Investigating the 2017 State of the Tropics report using Principal Component Analysis

Nikki Fitzherbert

Introduction

The 2017 State of the Tropics report on sustainable infrastructure was the third report released by a global alliance of 11 research institutions since 2014. These reports aimed to assess and report on some of the key issues facing countries and nations sitting within the topical region of this planet, using data from a broad range of pre-existing environmental, social and economic indicators. The ultimate aim of the project was to determine whether life in the tropics was getting better (or worse).[1]

The 2017 report focused on the current and historical status of infrastructure development in the tropics. It explored the challenges and opportunities countries and nations in this region faced with respect to improving the provision of adequate services and facilities to their populations, and highlighted how the extent to which nations in this region develop sustainable, resilient and inclusive infrastructure would influence whether the United Nations' Sustainable Development Goals can be realised.[2]

Data

The State of the Tropics reports analysed data from a broad range of economic, social and environmental indicators that fall under 14 major groups: population growth, poverty, education and literacy, children, undernourishment, carbon dioxide emissions, gross domestic product per capita, urbanisation, forests, life expectancy, slums, protected areas, health and tropical oceans.[3]

The analysis performed here used only a subset of the data collated for the State of the Tropics report. It included numeric data from the calendar year 2010 on 13 different indicators and 109 countries or nations. The included variables are listed in the table below:

Variable Name	**Description**
Life expectancy	Number of years.
Poverty under $1.25 per day	Per cent in extreme poverty.
Population under 15	Per cent.
Adult literacy	Per cent of literate people above 15 years of age.
Mean years of schooling	Number of years.
Unemployment	Per cent.
Youth literacy	Per cent literate people aged 15 to 24 years.
Tuberculosis	Number of cases.
Under 5 mortality	Number of cases per 1000 births.
Poverty under $2 per day	Per cent in moderate poverty.
Undernourishment	Per cent undernourished.
Urban population	Per cent living in an urban area.
Area of agricultural land	Amount of agricultural land.

[1] https://www.jcu.edu.au/state-of-the-tropics/project
[2] https://www.jcu.edu.au/__data/assets/pdf_file/0004/473503/SOTT-2017-Infrastructure-Report_V02.pdf
[3] https://www.jcu.edu.au/state-of-the-tropics/data

Exploratory analysis indicated that the dataset used contained 187 missing values across nine of the 13 indicators (four indicators contained values for every country or nation in the dataset). This represented 13.2 per cent of the total dataset. All countries and nations with at least one missing value across the nine indicators were removed from the dataset. Although complete case analysis is the simplest approach to dealing with missing values, it can lead to biased results if the data is Missing at Random.[4] Alternative approaches include the use of single imputation methods such as substituting all values with the sample mean, or multiple imputation with a statistical model. The final dataset used in the analysis comprised 56 countries and nations across the 13 previously-described indicators.

Exploratory analysis also indicated that the dataset needed to be standardised prior to performing the singular value decomposition. This ensured that indicators of larger orders of magnitude did not bias the resulting principal axes. In addition, it can be shown mathematically that principal component analysis can be performed using singular value decomposition (SVD) if and only if the starting matrix of values is centred and scaled. The standardisation was performed using the following equation:

$$X_i = \frac{\tilde{X}_i - \mu_i}{\sqrt{\sigma_i}}$$

where X_i is the standardised matrix, $\tilde{X}_i$ is the unstandarised matrix, μ_i is the sample mean and σ_i is the sample standard deviation.

Methods

This objective of this analysis was to reduce the dimensionality of the dataset using Singular Value Decomposition (SVD). SVD decomposes a matrix of numeric values into three sub-matrices:

$$X = USV^T$$

where S is a diagonal matrix with singular values on its diagonal, U is a matrix containing the 'left singular values' and V is a matrix containing the 'right singular values' or the principal component (PC) vectors.

The variances of the PCs are related to the singular values of S through the following equation. $\lambda_i = \sigma_i^2$ and were used to determine the proportion of variation explained by the individual principal components as well as the cumulative contribution. The latter was then used in conjunction with an empirically-determined cut-off value of 80 per cent to decide on the number of principal components (k) to retain for the reduced-dimensionality matrix.

The next step was to calculate the matrix of scores for each country and nation, which is given by $T = US$, such that $T = XV$. The values of the scores indicated how much of each PC contributed to each case and assisted in determining how much any particular country or nation in the data deviated from the general trend.

[4] Determining whether the data were Missing Completely at Random, Missing at Random or Missing Not at Random was outside the scope of this analysis and therefore not investigated further.

The penultimate step in this analysis involved obtaining a dimensionally reduced representation of the dataset using the number of PCs determined by the 80 per cent cut-off value of cumulative proportion of variance explained:

$$\hat{X} = U\hat{S}\hat{V}^T$$

where $\hat{S}$ and $\hat{V}$ are the first k columns of S and V respectively.

The 'residuals' of the countries and nations could then be defined with the reduced dimensionality matrix $\hat{X}$:

$$SPE_j = \sum_i \left(X - \hat{X}\right)_{ji}^2$$

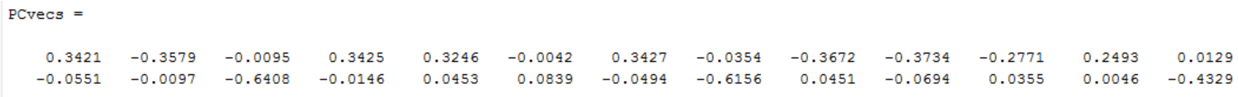
where SPE stands for square prediction error. If a country or nation had a large SPE, then it was not well described by the k PCs retained in the reduced dimensionality matrix. The residuals also helped identify countries and nations that were possible outliers in the dataset.

A copy of the Matlab code used can be found at Appendix A.

Results and Discussion

Principal component vectors

The PC vectors from matrix V provided information about the strength of the correlation between the variables in each principal component. The first two are shown below.



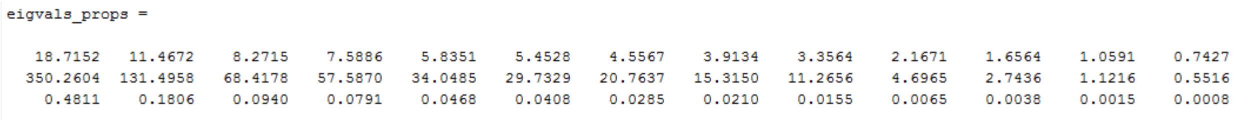
```
PCvecs =

    0.3421   -0.3579   -0.0095    0.3425    0.3246   -0.0042    0.3427   -0.0354   -0.3672   -0.3734   -0.2771    0.2493    0.0129
   -0.0551   -0.0097   -0.6408   -0.0146    0.0453    0.0839   -0.0494   -0.6156    0.0451   -0.0694    0.0355    0.0046   -0.4329
```

The first PC vector indicated that the strongest positive relationships were between 'Life expectancy', 'Adult literacy', 'Mean years of schooling', 'Youth literacy' and 'Urban population', and the strongest negative relationships were between 'Poverty under $1.25 per day (Extreme poverty)', 'Under 5 mortality', 'Poverty under $2 per day (Moderate poverty)' and 'Undernourishment'. The first PC vector therefore indicated that as life expectancy, educational levels and urbanisation increased, poverty/undernourishment and the mortality rate of children decreased.

The second PC vector indicated a negative relationship between 'Population under 15', 'Tuberculosis' and 'Area of agricultural land', with little connection to the other 10 indicators.

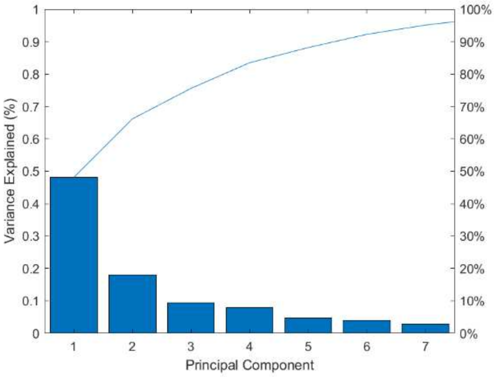
Singular values and proportion of variation

The SVD resulted in 13 singular values, which were used to derive the eigenvalues according to the equation defined in the methods section and the proportion of variation in the dataset due to each PC vector. These calculations showed that the first PC vector was able to account for just under 50 per cent (48.1) and the second 18.1 per cent of the variation in the dataset.



```
eigvals_props =

   18.7152   11.4672    8.2715    7.5886    5.8351    5.4528    4.5567    3.9134    3.3564    2.1671    1.6564    1.0591    0.7427
  350.2604  131.4958   68.4178   57.5870   34.0485   29.7329   20.7637   15.3150   11.2656    4.6965    2.7436    1.1216    0.5516
    0.4811    0.1806    0.0940    0.0791    0.0468    0.0408    0.0285    0.0210    0.0155    0.0065    0.0038    0.0015    0.0008
```

The plot below showed that retaining at least 90 per cent of the variation in the dataset required six PCs, whereas at least 80 per cent only required four. As such, a cut-off value of 80 per cent was deemed sufficient in this case.



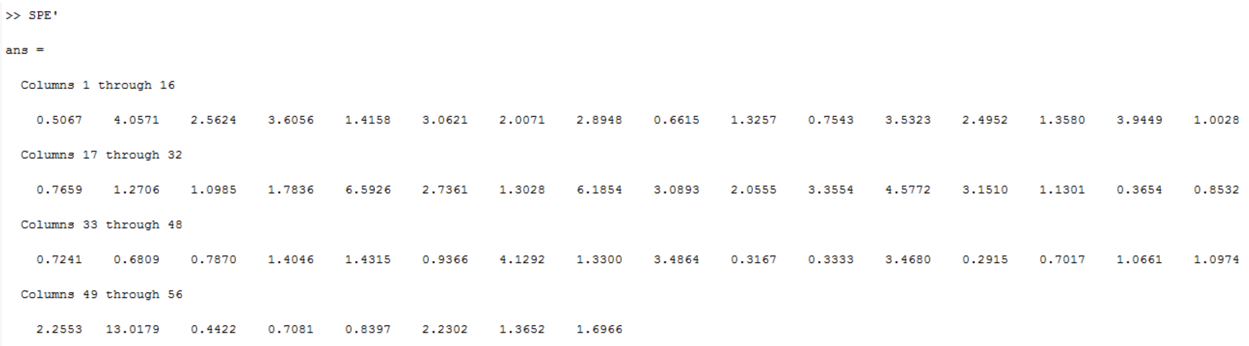
The matrix of scores

The matrix T containing the PCs contained 56 rows – one for each country or nation in the dataset. Since the first two PCs accounted for 83.5 per cent of the variation in the dataset, most of the 56 countries and nations should have had larger (absolute) scores for only these components. However, the matrix indicated that Ethiopia, Kenya, Malawi and Bangladesh could all be outliers as each of those four countries or nations had larger scores on the fifth and/or sixth PC.

The dimensionally-reduced representation

The reduced matrix was calculated by taking only the first four PCs, and the SPEs provided an indication of how well each country or nation was described by the reduced matrix. The SPEs were also able to confirm whether the countries possibly identified as outliers through the examination of the matrix of scores were actually outliers.

The 56 SPEs are shown below. They indicated that Bangladesh was the largest outlier ($SPE = 13.02$) and Kenya and Malawi were minor outliers ($SPE = 6.59$ and $SPE = 6.19$ respectively). However, the SPE metric did not consider Ethiopia to be an outlier as the residual was not particularly large.



```
>> SPE'

ans =

  Columns 1 through 16

    0.5067    4.0571    2.5624    3.6056    1.4158    3.0621    2.0071    2.8948    0.6615    1.3257    0.7543    3.5323    2.4952    1.3580    3.9449    1.0028

  Columns 17 through 32

    0.7659    1.2706    1.0985    1.7836    6.5926    2.7361    1.3028    6.1854    3.0893    2.0555    3.3554    4.5772    3.1510    1.1301    0.3654    0.8532

  Columns 33 through 48

    0.7241    0.6809    0.7870    1.4046    1.4315    0.9366    4.1292    1.3300    3.4864    0.3167    0.3333    3.4680    0.2915    0.7017    1.0661    1.0974

  Columns 49 through 56

    2.2553   13.0179    0.4422    0.7081    0.8397    2.2302    1.3652    1.6966
```

Appendix A – Matlab code

```
%% Assessment 4 - Principal Component Analysis
% Nikki Fitzherbert jc697513

%% Introduction
% The code in this assessment implements a principal component analysis
% using singular value decomposition, using a subset of the dataset
% aggregated for the 2017 Sustainable Infrastructure for the Tropics
% report.
%
% More information about the dataset can be found at:
https://www.jcu.edu.au/state-of-the-tropics

%% Preparatory work
% cleaning the workspace
clear all
clc

% loading in the data and determining the its dimensions
[ndata, text, alldata] = xlsread('SotTCombined2010.xlsx');
Xtild = ndata;

sum(isnan(Xtild))
sum(isnan(Xtild), 'all')

% removing missing values
Xtild_nomissing = rmmissing(Xtild);

% centring and scaling the data to create matrix X
X = (Xtild_nomissing - mean(Xtild_nomissing)) ./ std(Xtild_nomissing,1);

%% The Principal Component (PC) vectors and proportion of variation
% performing the SVD; that is, decomposing matrix X into its three
% component matrices
[U,S,V] = svd(X);

% V is a matrix with orthonormal rows and shows the relationships between
% the 13 indicators; that is, the columns of V are the right singular
% vectors of matrix X (or the PC vectors)
V;

% extracting the first two PC vectors
PCvecs = [V(:,1:2)]'

% S is a diagonal matrix showing the singular values of matrix X, and
% ordered from largest to smallest
S;
diag(S)';
```

```
% calculating the eigenvalues of matrix X. These are used to calculate the
% proportion of variation in the dataset due to each PC vector.
eigvals = diag(S).^2
prop_var = eigvals / sum(eigvals)

eigvals_props = [diag(S)'; eigvals'; prop_var']

% calculating and graphically displaying the cumulative proportion of
% variation explained
cumprop_var = cumsum(prop_var)

pareto(prop_var)
xlabel('Principal Component')
ylabel('Variance Explained (%)')

saveas(gcf, 'vars_plot', 'png')

%% The matrix of scores
% deriving the matrix of scores for each country/nation in the dataset.
% These are the principal components.
T = U*S
T2 = X*V;

%% Truncating the matrix
% deriving the reduced matrix by retaining only those PCs cumulatively
% explaining more than 80% of the variation in the data
k = 4;
Uhat = U(:, 1:k);
Shat = S(1:k, 1:k);
Vhat = V(:, 1:k);

Xhat = U(:, 1:k)*S(1:k, 1:k)*V(:, 1:k)'

% comparing the reduced matrix to the original dataset. If a case has a
% large squared prediction error then it isn't well described by the k PCs
% retained in Xhat
SPE = sum((X-Xhat).^2,2);
SPE_vec = SPE'
```