x_{pn} \end{bmatrix} = \left[x_{ij}\right] \tag{1}$$

Equation 1 corresponds to the data source notation in Algorithm 1 and, in this application, it describes the 11 numeric variables selected from Table 1, constituting the set

$$\mathcal{SDGI} = \{MA, NO, EB, MI, IN, MG, ME, MP, MO, UN, TR\} \subset \mathbb{R}^n \tag{2}$$

Extracted components are inferred from the correlations among the indicator variables, with each component being estimated as a weighted sum of the variables. That is, we can extract 11 components as random variables such that

$$\mathcal{PC}_k = \{w_{ik}MA, w_{ik}NO, w_{ik}EB, w_{ik}MI, w_{ik}IN, w_{ik}MG, w_{ik}ME, w_{ik}MP, w_{ik}MO, w_{ik}UN, w_{ik}TR\} \tag{3}$$

where $k = 1, 2, 3, \ldots, 10, 11$ denoting the number of components and $i = 1, 2, 3, \ldots, 10, 11$, denoting the number of variables. The vectors w_{ik} are chosen such that the following conditions are met.

1. $\|w_k\| = 1$
2. Each of the $\mathcal{PC}_k$, maximises the variance $V\left\{ w_k^{'} \mathcal{SDGI}_k \right\}$ and
3. The covariance $COV\left\{ w_k^{'} \mathcal{SDGI}_k \; w_r{}' \mathcal{SDGI}_r \right\} = 0, \; \forall k < r$

In other words, the principal components are extracted from the linear combinations of the original variables maximising the variance and have zero covariance with the previously extracted components. It can be shown that the number of such linear combinations is exactly 11. Our applications will adopt this method.

Underlying mechanics of data clustering

Another common unsupervised learning method is cluster analysis [24, 25] groups data according to some measures of similarity, and it is generally described as follows. Given

$\mathcal{SDGI}$ data in Eq. 2 and, assuming k distinct clusters for $\mathcal{SDGI}$, i.e., $\mathcal{C} = \{c_1, c_2, \ldots, c_k\}$, each with a specified centroid. Then, for each of the vectors $j = 1, 2, \ldots p$, we can obtain the distance from $\mathbf{v}_j \in \mathcal{SDGI}$ to the nearest centroid from the set $\{\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_k\}$ as

$$\mathcal{D}_j(\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_k) = \min_{1 \le l \le k} d\left(\mathbf{x}_l,\ \mathbf{v}_j\right) \tag{4}$$

where $d(.)$ is an adopted measure of distance and the clustering objective would then be to minimise the sum of the distances from each of the data points in $\mathcal{SDGI}$ to the nearest centroid. That is, optimal partitioning of $\mathcal{C}$ requires identifying k vectors $\mathbf{x}_1^*, \mathbf{x}_2^*, \ldots, \mathbf{x}_k^* \in \mathbb{R}^n$ that solve the continuous optimisation function in Eq. 5.

$$\min_{\{\mathbf{x}_1, \ldots, \mathbf{x}_k\} \in \mathbb{R}^n} f(\mathbf{x}_1, \ldots, \mathbf{x}_k) = \sum_{j=1}^{p} \mathcal{D}_j(\mathbf{x}_1, \ldots, \mathbf{x}_k) \tag{5}$$

Minimisation of the distances depends on the initial values in $\mathcal{C}$, hence if we let $z_{i=1,2,\ldots,n}$ be an indicator variable denoting group membership with unknown values, the search for the optimal solution can be through iterative smoothing of the random vector $x|(z = k)$, for which we can compute $\bar{\mu} = \mathbf{E}(x)$ and $\delta = \left\{\mu_k - \bar{\mu}|y = k \in \mathbf{c_z}\right\}$. In a labelled data scenario, $\left\{x_i, y_i\right\}$ $i = 1, 2, \ldots n$, Eq. 5 amounts to minimising Eq. 6

$$f(\theta) = \sum_{i=1}^{n} [y_i - g(x_i;\ \theta)]^2 \tag{6}$$

where x_i are described by the parameters $\left\{\bar{\mu} \text{ and } \delta\right\} \in \theta$ and $g(x_i;\ \theta)$ are fitted values. Equations 4 through 6 relate to the K-Means clustering algorithm [22, 23], which searches for clusters in numeric data based on pre–specified number of centroids. The decision on the initial number of centroids does ultimately impinge on the detected clusters and we shall be addressing this issue via the Algorithm in "The Sample-Measure-Assess (SMA) Algorithm" section.

Whether we are looking for variations among countries, SDGs or their indicators, interest is in their variant or invariant behaviour across the set and over time. Addressing spatio–temporal variations in SDG data appeals naturally to dealing with randomness in data [19, 20] and adopting interdisciplinary approaches to gaining a unified understanding and interpretation of data modelling. This work is not based on a complex high-dimensional dataset but, as explained in "Research question and objectives" section, it demonstrates the techniques in anticipation of such volumes and dynamics. The Sample-Measure-Assess (SMA) algorithm [26], described in "The Sample-Measure-Assess (SMA) Algorithm" section was developed to address variations in data due to inherent randomness.

The Sample-Measure-Assess (SMA) Algorithm

The SMA algorithm [15, 16] seeks to address issues of data randomness [19, 20]. It draws from existing modelling techniques such as the standard variants of cross–validation [27] and permutation feature importance [28]. Unlike many of its predecessors, the SMA has a built–in mechanism that allows it to handle data randomness more efficiently. Further, it is adaptable to a wide range of models and amenable to both clustering and classification problems. Its mechanics, described below, assume structured data [29], but can readily be extended to semi-structured and unstructured data. Its implementation is problem-specific and, in this case, the free parameters were chosen based on the assumptions made in the last paragraph of "Introduction" secion. For example, the decision to run the algorithm without the industrial manufacturing related variables was based on the prior knowledge that, relative to other African countries, these factors are predominantly influenced by South Africa.

Algorithm 1 SMA-Sample, Measure, Assess

```
 1: procedure SMA
 2:     Set $\mathbf{X} = [x_{i,j}]$ : Accessible Data Source
 3:     Learn $F(\phi) = \underbrace{(P)}_{x,y \sim D} [\phi(x) \neq y]$ based on a chosen learning model
 4:     Set the number of iterations to a large number $\kappa$
 5:     Initialise: $\Theta_{tr} := \Theta_{tr}(.)$ : Training Parameters
 6:     Initialise: $\Theta_{ts} := \Theta_{ts}(.)$ : Testing Parameters
 7:     Initialise: $\Pi_{cp} := \Pi_{cp}(.)$ : Comparative Parameters
 8:     Initialise: $s$ as a percentage of $[x_{\nu,\tau}]$ , say 1%
 9:     $s_{tr}$ : Training Sample $[x_{\nu,\tau}] \leftarrow [x_{i,j}]$ extracted from $\mathbf{X} = [x_{i,j}]$
10:     $s_{ts}$ : Test Sample $[x_{\nu,\tau}] \leftarrow [x_{l \neq i,j}]$ extracted from $\mathbf{X} = [x_{i,j}]$
11:     for $i := 1 \rightarrow \kappa$ do: Set $\kappa$ large and iterate in search of optimal values
12:         while $s \leq 50\%$ of $[x_{\nu,\tau}]$ do Vary sample sizes to up to the nearest integer 50% of $X$
13:             Sampling for Training: $s_{tr} \leftarrow X$
14:             Sampling for Testing: $s_{ts} \leftarrow X$
15:             Fit Training and Testing Models $\hat{\mathcal{L}}_{tr,ts} \propto \Phi(.)_{tr,ts}$ with current parameters
16:             Update Training Parameters: $\Theta_{tr}(.) \leftarrow \Theta_{tr}$
17:             Update Testing Parameters: $\Theta_{ts}(.) \leftarrow \Theta_{ts}$
18:             Compare: $\Phi(.)_{tr}$ with $\Phi(.)_{ts}$ : Plotting or otherwise
19:             Update Comparative Parameters: $\Pi(.)_{cp} \leftarrow \Phi(.)_{tr,ts}$
20:             Assess: $P(\Psi_{D,POP} \geq \Psi_{B,POP}) = 1 \iff \mathbb{E}[\Psi_{D,POP} - \Psi_{B,POP}] = \mathbb{E}[\Delta] \geq 0$
21:         end while
22:     end for
23:     Output the Best Models $\hat{\mathcal{L}}_{tr,ts}$ based on $\mathbb{E}[\Delta] \geq 0$
24: end procedure
```

The dataset $\mathbf{X} = \left[x_{i,j}\right]$ corresponds to Table 1 and the learning model $F(\phi)$ is, in this case, either PCA or K-Means. The constant κ used here is a free parameter, determined by the user. The algorithm draws samples from the full data, generating random training and testing samples with distributional parameters varying from sample to sample. The parameters $\Theta_{tr}(.) \leftarrow \Theta_{tr}$ and $\Theta_{ts}(.) \leftarrow \Theta_{ts}$ are updated by randomly drawn samples, $[x_{\nu,\tau}] \leftarrow \left[x_{i,j}\right]$, initialised in step 8, and sampled in steps 9 and 10. They are random and they remain stateless across all iterations. The same applies to $[x_{\nu,\tau}] \leftarrow \left[x_{l \neq i,j}\right]$. The notation $\hat{\mathcal{L}}_{tr,ts} \propto \Phi(.)_{tr,ts}$ represents multiple trained and tested machine learning models, adopted by the user. The loop from step 11through 19 involves sampling through the data with replacement, fitting the model and updating the parameters. The choice for the best performing model is carried out at step 20, where $P\left(\Psi_{D,POP} \geq \Psi_{B,POP}\right)$ is the probability of the population error being greater than the training error.

Analyses, results and discussions

Analyses are presented from both descriptive and inferential perspectives in order to, firstly, grasp an understanding of the data we are looking at and, secondly, deciding what to do with the data in the attributes. Due to constraints on time and space, we work only with a handful, selected indicators, some of which are shown in Table 1.

Exploratory data analyses

Expolaratory data analysis *(EDA)* is a common good practice that helps gain insight into the data at hand–typically via visual inspection through graphs and numerical results. These early investigations are extremely useful in that they serve as either warning, hints or both as to the overall behaviour of the data–e.g., presence of outliers, missing data or systematic patterns. Figure 1 shows box plots of 4 of the selected indicators for the six countries. Each of the boxes is built on sorted scores, forming equal sized groups referred to as quartiles. The median line divides your univariate data into 2 parts, forming the "inter-quartile" range and containing 50% of all data. The lines extending from the top and bottom edges of the boxes, known as *whiskers* represent data poits outside the middle 50% and they are indicators of outlying cases. For all four indicators the boxes show that South Africa has a high level of consensus on the data collected over the period, whereas there are relatively wide variations for Rwanda and Cameroon.

In all four cases, South Africa is well isolated from the remaining five countries, which suggests a fundamental difference between the country and the rest. With the exception of Rwanda, where there are outlying cases in the upper quartile for maternal and infant mortality indicators and Kenya with outlying cases on undernourishment, the remaining boxes are quite evenly distributed. These variations warrant further investigations.

Another common *EDA* method for investigating data behaviour is to look at the individual univariate densities and try to hypothesise what message they provide. Note that a density estimator seeks to model the probability distribution that generated the data and there cannot be a better example than the histogram. The challenges we face with histogram estimation–i.e., choosing the bin size and location are typical what you would encounter with density estimation. Figure 2 presents the same four indicators discussed above, drawn at a very small bandwidth of 0.01, the equivalent of choosing very small bin sizes for an histogram. The number of modes in each density suggests existence of a group or a distinctive feature within that dataset and altering the bandwidth changes the number of these features and, like with histograms, various choices can lead to data representations with distinctively different features.

Univariate analysis has many limitations which arise from factors outside that variable. With each SDG associated with hundreds of indicators, the need to explore interactions and variations among data attributes is apparent. One way to explore such relationships

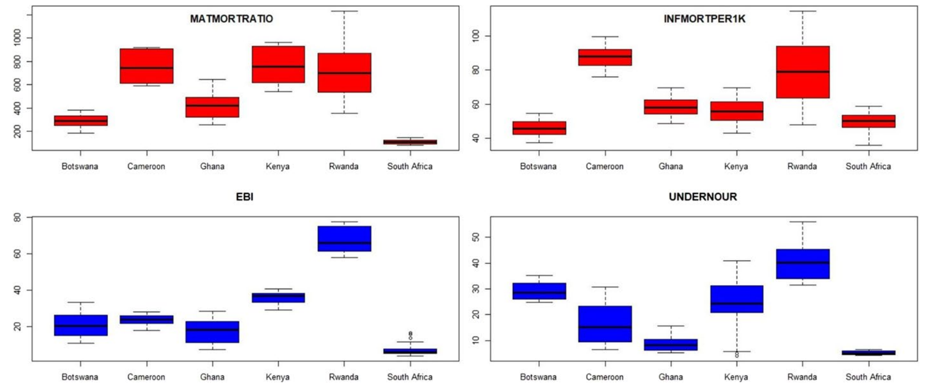

Fig. 1 Box plot distributions for four indicator variables selected from Table 1

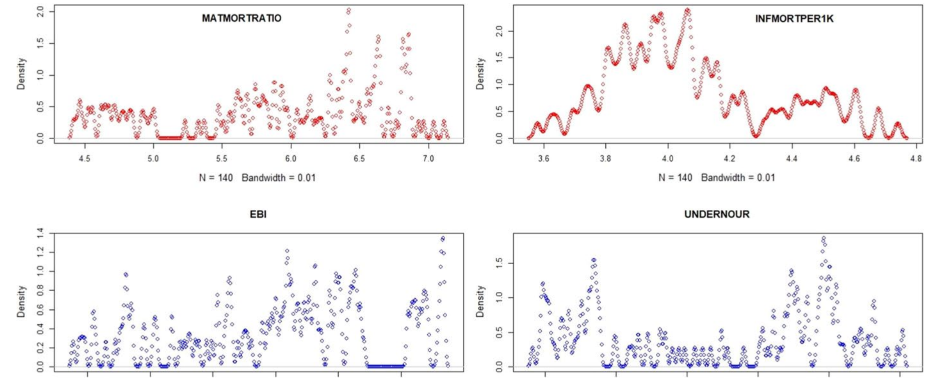

Fig. 2 Density estimations for the four variables plotted at bandwidth 0.01

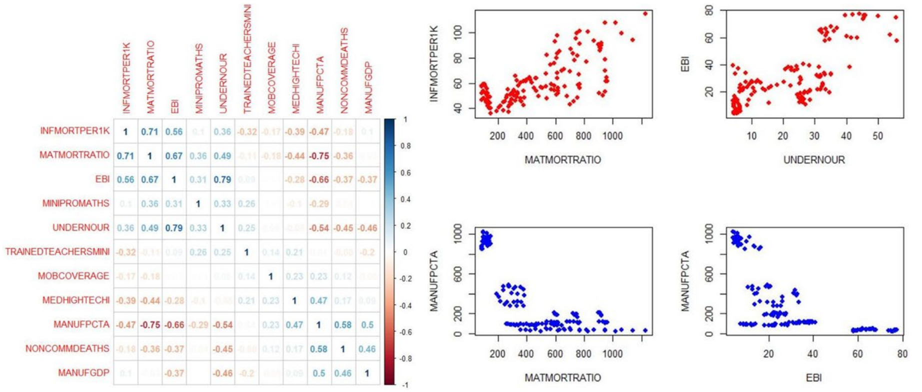

Fig. 3 Correlations among selected indicator variables described in Table 1

is through correlation analysis which measures the strength of a linear association between two variables. The left hand side panel in Fig. 3 shows all paired correlations, with the strongest positively being between UNDERNOUR and EBI (79%) and between MATMORTRATIO and INFMORTPER1K (71%). The indicators MANUFPCTA and MATMORTRATIO exhibit strong negative correlation (−75%) while MANUFPCTA and EBI stand at -66%. The correlation plots for these four indicators are given on the right hand side panel. Effectively, the correlation function attempts to draw a line of best fit through the paired indicators without regard to which influences the other–hence the old dictum, correlation does not mean causation.

Unsupervised learning

The EDA methods presented in "Exploratory data analyses" section provide good insights into the overall data behaviour, relationships and variations among the data attributes. This section focuses on *unsupervised learning*-a process of drawing inferences from unlabelled data. It implements two unsupervised learning models–PCA and clustering.

Principal Component Analysis

We adopt PCA for a simple illustration on how to use the data matrix in Eq. 2, to try and uncover triggers of the indicators, via similarities and dissimilarities among countries, SDGs and indicators. Since PCA transforms indicators into linear combinations of an underlying set of hypothesized or unobserved components, each component may be associated with 2 or more of the original indicators. That is, rather than measuring infant and maternal mortality, say, we may have a single measure on the state of health services in a particular country.

Table 2 exhibits PCA loadings-values that relate the specific association between factors and the original SDG indicators. In particular, the concept of "loadings" refers to the correlation between the indicators and the factors and they are key to understanding the nature of a particular factor. Loadings derive from the magnitude of the eigenvalues associated with the individual indicator. Squared factor loadings indicate what percentage of the variance in an original SDG indicator is explained by a component. Consequently, it is necessary to find the loadings, then solve for the factors, which will approximate the relationship between the original indicators and underlying factors. The directions of the SDG indicators here reflect the role each indicator played in forming each of the eleven components. In interpreting extracted components, the main consideration is about these values as each component is the directions which maximizes variance among all directions orthogonal to the previous component.

The two panels in Fig. 4 derive from the data in Table 1 and the methods in Eqs. 2 and 3. In the panel on the left, involving 140 observations on 11 SDG indicators, the highest component accrued a total of 38.2% of the total variation. Despite the small size, clear country-specific patterns emerge-on manufacturing per capita, for instance, South Africa dominates, with deaths from non-communicable diseases and other aspects of manufacturing taking the same direction. On the other hand, issues relating to poverty-maternal and infant mortality, undernourishment and the employed population below international poverty line are dominated by statistics from Rwanda. The right hand side panel, yielding a 39.4% explained variance by the first component, excludes manufacturing related variables-MANUFPCTA, NONCOMMDEATHS and MANUFGDP, which are mostly associated with South Africa.

In both cases in Fig. 4, multiple runs through the SMA algorithm, using randomly sampled data, gave Eigenvalue rule cut-off points of between 2 and 3 components. In searching for triggers of SDG indicators, a thorough understanding of these categories of variables is required. It is under such circumstances that interdisciplinarity plays a crucial role. For such a small study, it is important to interpret findings with care, as such patterns may arise from the level and quality of data. We take a closer look at variations in patterns attributable to data randomness.

Data clustering

The K-Means clustering algorithm [22, 23] was applied to the data in Table 1 for a range of between 2 and 7 clusters. The results for 2, 3, 4 and 5 clusters are summarised in Table 3, where it is evident that in all cases South Africa distinctively stands out in forming the clusters. With two clusters, the pattern is almost binary-South

Table 2 Loadings provide differentiation among extracted components

Indicator	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11
MATMORTRATIO	0.41	− 0.22	0.20	− 0.002	0.05	− 0.15	0.17	− 0.44	0.27	− 0.18	0.63
NONCOMMDEATHS	− 0.29	− 0.18	0.42	− 0.05	− 0.18	0.59	− 0.39	− 0.34	0.06	− 0.18	− 0.05
EBI	0.42	0.10	0.16	− 0.23	0.06	0.30	− 0.09	0.08	0.29	0.73	− 0.04
MINIPROMATHS	0.18	0.13	0.56	0.44	− 0.21	− 0.38	− 0.37	0.18	− 0.23	0.08	0.02
INFMORTPER1K	0.32	− 0.38	0.17	− 0.27	0.21	0.14	0.15	0.02	− 0.74	− 0.02	− 0.10
MANUFGDP	− 0.21	− 0.41	0.46	0.03	0.22	− 0.19	0.40	0.17	0.38	0.01	− 0.37
MEDHIGHTECHI	− 0.23	0.33	0.15	− 0.19	0.71	− 0.26	− 0.24	− 0.36	− 0.07	0.07	− 0.03
MANUFPCTA	− 0.43	0.01	0.17	− 0.11	0.12	0.14	0.08	0.49	− 0.11	0.12	0.66
MOBCOVERAGE	− 0.08	0.30	0.25	− 0.68	− 0.49	− 0.28	0.18	− 0.07	− 0.03	− 0.05	− 0.05
UNDERNOUR	0.37	0.29	0.13	− 0.17	0.24	0.20	− 0.12	0.45	0.21	− 0.59	− 0.07
TRAINTEACHMINI	0.01	0.53	0.22	0.36	0.01	0.33	0.60	− 0.19	− 0.15	0.01	− 0.04

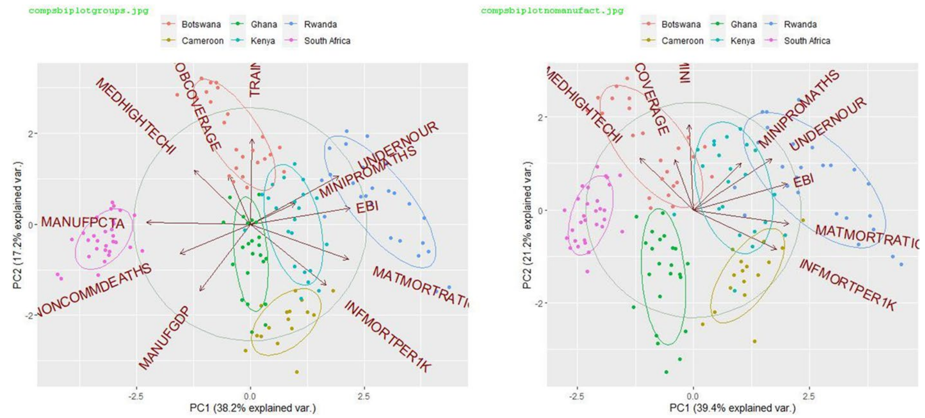

Fig. 4 The LHS and RHS panels are with and without manufacturing related variables respectively

Table 3 Country-by country participation in the formation of clusters

Centroids	Cluster	Botswana	Cameroon	Ghana	Kenya	Rwanda	South Africa	Total	$\frac{B_{ss}}{T_{ss}}$
K=2	C1	9	0	0	0	0	30	39	
	C2	15	16	23	22	25	0	101	28.1%
K=3	C1	18	0	23	1	1	0	43	
	C2	0	16	0	21	24	0	61	
	C3	6	0	0	0	0	30	36	40.0%
K=4	C1	0	16	7	9	0	0	32	
	C2	0	0	0	0	25	0	25	
	C3	24	0	16	13	0	0	53	
	C4	0	0	0	0	0	30	30	50.6%
K=5	C1	0	16	5	7	0	0	28	
	C2	0	0	0	0	25	0	25	
	C3	14	0	18	15	0	0	47	
	C4	10	0	0	0	0	0	10	
	C5	0	0	0	0	0	30	30	58.3%

Africa versus the rest, which is similar to the pattern observed with PCA in Figure 4. While South Africa still dominates under $K = 3$, Botswana and Ghana dominate one cluster while Cameroon, Kenya and Rwanda dominate the other cluster. Particularly important is the last column in Table 3, exhibiting the ratio B_{ss} (between-cluster sum of squares) over T_{ss} (the total sum of squares), giving what is basically the within-cluster sum of squares. The K-Means algorithm uses the minimum sum of squares to identify clusters. The algorithm iteratively updates cluster centres, allocating observations as it goes and stops only when the maximum number of iterations is reached or the change of within-cluster sum of squares in two successive iterations is less than the set threshold. Thus, the values in the last column of Table 3 measures the total variance in the dataset that is due to that level of clustering. That is, by assigning the samples to the specified clusters rather than the total number of samples, we are able to show the reduction in the sum of squares each cluster achieved.

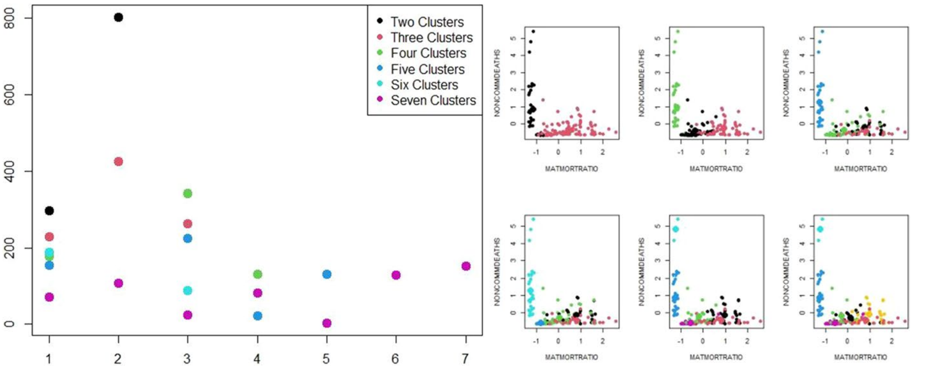

Fig. 5 The within cluster variations (LHS) and each of the six cluster categories (RHS)

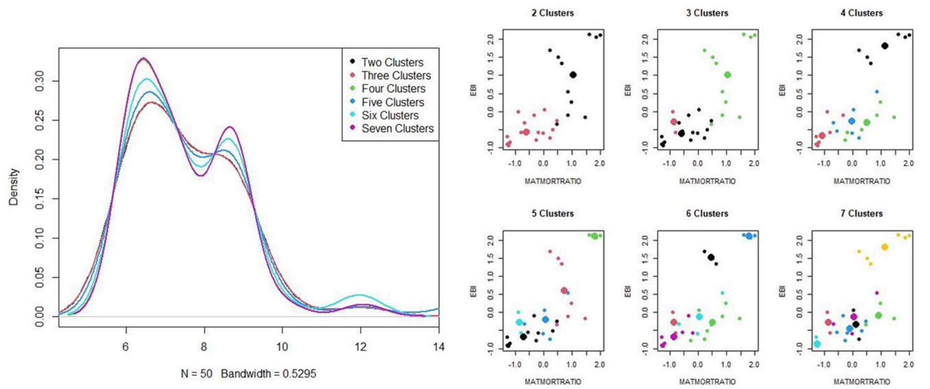

Fig. 6 Within and between cluster variations for the clusters on the RHS based on 30 sample size

The four clusters presented in Table 3 are from the total of six clusters generated from the full dataset. The left hand side panel in Fig. 5 exhibits the within cluster variations per cluster for each of the six cluster categories. Notice the huge variation within the category $K = 2$ and the relatively high variation for $K = 3$ and $K = 4$ as compared to the remaining categories. This pattern is reflected by the six small panels to the right, corresponding to each of the six cluster categories. As noted above, the influence of industrial manufacturing factors is felt heavily.

To suppress the dominant influence of South Africa, we remove the three variables-NONCOMMDEATHS, MANUFGDP and MANUFPCTA and run the K-Means clustering through the SMA Algorithm. Multiple samples of between 30 and 60 were randomly drawn from $\mathbf{X} = \left[x_{i,j}\right]$ fifty times, with parameters of interest being the variation within and between clusters. The densities to the left of Figures 6 and 7 exhibit the within cluster variations for the clusters shown in the right hand side panels. These were selected from 50 runs through the SMA algorithm. Both Figs. 6 and 7 provide insights into the naturally arising structures in the data in Table 1.

The six panels on the right hand side of Fig. 6 exhibit within and between cluster variations based on the 30 sample size runs. Each panel is a 2 dimensional plot of the proportion of maternal mortality (MARTMORTRATIO) versus the proportion of

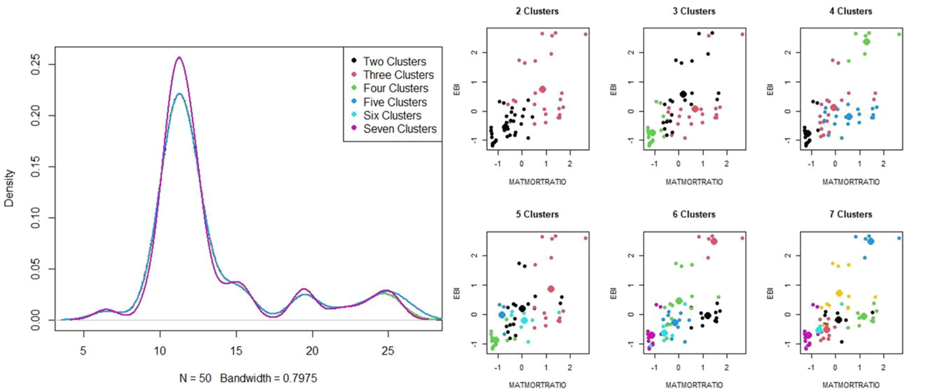

Fig. 7 Within and between cluster variations for the clusters on the RHS based on 60 sample size

employed population below the international poverty line (EBI), for 2, 3, 4, 5, 6 and 7 clusters. The plots indicate cluster overlaps as the number of clusters increases, making them increasingly less distinctive, which we can interpret as over–fitting.

The plots in Fig. 7 are similar to those in Figure 6, except that they are based on samples of size 60. They both provide insights into the naturally arising structures in the data in Table 1 and it can be seen that they exhibit consistent variations across samples, as the sample size increases–that is, the higher the number of clusters the more evident over–fitting becomes. Note that via the SMA, samples of different sizes can be drawn and implemented with different numbers of centroids. While, the final decision on the optimal number of clusters can be decided based on the set criteria for between–cluster variation, in practice it will also depend, *inter–alia* on the problem of interest. That is what entails interdisciplinarity, as domain knowledge outside modelling plays a crucial role here.

We can tell from Fig. 6 and 7 that while it is possible to capture key metrics on indicators, their triggers remain buried in the data. For example, the inclusion and omission of variables dominated by South Africa showed that socio-economic, cultural and geo-political variations make it impossible for an overarching strategy to be developed for the continent. Thus, attainment of agenda 2030 requires a unified understanding of the agenda at both low and high levels. Variations will typically arise from a wide range of causes and it is imperative that SDG stakeholders, like the Sustainable Development Goals Center for Africa *(SDGCA)* [30] in Kigali, engage with individual Governments through relevant departments to monitor SDG dynamics in a spatio-temporal context. Initiatives, geared towards accelerating attainment of agenda 2030, can be enhanced by adopting interdisciplinary approaches to providing relevant support to governments, civil society, businesses and academic institutions.

Summary of results

The analyses in this section sought to highlight the impact of data variation in addressing complex challenges in SDG monitoring. We considered soft and technical solutions- i.e., socio-economic and cultural variations and data interactions and dynamics. Initial EDA patterns isolated South Africa from the remaining five countries. The dominance of South Africa continued through PCA and clustering analyses. We attempted to iron out the impact of data randomness on variations by deploying the SMA algorithm, taking repeated samples of sizes 20 through 75, two of which are presented in Figs. 6 and 7, exhibiting consistent variations across samples, as the sample size increases. The indicators were not written on stone, and so a simple way to assess progress would be to think of how often they have been reviewed or whether there is a regular update forming posterior information. Posterior information forms the basis of prior knowledge which decision makers need for any interventions.

The decision to omit some of the manufacturing-related variables was based on a prior knowledge we notionally generated from the same data. For example, not all the variables in Table 1 were used in PCA and data clustering. Some variables, like MOBCOVERAGE, were used externally to provide "prior knowledge" for segmenting the SDG data. For example, our SDG #9 data showed that the mobile coverage across the selected countries was exceptionally high, making South Africa hardly distinguishable from the rest. This raises the question as to what triggers such development. Talk to different sections of the population and you might get different answers. Such disparate perceptions on the impact of mobile phones reflect the extent of data fragmentation among countries and even institutions within the same country.

The socio-economic, cultural and geo-political variations make it impossible for an overarching strategy to be developed for the African continent, or indeed elsewhere. It is on those premises that we emphasise a unified understanding, across disciplines and sectors, of the 2030 agenda at both low and high levels. Given the magnitude of SDG indicators, the dataset used in this section was a drop in the ocean. The foregoing results are therefore not geared towards establishing unknown patterns among SDGs, but rather to highlight novel paths towards attaining informative patterns. Our consilient approach was conceived in anticipation of infinitely many challenges that require data-driven solutions across the 17 SDGs and, particularly, our original ideas of the DSF [15, 16] that views each SDG as a source of Big Data. Success will come from sharing data, skills and resources and ensuring that open science became the norm. One of the most difficult tasks of this work was to collate the data attributes in Table 1 from the main database of SDG indicators, as the initial variable selection is problem-specific. The findings in this section should open new paths to interdisciplinary research for a unified understanding of the triggers of SDG indicators.

Concluding remarks

The paper proposed a robust machine learning approach to data segmentation, constituting what can be viewed as a consilient approach to expounding triggers of SDG indicators via interdisciplinary modelling. It examined a range of tools that have been developed to provide SDG visualisation [13, 14] which while they capture key metrics on indicators, they leave the triggers of those indicators buried in data. Using selected SDG indicators it fulfilled objective #1 by illustrating the impact of data randomness and variation through visual objects.

The analyses exhibited potential knowledge gaps that may arise from including or excluding different data attribute. On objectives #2 and #3 it underlined interdisciplinarity in identifying actual and potential triggers-a major step towards attaining agenda 2030. Understanding the overall behaviour and development of SDGs requires taking a much broader perspective than just exploring individual SDGs. While interdisciplinarity may not have significantly featured in this work, the strong correlation among the SDGs and their span across disciplines and sectors imply that future applications of the proposed methods will adopt more interdisciplinary approaches.

While the number of SDGs and indicators used in this paper may not represent highly voluminous data, the indicators and targets in the entire set of 17 SDGs do. Apparently, the patterns in Figures 1 through 7 could have been fundamentally different if different sets of indicators, different SDGs or different countries had been used in the analyses. It is that level of intricacy that calls for interdisciplinary approaches in addressing SDG and underlines the role of objective #2 in SDG modelling. All the three objectives in the paper sought to promote interdisciplinary research. Objective #1 provided visualisation, while objectives # 2 and #3, effectively, focused on a unified, interdisciplinary understanding of the visual images and together, they highlighted the importance of clear definition of the problem to be tackled.

Through objectives #2 and #3, the paper calls for SDG monitoring teams to realise that attainment of agenda 2030 hinges on looking at all 17 SDGs as a Big Data challenge. The proposed methods are readily upscalable to higher data volumes. The paper's main output was not a pack of triggers, but a robust, unified approach driven by the three objectives. Data scientists may be content with the technical output of the adopted method, but it is combining domain knowledge and the power of modelling that yields reliable results. Agenda 2030 hinges heavily on this combination and future research paths should focus on it.

Abbreviations

BDMSDG: Big Data Modelling of Sustainable Development Goals; DSF: Development Science Framework; EDA: Exploratory Data Analysis; PCA: Principal Component Analysis; RDBMS: Relational Database Management Syststem; SDG: Sustainable Development Goals; SDGCA: Sustainable Development Goals Centre for Africa; SDGI: Sustainable Development Goals Indicators; SMA: Sample-Measure-Assess; UN-SDG: United Nation's Sustainable Development Goals.

Acknowledgements

This paper is part of on-going initiatives towards Big Data Modelling of Sustainable Development Goals (BDMSDG). We would like to thank many individuals and institutions who have discussed these initiatives with us at different stages of development. We particularly acknowledge the Data Intensive Research Initiative of South Africa (DIRISA), through the Council for Scientific and Industrial Research (CSIR) (https://www.csir.co.za/), who have invited us a couple of times to Pretoria to present our findings. We are also grateful to the Joint Support-Center for Data Science Research (DS) and the Polar Environment Data Science Center (PEDSC) of Japan (http://pedsc.rois.ac.jp/en/), the United Nations World Data Forum (UNWDF) (https://unstats.un.org/unsd/undataforum/index.html) and the University of Sussex's Sustainability Research Programme (https://www.sussex.ac.uk/ssrp/research/sdg-interactions). We also thank Sheffield Hallam University (https://www.shu.ac.uk/research/industry-innovation-research-institute) for meeting Springer's Gold Open Access publication costs.

Authors' contributions

The individual contributions of the authors heavily overlapped, from previous joint work on SDGs. For this work, KS carried out most of the data cleaning and automated selection, estimated at approximately 40%. IM provided insights into designing the analyses layout based on his experiences with the education system in the Republic of Rwanda and particularly his connections at the SDG Centre for Africa in Kigali, Rwanda [30], estimated at about 30% and BG carried out a background study on the impact of SDG initiatives within the Southern African Development Community (SADC) member states, in which he has previously published work-estimated at 30%.

Funding

This work has not been supported by any grant, but rather it is an outcome of ordinary Research and Scholarly Activities (RSA) allocation to each of the three authors by their respective institutions.

Data availability

As noted in "Data sources" section, raw secondary data attributes on hundreds of SDG indicators in six African countries—Botswana, Cameroon, Ghana, Kenya, Rwanda & South Africa, were extracted from the United Nation's Sustainable Development Indicators repository [2]. The data only relate to SDGs # 1, 2, 3, 4 and 9, covering the period 2000 to 2018. The source of the raw SDG indicators [2] is open-source, hence freely available to the public upon request. However, the data attributes used in this paper, were obtained via a semi-automated selection and cleaning process by the authors. They were reformatted to fit in with the adopted modelling strategy-hence, the data is only available from the authors, who have retained both the raw and modified copies, should they be requested.

Competing interests

All the three authors declare that there are no competing interests in publishing this paper, be they financial or non-financial and, as a co-authored paper, the costs of publishing will be shared among the three institutions.

Author details

[1] College of Business, Technology and Engineering, Sheffield Hallam University, Sheffield, United Kingdom. [2] Ministry of Education, Kigali, Republic of Rwanda. [3] Department of Applied Information Systems, University of Johannesburg, Johannesburg, South Africa.

Received: 19 May 2020 Accepted: 1 November 2020
Published online: 11 November 2020

References

1. SDG, Sustainable Development Goals. 2015; https://www.un.org/sustainabledevelopment/sustainable-development-goals/
2. SDGI, Sustainable Development Goals Indicators. 2017; https://unstats.un.org/sdgs/indicators/database/
3. Kharrazi A. Challenges and opportunities of urban big-data for sustainable development. Asia Pacific Tech Monitor. 2017;34:17–211.
4. Kruse CS, Goswamy R, Raval Y, Marawi S. Challenges and opportunities of big data in health care: a systematic review. JMIR Med Inf. 2016;4:e38.
5. Yan M, Haiping W, Lizhe W, Bormin H, Ranjan R, Zomaya A, Wei J. Remote sensing big data computing: challenges and opportunities. Future Gener Comput Syst. 2015;51:47–60.
6. IUCN, In the spirit of nature, everything is connected. 2018; https://www.iucn.org/news/europe/201801/spirit-nature-everything-connected
7. Mwitondi KS. Tracking the Potential, Development, and Impact of Information and Communication Technologies in Sub-Saharan Africa; International Council for Science (ICSU-ROA);2018
8. Meusburger P. In Knowledge and the Economy; Meusburger, P., Glückler, J., el Meskioui, M., Eds.; Springer Netherlands: Dordrecht, 2013; pp 15–42
9. Parr M, Musker R, Schaap B. GODAN'S Impact 2014 to 2018 - Improving Agriculture, Food and Nutrition with Open Data;2018
10. UN-Global-Pulse, Big Data for Development: Challenges and Opportunities.UN Global Pulse. **2012**
11. UN-Global-Pulse, Big Data for Development and Humanitarian Action: Towards Responsible Governance. 2016
12. Bamberger M. Integrating Big Data Into the Monitoring and Evaluation of Development Programmes. **2016**
13. Roser M, Ortiz-Ospina E, Ritchie H, Hasell J, Gavrilov D. Our World in data: Research and interactive data visualizations to understand the world's largest problems;2018
14. WBGroup, Atlas of Sustainable Development Goals From World Development Indicators. 2018
15. Mwitondi K, Munyakazi I, Gatsheni B. An interdisciplinary data-driven framework for development science. DIRISA National Research Data Workshop, CSIR ICC, 19-21 June 2018, Pretoria, RSA2018
16. Mwitondi K, Munyakazi I, Gatsheni B. Amenability of the United Nations Sustainable Development Goals to Big Data Modelling. International Workshop on Data Science-Present and Future of Open Data and Open Science, 12-15 Nov 2018, Joint Support Centre for Data Science Research, Mishima Citizens Cultural Hall, Mishima, Shizuoka, Japan**2018**
17. Ishikawa K. Guide to auality control; Asian Productivity Organization;1976
18. Primmer E, Furman E. Operationalising ecosystem service approaches for governance: do measuring, mapping and valuing integrate sector-specific knowledge systems? Ecosyst Serv. 2012;1:85–92.
19. Mwitondi KS, Said RA. A data-based method for harmonising heterogeneous data modelling techniques across data mining applications. J Stat Appl Probab. 2013;2(3):293–305.
20. Mwitondi KS, Moustafa RE, Hadi AS. A data-driven method for selecting optimal models based on graphical visualisation of differences in sequentially fitted ROC model parameters. Data Sci J. 2013;12:WDS247–WDS253.
21. SDGTI, Sustainable Development Goals Targets & Indicators. 2020; https://unstats.un.org/sdgs/metadata/
22. Lloyd SP. Least squares quantization in PCM. Technical Report RR-5497, Bell Laboratories. 1957.
23. MacQueen JB. Some methods for classification and analysis of multivariate observations. 1967;1:281–97.
24. Chapmann J. Machine learning algorithms; CreateSpace Independent Publishing Platform, 2017
25. Kogan J. Introduction to clustering large and high-dimensional data. Cambridge: Cambridge University Press; 2007.
26. Mwitondi KS, Zargari SA. An iterative multiple sampling method for intrusion detection. Inf Secur J. 2018;27:230–9.
27. Bo L, Wang L, Jiao L. Feature scaling for Kernel Fisher discriminant analysis using leave-one-out cross validation. Neural Comput. 2006;18:961–78.
28. Galkin F, Aliper A, Putin E, Kuznetsov I, Gladyshev VN, Zhavoronkov A. Human microbiome aging clocks based on deep learning and tandem of permutation feature importance and accumulated local effects. BioRxiv. 2018;1:507780.
29. Codd EF. A relational model of data for large shared data banks. Commun ACM. 1970;13:377–87.
30. SDGCA, Sustainable Development Goals. 2015; https://sdgcafrica.org/

Publisher's Note

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