PROFESSUR FÜR ANGEWANDTE STATISTIK

DER FREIEN UNIVERSITÄT BERLIN

Seminararbeit

Application of Multiple Factor Analysis on the Sustainable Development Goals Dataset for African Countries

Gutachter(in):

Verfasser:

Matrikel-Nr.:

Adresse:

Email:

Telefon:

Studienfach:

Abgabetermin: 28. Juli 2021

Table of Contents

List of Figures

List of Abbreviations

UN	United Nations
SDG	Sustainable Development Goals
MFA	Multiple Factor Analysis
PCA	Principal Component Analysis
PC	Principal Component

1 Introduction

As of 2021 there are still glaring disparities concerning the living conditions throughout the world. While humans born in wealthy countries like Hong Kong, Japan or Switzerland have a life expectancy of around 85 years, humans born in some of the least developed countries like Central African Republic or Chad only have an expectancy of just under 55 years (World Bank, 2021). When it comes to education, the aforementioned countries Chad and Central African Republic again rank among the worst in the world, as evident from looking at the Education Index measuring mean schooling years before the age of 25. While some of the highest ranked countries like Germany or Norway achieve scores above 0.9 on the index, some of the poorest African countries show scores below 0.3. Comparing the gross domestic product between countries to capture a decent standard of living unsurprisingly paints the same picture. The Human Development Index by the United Nations (UN) combines these aspects into one score and shows significant deficiencies for some countries, particularly highlighting the continent of Africa as problematic (United Nations, 2020).

For this reason, the United Nations General Assembly initiated the 2030 Agenda for Sustainable Development, which was adopted by all United Nations Members in 2015. Core to it are 17 Sustainable Development Goals (SDGs), with 169 targets, following on from the previous Millennium Goals of trying to eradicate all forms of poverty. Encompassing economic, social, and environmental factors the agenda is supposed to provide a "… shared blueprint for peace and prosperity for people and the planet, now and into the future" (United Nations, 2015). The goals "… seek to realize the human rights of all and to achieve gender equality and the empowerment of all women and girls" (United Nations General Assembly, 2015).

1.1 Motivation

As stated in the resolution, uneven progress on the Millennium Goals was achieved, as especially African and least developed countries remained off target (United Nations General Assembly, 2015). The increased emphasis on these countries in the new agenda for 2030 provides us with the incentive to make an assessment of the current situation and provide indications as to which countries are on track or which may need more support. By applying a multiple factor analysis (MFA) on a selection of 20 indicators from 7 goals regarding the most basic human needs, we aim to explain how the African countries stack up in their development.

MFA as an extension of principal component analysis (PCA), is a method from Multivariate statistics with the purpose of summarizing and visualizing data tables with continuous and categorical variables structured into groups (Pagès, 2004, p.1). MFA allows us to use quantitative data from several different goals, while also focussing on qualitative aspects. These qualitative aspects comprise of regional differences within Africa as well as differences between countries which are on the list of least developed countries and countries which are officially labelled as developing. This enables insightful visualization of the 20 quantitative and 2 qualitative variables for 44 African countries and allows interpretation. We want to utilize this to make a general assessment of the development of the African countries relative to each other and provide an indication of which countries are in danger of not achieving the goals stated in the 2030 agenda.

1.2 Structure of the Paper

This paper is structured as follows. Section 2 contains a general explanation of MFA and PCA and the underlying mathematical theory. In Section 3, we describe the Sustainable Development Goals dataset, our process of selecting goals and indicators, and our data pre-processing. Section 4 is the main focus of our paper, where we show the results of our MFA and interpret our findings. Section 5 concludes.

2 Theory of Multiple Factor Analysis and Principal Component Analysis

When dealing with multivariate data a problem frequently occurs, which is widely referred to as the curse of dimensionality (Bellman, 1966). In essence, it describes a problem with high-dimensional data: it contains too many variables for graphical illustration and easy interpretation and inhibits other multivariate techniques. If we take a data set with two quantitative characteristics describing our observations, we can easily visualize our data in two-dimensional graphs. For example, we could create a scatterplot where each observation (human) is represented by a dot depending on its values regarding the two characteristics height and weight as in Figure 2.1. This provides information about the correlation and shows outliers.

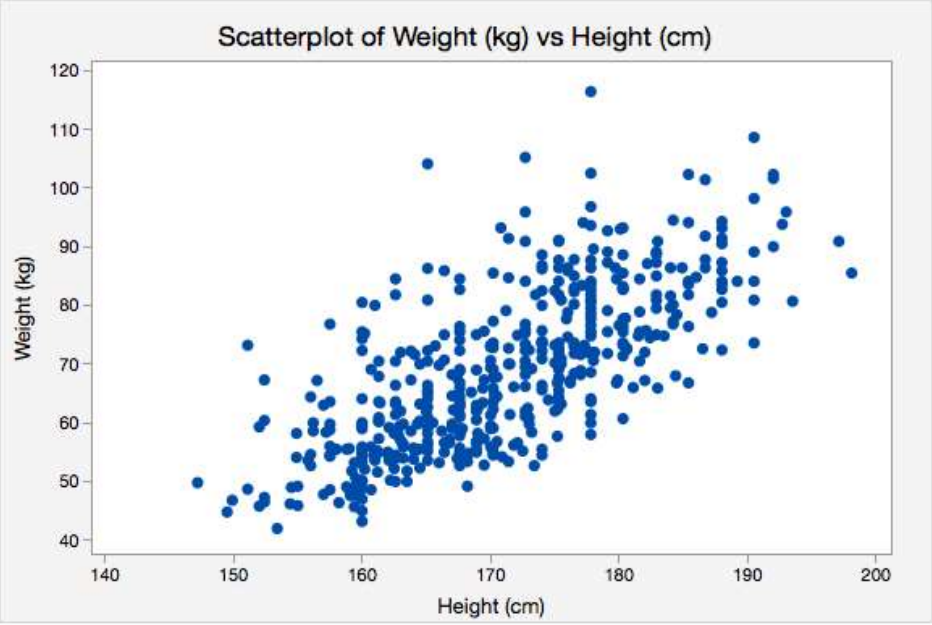

Figure 2.1: Scatterplot of Height against Weight (Source: online.stat.psu.edu)

If we added a third descriptional characteristic such as shoe size to our example, we could still display the data by coding the dots in different colors or shapes or by creating a three-dimensional graph. If we added more variables however, these techniques fail. These issues with big data can be resolved by reducing the dimensionality of the data using PCA. Whereas PCA is applied to solely continuous data, MFA pursues the same goal for mixed data including categorical data,

where the data is structured into groups (Pagès, 2004, p.1). These methods find linear combinations of the original variables, which maintain as much variation of the dataset as possible while reducing the number of dimensions and therefore eliminating interrelated variables. They can therefore be classified as an unsupervised machine learning type of algorithm. The results of a PCA or an MFA are a new set of variables called principal components (PC's), which are a smaller set of uncorrelated and ordered variables. This eases summarization of the data and enables two-dimensional graphical representation and interpretation thereof.

2.1 Mathematical Derivation of Principal Component Analysis

We can write any set of data as a matrix, where the variables $x_1, \dots, x_p$ are the columns and the n observations are the rows. If $\boldsymbol{x}$ denotes a vector of p random variables and α_k denotes a vector of constants, then $\alpha'_k x = \sum_{j=1}^{p} \alpha_{kj} x_j$. In PCA, we are looking for the linear function $\alpha'_1 x$ of the vector of variables $\boldsymbol{x}$ with the highest possible variance. This process is then repreated to find a second linear function $\alpha'_2 x$ of $\boldsymbol{x}$ with the highest possible variance, but under the condition that it is uncorrelated with the first one. This is iterated for the number of variables p, but the goal is to end up with $< p$ principal components, that retain as much variation in the dataset as possible.

To execute this, the first step of a PCA or MFA is usually to calculate the mean for every variable to determine the covariance matrix Σ for the data set. A covariance matrix is a matrix with the variances of the variables on the main diagonal and the covariances between the different variables arranged otherwise. With Σ we can determine pairs of eigenvectors α and eigenvalues λ and sort them in descending order according to the value of the eigenvalue.

PCA can be performed using a correlation matrix or a covariance matrix. If the data uses variables which are not scaled equivalently, it is necessary to standardize the data by subtracting the mean from each value and scaling the unit variance to one. Otherwise, the first principal components will be dominated by variables with high variance (Jolliffe, 2011).

Mathematically, there are some conditions that need to be satisfied. According to the variance condition, the $Var(z_k) = \alpha'_k \Sigma \alpha_k$ for each $k = 1, \dots, p$ variables. Due to the data being centered ($E[x] = 0$), the formulas for the variance and covariance are

$$Var(x_1) = E[x_1^2] - 0$$

and

$$Cov(x_{1,}x_2) = E[x_1x_2] - 0$$

$$\Rightarrow Cov(x) = E[xx']$$

which in terms of the $Var(z_k)$ for our linear function results in

$$Var(z_k) = E[({\alpha'}_k x)^2] = E[({\alpha'}_k x)(x'\,\alpha_k)] = E[{\alpha'}_k (xx')\,\alpha_k]$$

$$= {\alpha'}_k E[xx']\,\alpha_k = {\alpha'}_k Cov(x)\,\alpha_k$$

$$= {\alpha'}_k \Sigma\,\alpha_k. \qquad (1)$$

The covariance condition $Cov(z_i, z_j) = \alpha'_i \Sigma a_j = 0$ ensures that the different principal components are uncorrelated with each other. Lastly, the unit vector condition states that ${\alpha'}_k \alpha_k = 1$. Otherwise, the variances of new variables could be made arbitrarily large. The PCA now unfolds as an optimization problem, where we maximize the variance ${\alpha'}_k \Sigma\,\alpha_k$ under the constraint of the unit length condition:

$$\mathcal{L}(\alpha_k, \lambda_k) = {\alpha'}_k \Sigma\,\alpha_k - \lambda_k({\alpha'}_k \alpha_k - 1) \qquad (2)$$

The first order conditions for the Lagrangian are

$$\frac{\partial}{\partial \alpha_k}\,\mathcal{L}(\alpha_k, \lambda_k) = 2\Sigma\alpha_k - 2\lambda_k\alpha_k \overset{!}{=} 0 \qquad (3)$$

$$\frac{\partial}{\partial \lambda_k}\,\mathcal{L}(\alpha_k, \lambda_k) = 1 - {\alpha_k}'\,\alpha_k \overset{!}{=} 0\,. \qquad (4)$$

Vector α_k represents an extremum if both conditions are satisfied. The second constraint can be rewritten as

$${\alpha'}_k \alpha_k = 1. \qquad (5)$$

Simplyfing the first constraint

$$2\Sigma\alpha_k - 2\lambda_k \alpha_k = 0 \qquad (6)$$

as

$$\Sigma\alpha_k = \lambda_k\alpha_k \qquad (7)$$

yields an eigenvector equation, where α_k is an eigenvector of the covariance-matrix Σ and λ_k is the associated eigenvalue. The eigenvector α of the square matrix Σ is a non-zero vector that only changes by a scalar factor λ, when that matrix is applied to it

$$\Sigma\alpha = \lambda\alpha \Leftrightarrow (\Sigma - \lambda I)\alpha = 0 \qquad (8)$$

and the scalar λ is called eigenvalue. The equation has a solution α if the determinant of the matrix $(\Sigma - \lambda I)$ is zero. Next, we look for the eigenvector with the linear combination that has the highest variance. If equations (5) and (7) are satisfied, it follows that

$$\alpha'_k \Sigma\, \alpha_k = \alpha'_k \lambda_k \alpha_k = \lambda_k \alpha'_k \alpha_k = \lambda_k$$

and it becomes apparent that the eigenvalue λ_k of the corresponding eigenvector α_k is the variance of the linear combination $\alpha'_k x$. Thus, simply choosing the eigenvector α which belongs to the highest eigenvalue λ gives us the solution to the first principal component of $\boldsymbol{x}$. They receive the Index 1, so that $\Sigma\alpha_1 = \lambda_1\alpha_1$ is the first principal component. All other pairs of eigenvalues and eigenvectors are ordered in descending order. They are determined with an additional constraint, as they must be uncorrelated with the previous components according to the covariance condition. So, for the second principal component, this constraint can be denoted as

$$cov(\alpha_1' x,\ \alpha_2' x) = \alpha_1' \Sigma\alpha_2 = \alpha_2' \Sigma\alpha_1 = \lambda_1\alpha_1' \alpha_2 = 0.$$

Solving the Langrangian again with this additional constraint regarding the covariance yields the second principal component, with the second highest eigenvalue λ_2 and its eigenvector α_2, which are orthogonal to the first eigenvector. This is repeated for $k = 1 \ldots p$ with p different combinations of eigenvectors of Σ and eigenvalues $\lambda_1, \ldots, \lambda_p$.

To briefly summarize the procedure, to find a linear combination $\alpha'_k x$ with the highest variance, we needed to calculate the eigenvalues and eigenvectors of the covariance matrix Σ. The eigenvector α_1 corresponding with the highest eigenvalue λ_1 determines the weighting for the linear combination. The second highest eigenvalue λ_2 and eigenvector α_2 are orthogonal to the first one.[1]

2.2 Additional considerations in MFA

In MFA, as in PCA, it is important to balance the influence of variables. In PCA, this is done to prevent variables with greater variance from having more influence in the analysis. Similarly, in the case of MFA, groups of variables need to be balanced to avoid a uni-dimensional group from influencing the analysis too strongly and hiding the information contributed by multidimensional groups (Pagès, 2015).

To achieve an equilibrium, we divide the values of a variable by the square root of their group's largest eigenvalue:

$$\frac{X_j}{\lambda_1^j}, where\ X_j\ corresponds\ to\ the\ jth\ standardized\ group\ of\ variables$$

In this way, each group makes the same contribution to the construction of the MFA dimensions and the richness of information provided by multidimensional groups is retained.

[1] The mathematical derivation and explanation are not part of our own work. We refer to the course Multivariate Verfahren at FU Berlin as well as the book Multivariate Analysemethoden (2017) by Andreas Handl and Torben Kuhlenkasper (pp.126-129). We used the notation from the lecture slides from the course. Some phrasing was quoted verbatim to keep the articulation concise.

3 Data Set and Data Preparation

The Sustainable Development Goals by the United Nations comprise 17 goals covering various economic, social, and environmental factors. Each goal consists of multiple targets with one or more corresponding indicators. The indicators contain the data describing a country's development with regards to that target and can be compared over time or against other countries.

Each year the United Nations prepares a report on the progress towards the SDGs. To that end, it keeps a publicly accessible database, known as the Global SDG Database, with data for each indicator. Due to the immense volume of the data, the completeness and the nature of reporting varies between indicators, countries and over time. While some indicators are determined each year for all the countries, others only have very sporadic or close to no entries. In terms of the nature, the data includes country data and country adjusted data, estimated data, globally monitored data and modelled data. Some indicators additionally have differentiated entries for different age groups, sex and for rural or urban areas.

3.1 Selection of Goals and Indicators

The Global SDG Database has been compiling data since 1959 for 17 goals, 169 targets and several hundred indicators. Given the vastness of the data, we focused on a smaller set of goals and indicators for a discerning analysis of sustainable development in Africa. Choosing the goals and indicators was an iterative process involving prioritization based on the perceived importance and completeness of the data.

As many countries on the African continent have a particularly low score on the Human Development Index, goals connected to the most fundamental human needs were shortlisted: Goals 1 (No Poverty), 2 (Zero Hunger), 3 (Good Health and Well-being), 4 (Quality Education), 6 (Clean Water and Sanitation), and 7 (Affordable and Clean Energy). Inspection of the data revealed many indicators with missing values and redundant information. For each indicator, the most recent year with the most complete data was prioritized and to it an arbitrary boundary of 75% data completeness was set. Indicators outside of this boundary were excluded. The shortlist was thereby reduced to 4 goals, 20 indicators and 44 countries. In the final data set,

missing values were replaced with the column median. The median was preferred over the mean due to the high variable variance of certain indicators.

3.2 Creation of Variable Groups

As explained in Section 1, variables are grouped in MFA. It was intuitive to organize variables (indicators) by their respective goals. This produced four groups: "Poverty", "Health", "Sanitation" and "Energy". In addition, one group was created manually from indicators belonging to different goals. This artificial group labelled "Development Assistance" encompasses variables pertaining to the amount of monetary foreign aid received by the country. This group was deliberately pieced together to allow for the inclusion of information about foreign aid, which is highly relevant in the assessment of a country's sustainable development. After all, many African countries do not have the means to drastically improve conditions by themselves and are dependent on financial assistance from abroad. We would generally expect countries in a poor state of development to be receiving more assistance to improve the situation.

Correlation matrices were created to visualize variable relationships. Generally, high intra-group correlation was shown and at times high inter-group correlation. Especially noteworthy was the extremely high correlation of the Sanitation and Energy groups. Given this relationship, the two groups were combined into one, labelled: "Infrastructure".

Finally, a supplementary group "Regions" composed of two categorical variables was created. The first variable specifies to which region in Africa an individual country belongs according to the United Nations geoscheme for Africa. The second one indicates whether the country is officially classified as a Developing Country or Least Developed Country by the UN.[2] A comprehensive table with all groups, indicators and years can be found in Appendix A.

[2] The List of Least Developed Countries can be found at: https://www.un.org/development/desa/dpad/wp-content/uploads/sites/45/publication/ldc_list.pdf (last accessed July 28, 2021).

4 Results and Analysis

We conducted our analysis using the statistical software R. A first look at the Scree Plot depicting explained variances of the dimensions in the MFA is encouraging. As shown in Figure 4.1, the first two principal components explain 60.46% of the variance present in the data set. The first component, with an eigenvalue of 2.72, explains 43.25% of the variance and the second, with an eigenvalue of 1.08, explains 17.22%. After that a significant drop-off can be observed, with eigenvalues ≤ 0.51 and explained variance $\leq 8\%$ for the third principal component onward.

This poses the question: how many principal components are required to appropriately reproduce the data? Usually, the goal is to explain 70-90% of the total variation of the dataset. In our analysis four dimensions would lead to 75.43% cumulative variance, five to 80.56%. A different criterion proposed by Kaiser (1960) suggests using all principal components with eigenvalues larger than the average eigenvalue. For our model this would be 0.314. As the eigenvalue of Dimension 5 is 0.322, it would be reasonable to retain the first five principal components and thus explain 80% of the total variance in the data set. Our analysis is focused mainly on the first two PC's, as this allows us to plot the individual countries and interpret their position in the graph.

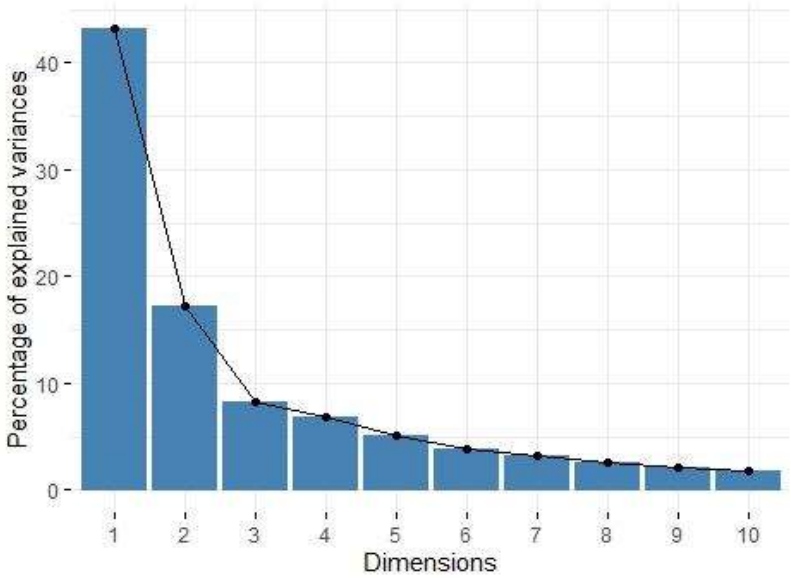

Figure 4.1: Screeplot of the Explained Variances Against the First 10 Dimensions

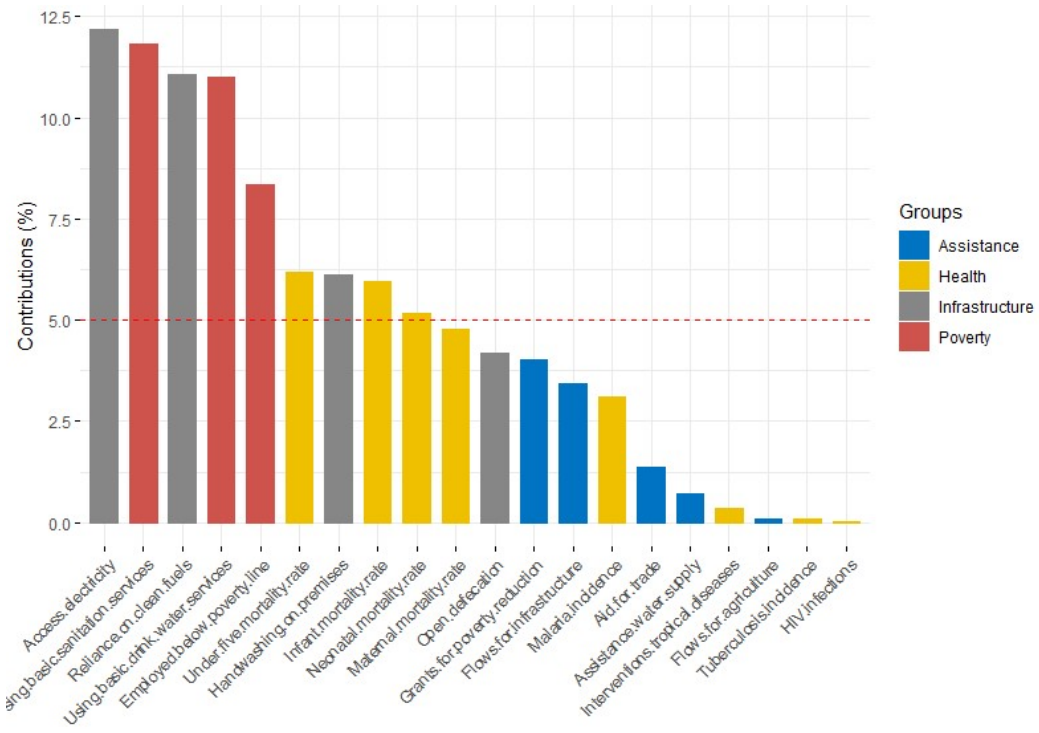

Figure 4.2: Screeplot of the Contribution of Variables to Dim-1

Focusing on Dimension 1 in Figure 4.2 reveals which variables and groups contribute the most to it. It appears to be dominated by the Infrastructure and Poverty groups; the variables measuring access to electricity and the proportion of population using basic sanitation services contribute the most. Variables from the Health group connected to the mortality rate at birth also have some influence. The first PC seems to encompass the curret state of development of African countries with regards to basic living conditions.

In contrast, the second Dimension (see Appendix B) is dominated by the variables from the Development Assistance group, with some contributions from the remaining health variables uncorrelated to the first PC. The second PC appears to mostly depicts how much financial aid the countries are receiving to improve these conditions.

Observations from Figure 4.2 are reflected in the Variable Groups graph (Figure 4.3). The graph quantifies the link between each variable group and the first two PC's. As the Poverty and

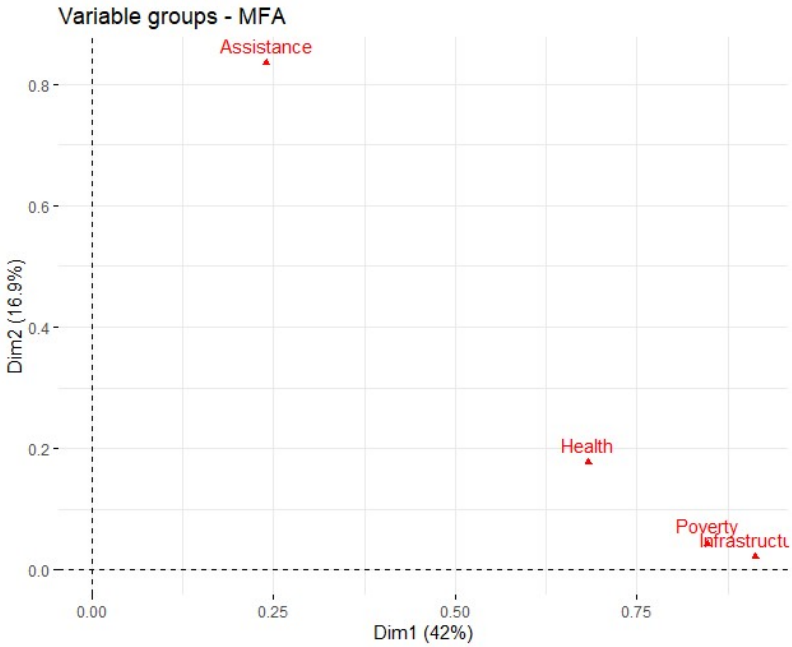

Figure 4.3: Contribution of the Groups of Variables to the First and Second Principal Component

Infrastructure groups have an x-axis value close to 1, they are strongly linked to the first dimension. The Development Assistance group has a value close to 1 on the y-axis, meaning that it is closely linked to the second dimension. The Health group, in turn, is more multidimensional. While it is more strongly aligned with the first dimension, it is further from the extremes of both axes.

The Variable Groups graph additionally visualizes similarity via spatial proximity. The closeness of the Poverty and Infrastructure coordinates suggests that these two groups are broadly similar.

4.1 Variables Plot

Graphing the variables in a circle of correlation reveals information about the relationships among them (Figure 4.4). The closer together two variables appear in the graph, the greater their correlation. Variables with a 90° angle between them are uncorrelated and variables that oppose each other are negatively correlated. Likewise, the correlation between the variables and the first two principal components can be observed from their distance to the x-axis and y-axis. The circle of correlation also visualizes the quality of representation of each variable, that is, how well the

two dimensions of the MFA capture the variance of a variable. The longer the vector, the better the quality of representation.

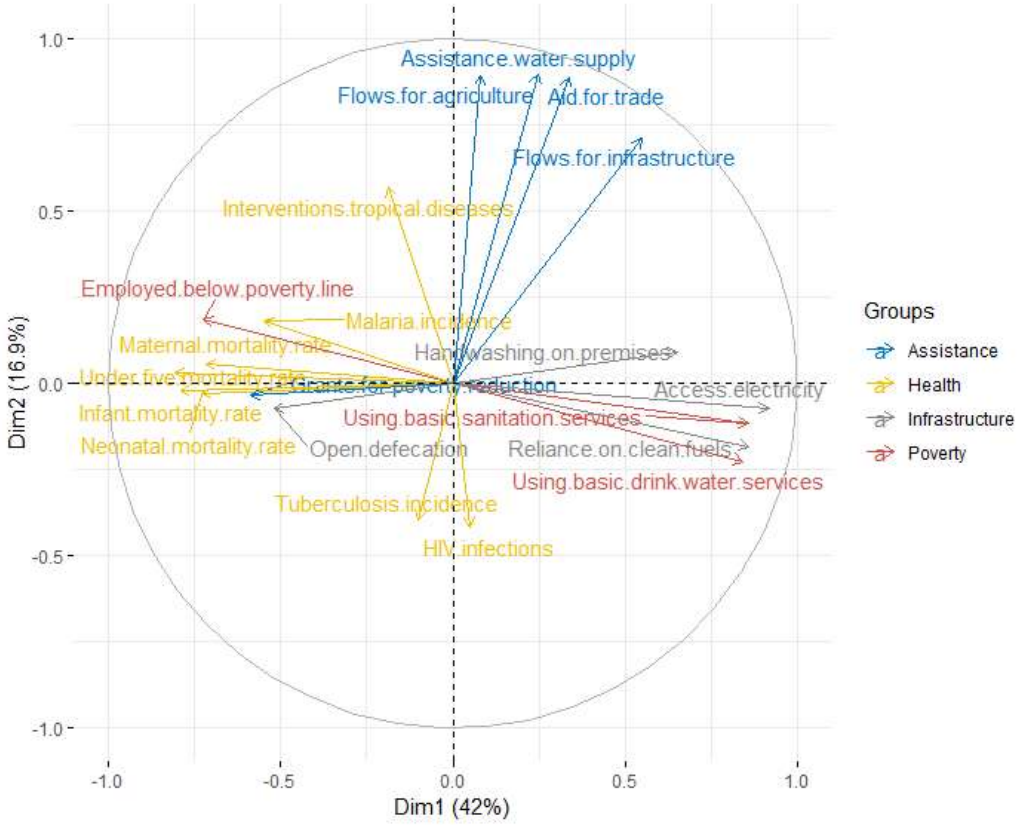

Figure 4.4: Correlation between Quantative Variables and Dimensions

Beginning with the Poverty group, a strong correlation can be observed between two of its variables. The proportion of population using basic sanitation services and basic drink water services are highly correlated and they both have a strong negative correlation with the variable 'Employed population below the international poverty line'. This relationship is not suprising, as for the first two variables a high percentage indicates better human development whereas for the last variable, a low value does. The high quality of representation of the Poverty variables increases confidence in the accuracy of the correlation.

Turning to the Infrastructure group, a similarly strong correlation between the variables is depicted. Access to electricity, access to handwashing facilities on premises and reliance on clean fuels are highly positively correlated, and all three variables have a strong negative correlation

with practicing open defecation. This relationship again intuitively makes sense, as one would expect a person with access to decent infrastructure to have plumbing and therefore not defectate openly.

Variables from both of those groups are also highly correlated with each other. This confirms the intuition that people with access to electricity will also have acces to basic drink water and sanitation services and vice versa.

Overall, the Poverty and Infrastructure groups stand out for being highly correlated with the first dimension of the MFA. The exact correlation between the first dimension and the Poverty group is 0.92, and even 0.96 between the first dimension and the Infrastructure group (see Appendix A). This finding is not surprising, as the Scree Plot in Figure 4.2 showed that variables from both groups contributed the most to the first dimension.

Following from the above observations, the first dimension of the MFA (the x-axis of the circle of correlation) can be interpreted as constituting a combined measure for the basic living conditions of a country. To interpret the direction: a country that is positioned further to the right in the graph has better basic living conditions than countries further to the left.

Returning to the group analysis, in contrast to the from the Poverty and Infrastructure groups, the Health group is more distributed between dimensions and an interpretation of it is more ambiguous. On the one hand, variables relating to mortality rates are highly correlated amongst each other and with the first dimension. A possible interpretation is that higher levels of poverty (i.e. worse living conditions) are positively correlated with higher mortality rates around the time of birth. One might speculate, for example, that pregnant women living below the poverty line do not go to hospitals to give birth, resulting in more deaths during delivery.

Secondly, the Health variables relating to illness and disease show mixed results. The Malaria variable appears positively correlated with the mortality Health variables and the "bad" infrastructure and poverty variables. This could suggest that Malaria is an illness that disproportionately affects deprived communities and perhaps is even a contributor to high infant mortality in those areas. Apart from the Malaria variable, however, the disease Health variables are mostly uncorrelated with the first dimension. While they appear correlated to the second

dimension, this correlation is limited, as is the variables' quality of representation. It is likely that different dimensions account for the relative positions of the HIV, tuberculosis, and tropical disease variables. More conclusive interpretations about the relationships between these variables is therefore not possible.

The fourth and final group, Development Assistance, shows a high positive correlation between four of its five variables. These appear negatively correlated from some Health variables, but as previously mentioned, due to the limited quality of representation of the HIV and tuberculosis variables, a negative correlation cannot be inferred with certainty.

Interestingly, the fifth Development Assistance variable (Grants for Poverty Reduction) is uncorrelated from the other Development Assistance variables. Instead, it is negatively correlated with the first dimension. As the first dimension represents the development of a country in terms of basic living conditions, it is reasonable that this is the relationship between this Development Assistance variable and the first dimension. In other words, the worse a country's basic living conditions, the more poverty reduction grants it should receive. Curiously, all other Development Assistance variables tip towards the correlation circle's right bisector, which represents countries that have somewhat better basic living conditions.

Despite one uncorrelated Development Assistance variable, given that all other group variables are highly correlated with the second dimension and additionally have a high quality of representation, the second dimension can be interpreted as a representation of how much financial aid a country receives.

4.2 Individuals Plot

Following the variable cloud analysis, an individual cloud analysis addresses the paper's original goal of visualizing African countries according to their level of sustainable development.

To reiterate, in Figure 4.5, countries further to the right have better basic living conditions, as they have better access to drink water services, sanitation services and electricity, and lower mortality rates at birth. How much a country is affected by certain illnesses also has a slight effect on its performance on the first dimension. In contrast, countries with a position higher up on the

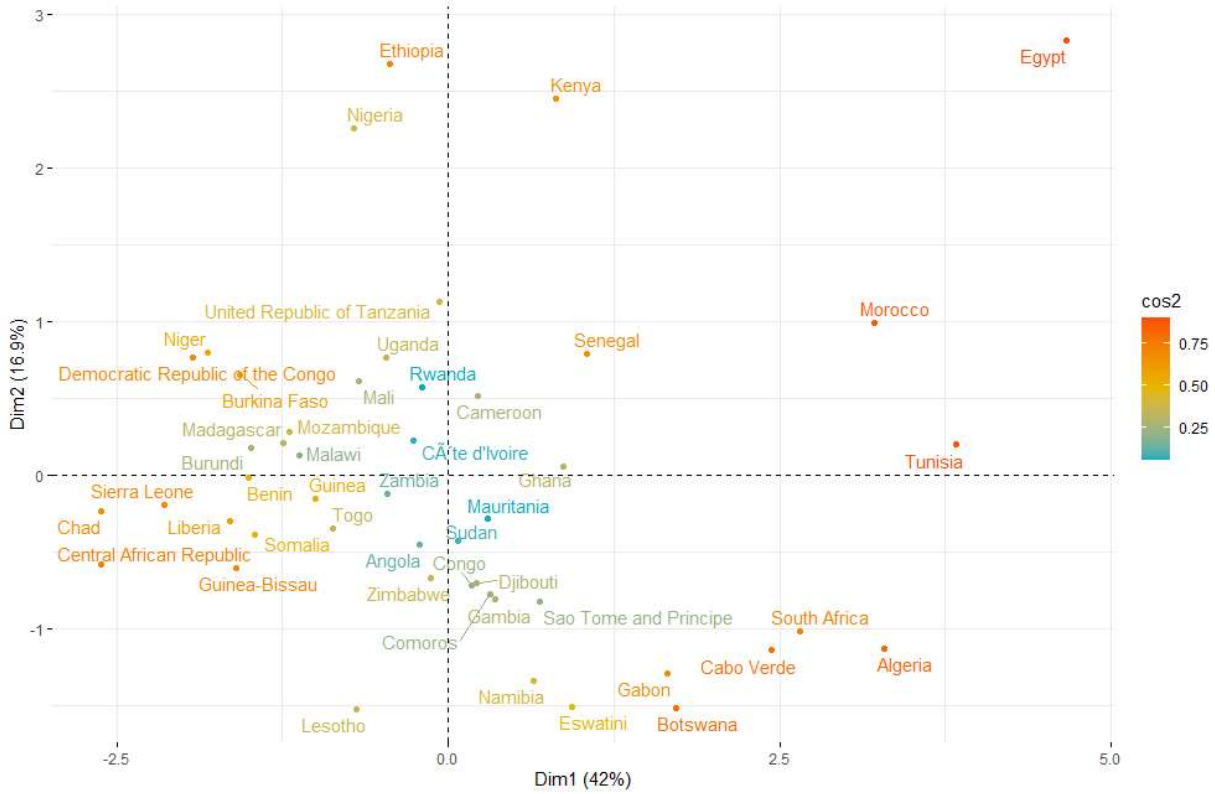

Figure 4.5: Individuals Plot Showing Quality of Representation

graph receive more development assistance for infrastructure, water supply, agriculture, trade and poverty reduction, although the prevalence of diseases such as HIV, tuberculosis and tropical diseases, can somewhat mitigate this.

According to the graph, Egypt, Tunisia, Algeria, and Morocco are the most developed countries, followed closely by South Africa. In contrast, Chad, Central African Republic, and Sierra Leone struggle the most. Furthermore countries with a higher cos^2 value illustrated in orange colour have the highest quality of representation in our graph.

The Individuals Graph in Figure 4.5 can be enriched by projecting information from the categorical variables onto it. The left graph in Figure 4.6 colors the Least Developed Countries (LDCs) in green and the Developing Countries in red. As expected, the mean LDC coordinate lies in the left bisector, whereas the Developing Country coordinate is found in the right bisector. Figure 4.6 furthermore reveals that the first dimension of the MFA almost perfectly separates

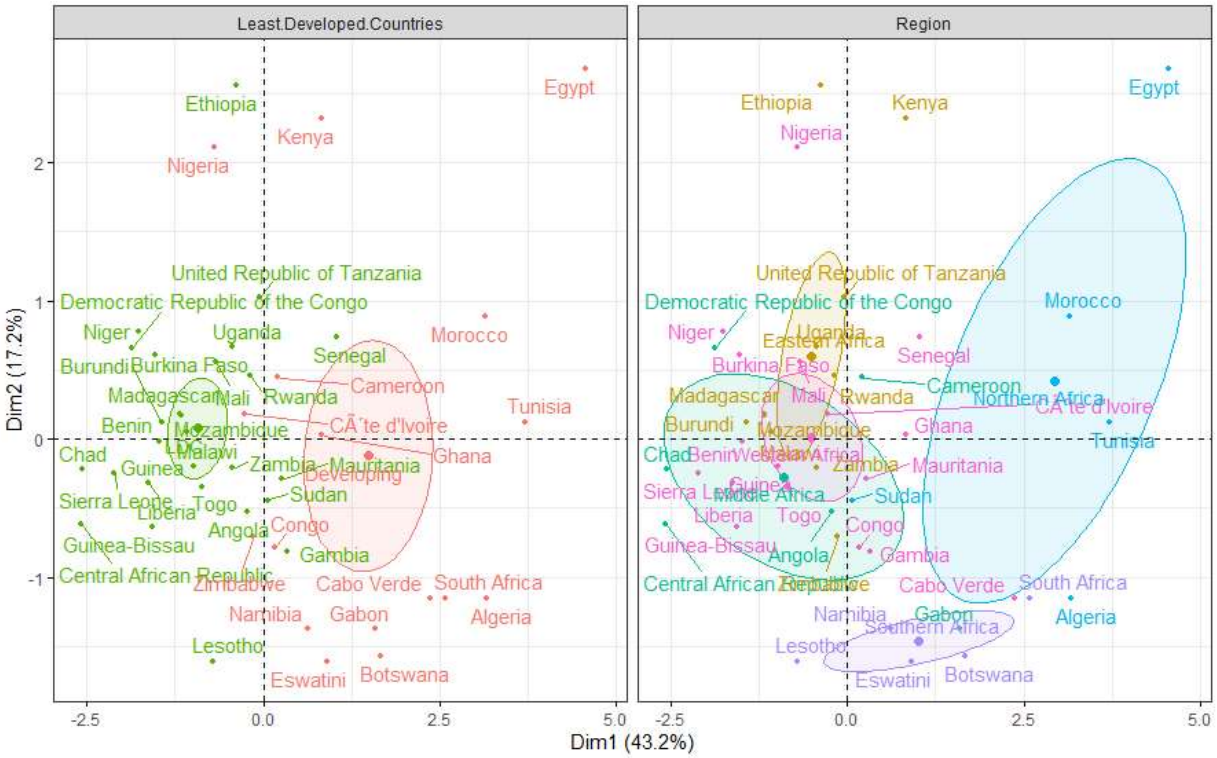

Figure 4.6: MFA factor map

LDCs from Developing Countries. As LDC is a supplementary variable that did not contribute to the construction of dimensions, its high correlation with the first dimension is noteworthy.

The right graph of Figure 4.6 groups the countries by region (Northern Africa, Southern Africa, Middle Africa, Eastern Africa, or Western Africa). This representation shows that almost all Northern and Southern African countries enjoy above average standards of living by African standards. Middle Africa is the region struggling the most. Given that many countries from Eastern and Western Africa have a low quality of representation on the plane, an analysis of these regions is not conclusive.

Turning to the second dimension, Egypt, Ethiopia, Kenya and Nigeria seem to be receiving the highest official development assistance, at least for the five indicators captured in the MFA. Egypt stands out for being exceptionally well projected on both dimensions, suggesting it is doing better than most other African countries in terms of basic livings standards and also receiving a lot of financial assistance. The combination seems paradoxical, as one would expect countries far to the left to receive the most aid. Actually, Egypt has always benefitted from a lot of foreign

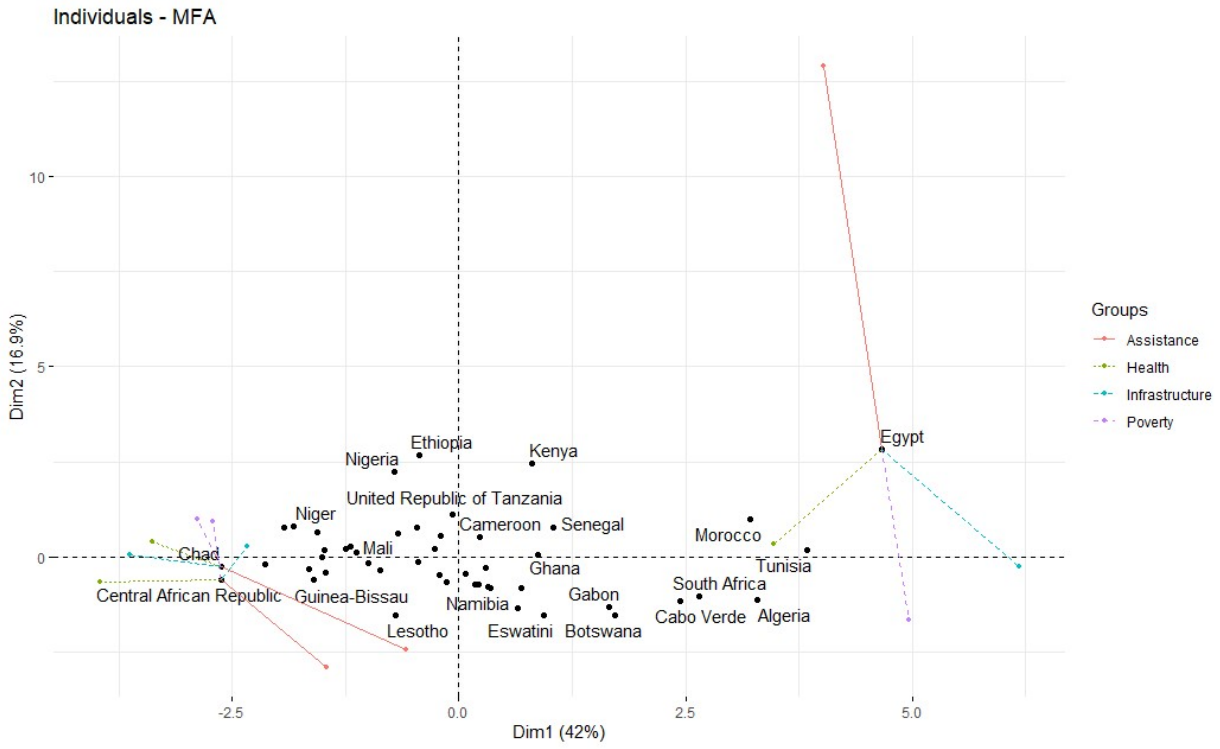

Figure 4.7: Partial Individuals Graph

assistance, especially from the USA: "Historically, Egypt has been an important country for U.S. national security interests based on its geography, demography, and diplomatic posture. ... Since 1946, the United States has provided Egypt with over $84billion in bilateral foreign aid" (Sharp, 2020).

The graph of partial individuals (Figure 4.7) visualizes the extent to which Egypt's center of gravity has been pulled upwards by the Assistance group. Without it, Egypt would be positioned far closer to the x-axis, given that the other partial individuals are lower on the graph.

Analyzing the individuals from a group perspective shows that Developing Countries are more varied in terms of the development assistance they receive (left graph in Figure 4.6). This range can be attributed to the Northern African countries, as Southern African countries are quite similar in the low amount of aid they receive. In fact, the Southern African region appears to be the most homogenous, with individual countries being closely grouped on both dimensions.

Overall, a surprising observation can be made: the amount of development assistance received by Least Developed Countries is only marginally higher than that received by Developing Countries. The graph of partial individuals (Figure 4.7) highlights this oddness. Chad and the Central African Republic are the least developed countries, particularly suffering in terms of Health indicators, yet the amount of aid they receive is even lower than suggested by their mean coordinate point.

5 Conclusion

Our goal in this paper was to assess how African countries are positioned relative to each other with regards to their development concerning basic living conditions and how much foreign assistance they receive to improve them. This was in order to gain an indication if the countries will likely meet the targets stated in the Sustainable Development Goals by the United Nations for 2030. We selected 20 indicators from 7 different goals for 44 African countries and added two qualitative variables for region and development status. After describing the theory of PCA and MFA and our process of data preparation, we performed an MFA on our data. In our analysis we managed to transform the 20 original variables into 5 new ones that retained 80% of the variation in our data set. The two-dimensional graphical representations of the data with 60% explained variance showed how Northern African countries generally are more developed and that some countries receive significantly more support to combat deficits. This surprisingly revealed that the countries with the worst conditions (particularly Middle Africa) don't receive more support. This could indicate that these countries might lag behind the targets set by the UN. Judging from our graphs we are inclined to believe that countries like Chad and Central African Republic will need more aid to accelaerate their development to ultimately meet the SDGs by 2030. As this MFA only focuses on one area and point in time, no conclusions can be drawn from our work about the level of human development for Africa compared to other contintents or about the development of individual countries over time. This could be subject of future work on this topic.

Bibliography

Bellman, R. (1966). *Dynamic programming*. Science. 153(3731), pp. 34-37.

Handl, A. and Kuhlenkasper, T. (2017): *Multivariate Analysemethoden, Theorie und Praxis mit R*. 3rd edn. Springer Spektrum.

Jolliffe, I. (2011) *Principal Component Analysis*. In: Lovric M. (eds) International Encyclopedia of Statistical Science. Springer, Berlin, Heidelberg. Available at: https://doi.org/10.1007/978-3-642-04898-2_455 (Accessed: 21 July 2021)

Kaiser, H. F. (1960): *The application of electronic computers to factor analysis*. Educ. Psychol. Meas., 20: pp. 141-151.

Pagès, J. (2004), *Multiple Factor Analysis: Main Features and Application to Sensory Data*. Revista Colombiana de Estadística, 27(1), pp. 1-26.

Pagès, J. (2015), *Multiple Factor Analysis by Example Using R*. CRC Press, Boca Ranton.

Sharp, J. (2020). *Congressional Research Service Report (RL33003): Egypt: Background and U.S. Relations*. U.S. Congress Printing Office.

United Nations, Department of Economic and Social Affairs, *THE 17 GOALS*, Available at: https://sdgs.un.org/goals (Accessed: 24 July 2021)

United Nations. Human Development Reports. *Education index.* Available at: http://hdr.undp.org/en/indicators/103706. (Accessed: 24 July 2021)

United Nations. Human Development Reports. *Human Development.* Availably at: http://hdr.undp.org/en/indicators/137506 (Accessed: 24 July 2021)

UN General Assembly. *Transforming our world: the 2030 Agenda for Sustainable Development*, 21 October 2015, A/RES/70/1, available at: https://www.un.org/en/development/desa/population/migration/generalassembly/docs/globalcompact/A_RES_70_1_E.pdf, p.1, (Accessed 23 July 2021).

World Bank. (2021). *Ranking of the 20 countries with the lowest life expectancy as of 2019.* Statista. Statista Inc.. Available at: https://www.statista.com/statistics/264719/ranking-of-the-20-countries-with-the-lowest-life-expectancy/ (Accessed: 27 July 2021)

Figures

online.stat.psu.edu: *3.4.1 - Scatterplots. Scatterplot of Weight (kg) and Heigh(cm).* Available at: https://online.stat.psu.edu/stat200/book/export/html/66 (Accessed 28 July 2021)

Appendix A

Table 1: Composition of Variable Groups

	Poverty Group	Health Group	Infrastructure Group	Development Assistance Group	Region (Supp.) Group
SGDs	1	3	6, 7	1, 2, 6, 8, 9	-
Variables	Employed.below.poverty.line, Using.basic.drink.water.services, Using.basic.sanitation.services	Infant.mortality.rate, Malaria.incidence, Maternal.mortality.rate, Neonatal.mortality.rate, HIV.infections, Interventions.tropical.diseases, Tuberculosis.incidence, Under.five.mortality.rate	Open.defecation, Handwashing.on.premises, Access.electricity, Reliance.on.clean.fuels	Grants.for.poverty.reduction, Assistance.water.supply, Aid.for.trade, Flows.for.agriculture, Flows.for.infrastructure	Region, Least.Developed.Country
% Missing Data	0,02	0,09	0,15	0	0
Year of Data	2018	2017	2019	2019	-

Table 2: Correlation between Variable Groups and Dimensions

	Dim. 1	Dim. 2
Poverty	0.92	0.21
Health	0.84	0.69
Infrastructure	0.96	0.33
Development Assistance	0.61	0.92

Appendix B

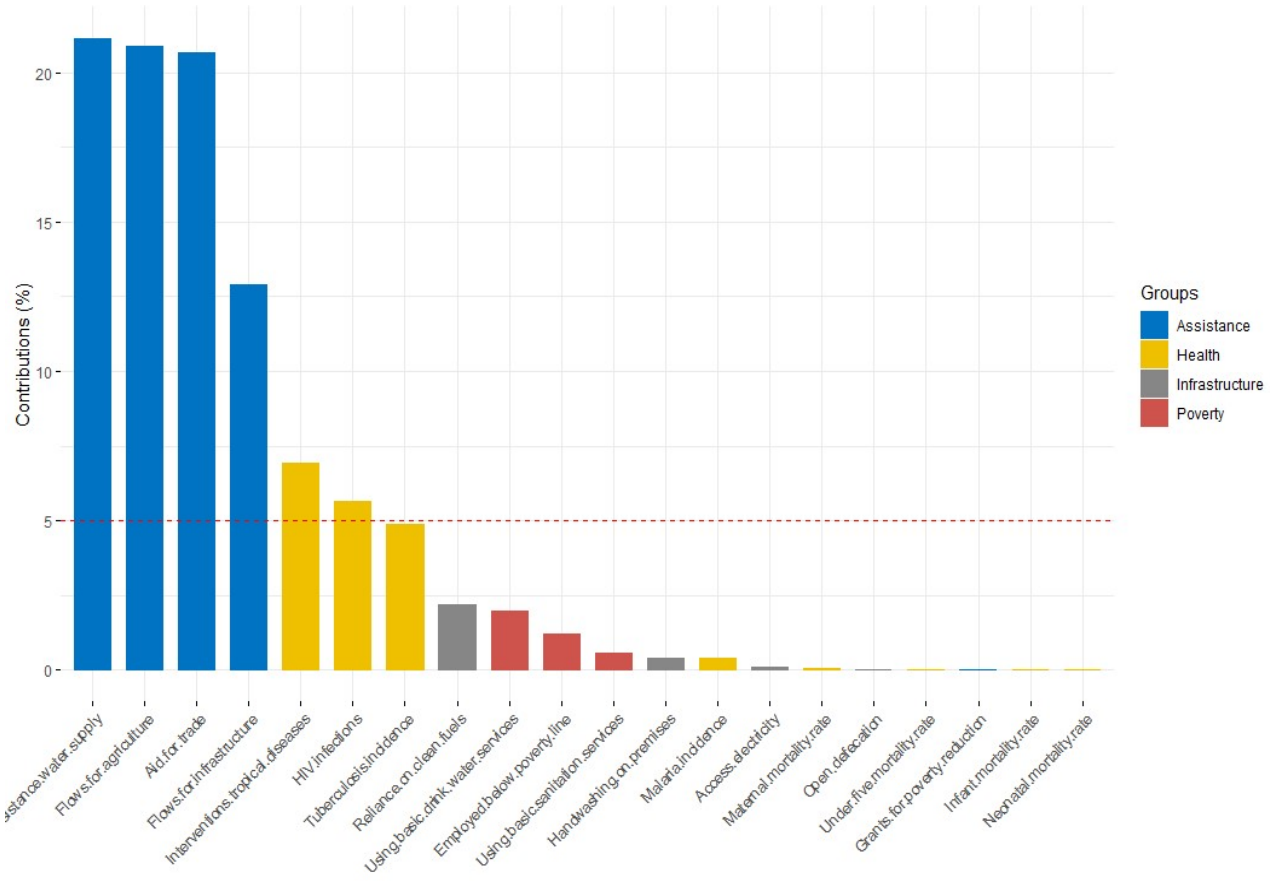

Figure: Contribution of Variables to Dimension 2