CxC 2025

SAP Sustainability & Multidimensional Poverty Index Challenge

By Shami-uz Zaman, Tony Ngo, Roah Cho, Angad Ahluwalia

Table of Contents

1. Executive Summary

This report details a data-driven analysis aimed at constructing a Multidimensional Poverty Index (MPI) by leveraging key socioeconomic indicators across countries. The study identifies and normalizes three core dimensions—education, health, and living standards—to rank countries by poverty levels. Advanced statistical techniques such as Principal Component Analysis (PCA) and Random Forests are employed to determine indicator influence, while clustering analysis groups countries for targeted policy interventions. The final recommendations focus on improving access to clean cooking fuels and technologies, alongside integrated strategies addressing education and health, in alignment with SAP's sustainability goals.

2. Introduction

Global poverty is a multifaceted challenge that requires an integrated understanding of its various dimensions. This report outlines the process of building an MPI by selecting key indicators related to education, health, and living standards. The analysis not only ranks countries based on their composite poverty score but also highlights the most influential factors contributing to poverty. By combining data preprocessing, normalization, and advanced analytical techniques, the study provides actionable insights to guide policy interventions.

3. Data Overview and Preparation

3.1. Data Loading and Cleaning

The dataset originates from an Excel file (SAP Datasets.xlsx), with the primary sheet labeled "AFE." The initial steps included:

- Loading the data using pandas and examining the first few rows.
- Dropping extraneous columns such as "short description" and "long description" to focus on essential variables.
- Converting the year columns (2000 to 2023) to numeric types and cleaning text fields for consistency.
- Evaluating missing values to understand data quality.

These steps ensure that the dataset is both accurate and ready for analysis.

The dataset was first loaded and cleaned to remove unnecessary columns and handle missing values. Below is the code used to preprocess the data before further analysis.

```
import pandas as pd
import numpy as np

df = pd.read_excel("SAP Datasets.xlsx", sheet_name="AFE")

# Drop unnecessary columns
df = df.drop(columns=["short description", "long description"],
```

```
errors='ignore')

# Convert year columns to numeric
year_columns = [col for col in df.columns if col.isdigit()]
df[year_columns] = df[year_columns].apply(pd.to_numeric,
errors='coerce')

# Check missing values
print(df.isnull().sum())
```

3.2. Exploratory Analysis: Time Series for Afghanistan

As an initial exploratory analysis, the indicator “Control of Corruption: Estimate” for Afghanistan was filtered and interpolated over the years 2000–2023. The interpolation provided a continuous time series, allowing visualization of trends despite missing data. This exercise underscored the importance of data completeness and set the stage for more complex multidimensional analysis.

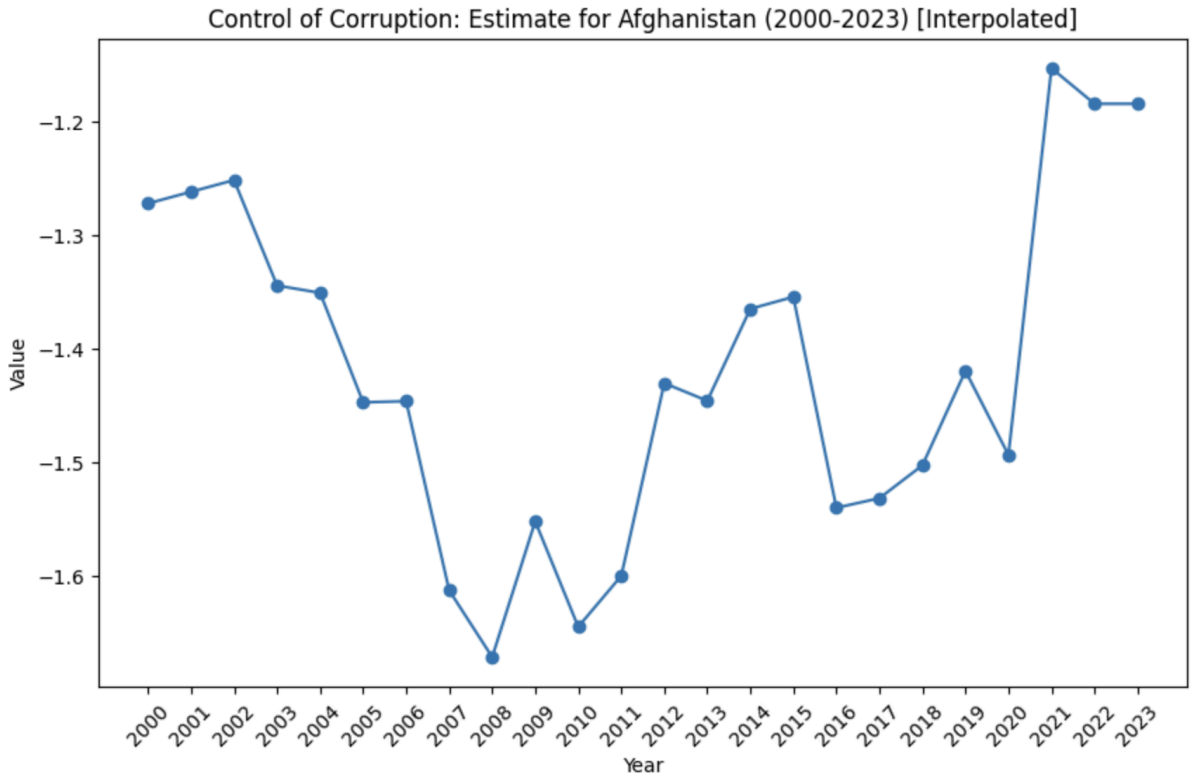

Figure 1: *Interpolated time series of 'Control of Corruption: Estimate' for Afghanistan (2000-2023). Missing values were filled using linear interpolation to enable trend analysis.*

To understand data completeness and historical trends, we examined the time series of 'Control of Corruption: Estimate' for Afghanistan, interpolating missing values for better visualization.

```
import matplotlib.pyplot as plt

afghanistan_data = df[df["Country Name"] == "Afghanistan"]
indicator = "Control of Corruption: Estimate"

afghanistan_ts = afghanistan_data[["Country Name"] +
year_columns].set_index("Country Name").T.interpolate()

plt.figure(figsize=(10,5))
plt.plot(afghanistan_ts.index, afghanistan_ts[indicator], marker='o',
linestyle='-')
plt.xlabel("Year")
plt.ylabel("Control of Corruption Estimate")
plt.title("Time Series of Control of Corruption in Afghanistan
(2000-2023)")
plt.grid(True)
plt.show()
```

4. Construction of the Multidimensional Poverty Index (MPI)

4.1. Selection of Key Indicators

For the MPI, three indicators were selected to capture the following dimensions:

- **Education:** Literacy rate, youth (ages 15-24), gender parity index (GPI)
- **Health:** Current health expenditure per capita (current US$)
- **Living Standards:** Access to clean fuels and technologies for cooking (% of population)

These indicators were chosen because they collectively offer insights into the educational, health, and living standards dimensions that are critical to understanding poverty.

4.2. Normalization and Inversion of Indicators

Since the selected indicators are “positive” indicators (where higher values indicate better outcomes), they were first normalized using the MinMaxScaler. Then, an inversion was performed (i.e., 1 – normalized_value) so that higher values reflect higher levels of poverty. This step is crucial for ensuring that all components of the MPI are aligned in direction, making the composite index intuitive.

Since the selected indicators had varying scales, Min-Max normalization was applied. Additionally, inversion was performed so that higher values consistently indicate higher levels of poverty.

```
from sklearn.preprocessing import MinMaxScaler
```

```
selected_indicators = [
    "Literacy rate, youth (ages 15-24), gender parity index (GPI)",
    "Current health expenditure per capita (current US$)",
    "Access to clean fuels and technologies for cooking (% of
population)"
]

scaler = MinMaxScaler()
df_normalized = df[selected_indicators].copy()
df_normalized[selected_indicators] =
scaler.fit_transform(df_normalized[selected_indicators])

# Invert so higher values indicate more poverty
df_normalized = 1 - df_normalized

df_normalized["Country Name"] = df["Country Name"]
```

4.3. Composite MPI Calculation

Two MPI scores were computed:

- **Simple Average MPI:** The mean of the inverted scores for the three selected indicators.
- **PCA-Weighted MPI:** A composite index where weights are determined via PCA loadings. The PCA-based approach helps mitigate the dominance of any single indicator by assigning data-driven weights.

Both methods yielded country rankings that identified nations such as Niger, Chad, and Guinea-Bissau as highly affected by multidimensional poverty.

5. Determining the Most Influential Indicators

5.1. PCA-based Weighting and Analysis

PCA was applied to the inverted indicators to extract the principal component that best explains the variance. The resulting loadings were normalized to generate weights. The analysis revealed that the “Access to clean fuels and technologies for cooking (% of population)” dominated the PCA, suggesting it is the most influential factor.

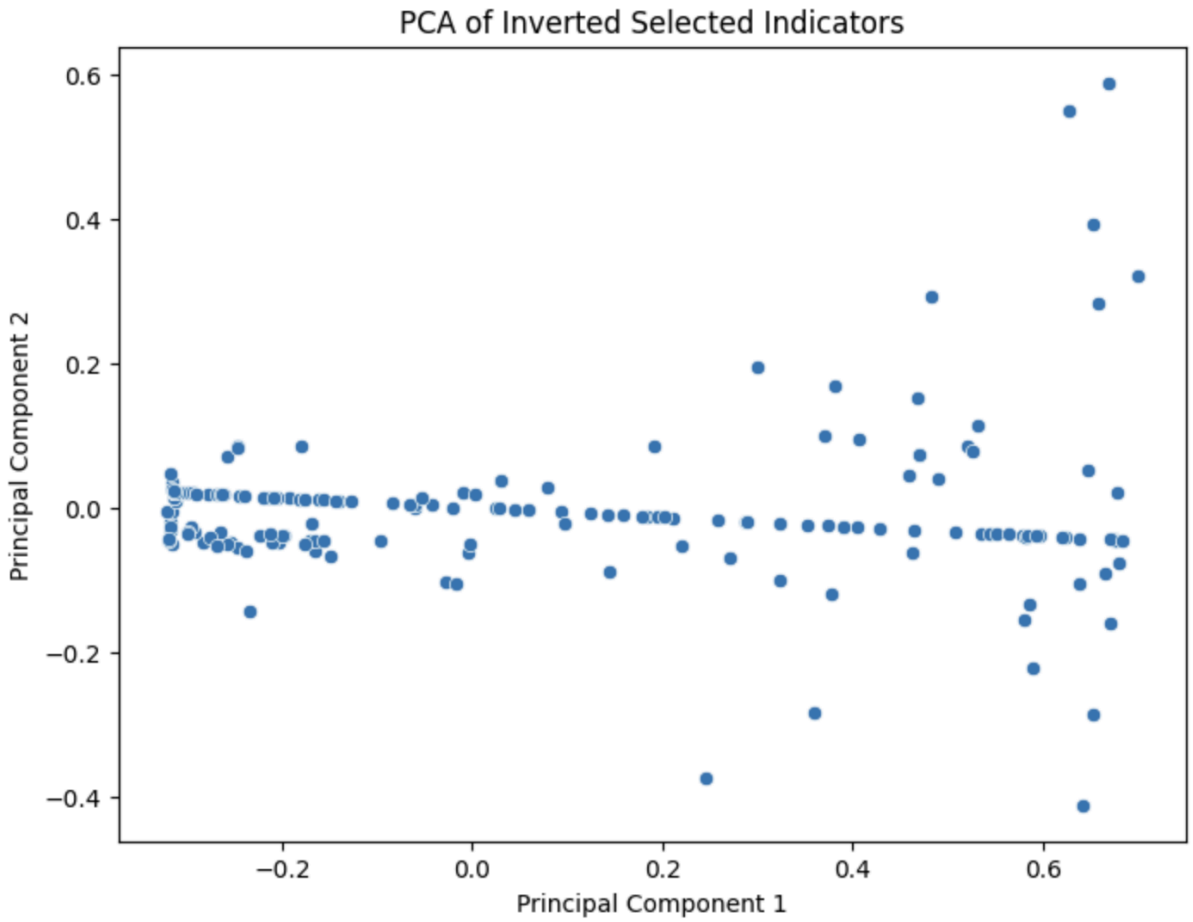

Figure 2: *Scatter plot of the first two principal components of the inverted selected indicators. The horizontal spread indicates the variance captured by Principal Component 1, which was used to determine the weights for the Multidimensional Poverty Index (MPI).*

PCA was applied to derive weights for each indicator based on variance explained, allowing an objective method for constructing the MPI.

```
from sklearn.decomposition import PCA

pca = PCA(n_components=1)
pca_weights = pca.fit_transform(df_normalized[selected_indicators]) /
pca.explained_variance_ratio_.sum()

df_normalized["MPI_PCA"] = (df_normalized[selected_indicators] *
pca_weights.flatten()).sum(axis=1)

# Display top affected countries
df_normalized[["Country Name", "MPI_PCA"]].sort_values(by="MPI_PCA",
ascending=False).head(10)
```

5.2. Correlation Analysis

Correlation coefficients between each inverted indicator and the simple average MPI were computed. The indicator related to clean cooking fuels had the highest positive correlation with the MPI, reinforcing its critical role in the poverty landscape.

5.3. Random Forest Feature Importances

A Random Forest regressor was also utilized to predict the MPI. The feature importances derived from this model confirmed that “Access to clean fuels and technologies for cooking (% of population)” is the most significant predictor, followed by the other two indicators.

A Random Forest model was trained to determine which indicators most significantly impact poverty levels. The resulting feature importance values highlight the strongest predictors of MPI.

```
from sklearn.ensemble import RandomForestRegressor

X = df_normalized[selected_indicators]
y = df_normalized["MPI_PCA"]

rf = RandomForestRegressor(n_estimators=100, random_state=42)
rf.fit(X, y)

feature_importances = pd.Series(rf.feature_importances_,
index=selected_indicators)
print(feature_importances.sort_values(ascending=False))
```

6. Clustering Analysis for Tailored Interventions

K-Means clustering was employed on the inverted indicators to group countries with similar poverty profiles. The clustering, visualized in PCA space, produced three distinct groups. This segmentation allows for the development of targeted policy interventions, acknowledging that different groups of countries may require tailored strategies to address their unique challenges.

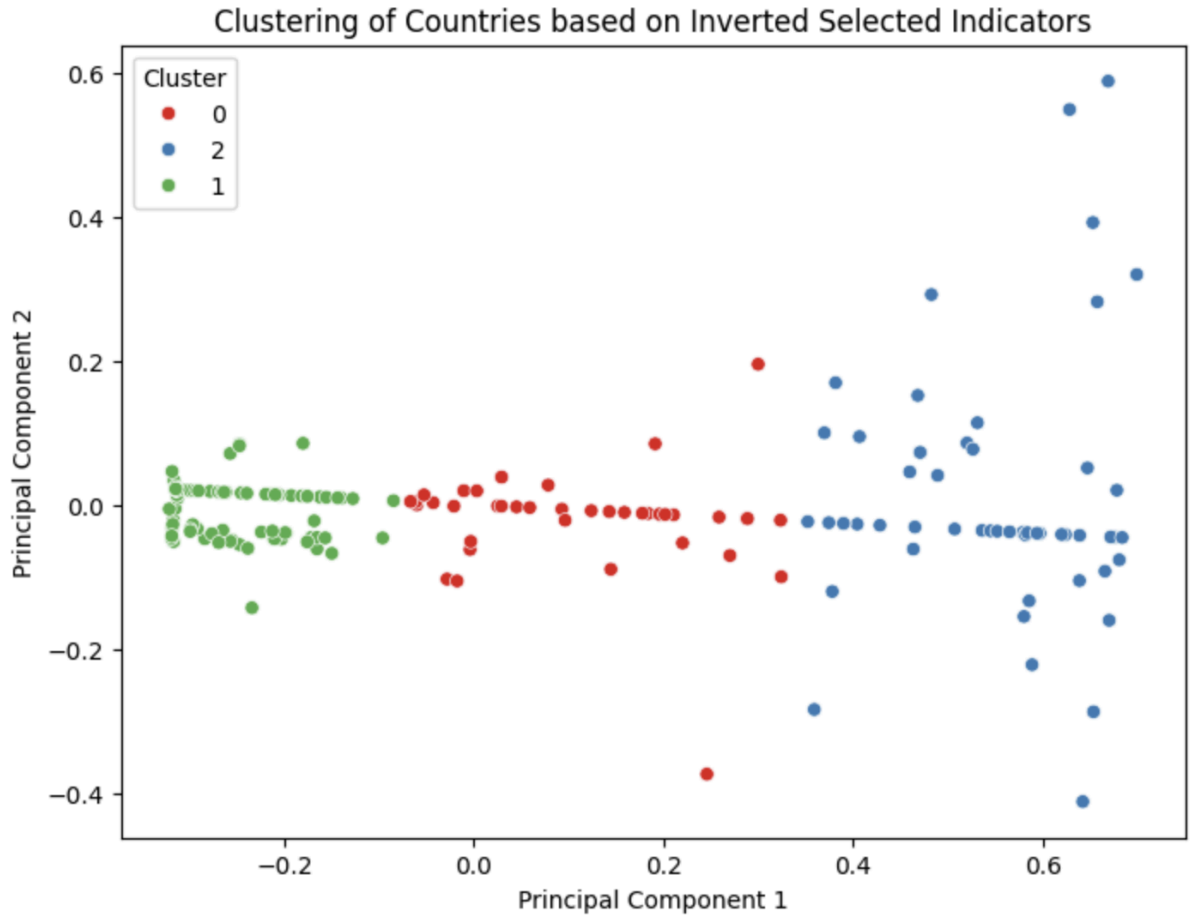

Figure 3: *Clustering of countries based on the first two principal components of the inverted selected indicators. Each color represents a different cluster, allowing for targeted policy recommendations tailored to similar poverty profiles.*

To tailor policy interventions, countries were grouped using K-Means clustering based on their MPI-related indicators. This allows region-specific recommendations for poverty alleviation.

```
from sklearn.cluster import KMeans
import seaborn as sns

kmeans = KMeans(n_clusters=3, random_state=42, n_init=10)
df_normalized["Cluster"] = kmeans.fit_predict(X)

plt.figure(figsize=(10,6))
```

```
sns.scatterplot(x=df_normalized[selected_indicators[0]],
                y=df_normalized[selected_indicators[1]],
                hue=df_normalized["Cluster"], palette="viridis")
plt.xlabel(selected_indicators[0])
plt.ylabel(selected_indicators[1])
plt.title("Country Clusters Based on MPI Indicators")
plt.legend(title="Cluster")
plt.show()
```

7. Policy Recommendations

To focus interventions on the most vulnerable populations, the top 10 countries with the highest MPI scores were identified.

```
top_n = 10
top_countries = df_normalized.nlargest(top_n, "MPI_PCA")[["Country
Name", "MPI_PCA"]]

print("Top 10 Most Affected Countries:\n", top_countries)
```

Based on the analysis, several data-driven policy recommendations are proposed:

- **Prioritize Clean Cooking Solutions:** Given the dominant influence of access to clean fuels, initiatives aimed at expanding access to clean cooking technologies should be a top priority.
- **Integrated Interventions:** Develop programs that simultaneously address educational improvements and health expenditure alongside efforts to improve living standards.
- **Tailored Regional Strategies:** Use clustering insights to design region-specific policies. For example, countries in clusters with particularly low access to clean fuels may benefit from subsidized clean energy programs and infrastructure investments.
- **Monitor and Evaluate:** Implement monitoring systems to continuously evaluate the impact of these interventions and adjust policies as needed, ensuring alignment with sustainability and poverty alleviation goals.

8. Conclusion

This report presents a comprehensive approach to constructing a Multidimensional Poverty Index that integrates key indicators of education, health, and living standards. By normalizing and inverting these indicators, and employing advanced analytical methods such as PCA and Random Forests, the study identifies critical areas for intervention. Clustering analysis further supports the development of tailored policy recommendations. The findings not only rank countries by poverty but also provide a clear roadmap for policy actions aligned with sustainable development goals.

9. Appendix: Code Overview

The complete code used in this analysis is structured as follows:

- **Data Loading and Cleaning:** Reading the Excel file, cleaning text fields, and handling missing data.
- **Exploratory Analysis:** Time series interpolation for a selected country and indicator.
- **MPI Construction:** Pivoting data, normalization, inversion of indicators, and computing both simple and PCA-weighted MPI.
- **Influence Analysis:** PCA loadings, correlation analysis, and Random Forest feature importances.
- **Clustering Analysis:** K-Means clustering and visualization in PCA space.
- **Policy Recommendations and Reporting:** Summarizing findings and implications for policy.