ROYAUME DU MAROC

HAUT COMMISSARIAT AU PLAN

INSTITUT NATIONAL DE STATISTIQUE ET D'ECONOMIE APPLIQUEE

INSEA

RAPPORT - ANALYSE DE DONNÉES

Utilisation de l'ACP pour Identifier les Pays Prioritaires

En Fonction des Indicateurs Socio-économiques et Sanitaires

Réalisé par :
ELFATINE M'barka
LAOUAD Ayoub

Encadré par :
Pr. SADKI Marya

Master Systèmes d'Information et Systèmes Intelligents

M2SI M1
2024 - 2025

Table des matières

1 Introduction

1.1 Contexte et problématique

Dans un contexte mondial où les inégalités socio-économiques et sanitaires persistent, l'allocation efficace des ressources humanitaires représente un défi majeur pour les organisations non gouvernementales (ONG). Face à un budget humanitaire limité de 10 millions de dollars, une ONG internationale doit identifier de manière objective et scientifique les pays nécessitant une aide prioritaire.

Cette problématique soulève plusieurs questions fondamentales :

— Comment classer objectivement les pays selon leurs besoins ?
— Quels indicateurs utiliser pour mesurer le niveau de développement ?
— Comment optimiser l'impact des fonds disponibles ?
— Comment garantir une approche transparente et équitable ?

1.2 Objectifs de l'étude

L'objectif principal de cette étude est de développer une méthodologie basée sur l'Analyse en Composantes Principales (ACP) pour :

1. Analyser les relations entre les différents indicateurs socio-économiques et sanitaires
2. Identifier les dimensions principales qui caractérisent le développement des pays
3. Classer les pays selon leur niveau de priorité pour l'aide humanitaire
4. Fournir une base scientifique pour la prise de décision stratégique

1.3 Méthodologie générale

Notre approche s'appuie sur l'utilisation de l'Analyse en Composantes Principales (ACP), une méthode statistique multivariée qui permet de :

— Réduire la dimensionnalité des données tout en conservant l'information essentielle
— Identifier les structures sous-jacentes dans les données
— Visualiser les relations complexes entre variables et individus
— Classifier les pays selon des critères objectifs

2 Revue de littérature

2.1 L'Analyse en Composantes Principales : fondements théoriques

L'Analyse en Composantes Principales, développée par Pearson (1901) et Hotelling (1933), est une technique statistique fondamentale pour l'analyse de données multidimensionnelles. Selon Jolliffe (1986), l'ACP vise à transformer un ensemble de variables corrélées en un nombre plus restreint de variables non corrélées appelées composantes principales.

2.2 Applications de l'ACP dans le développement international

Plusieurs études ont démontré l'efficacité de l'ACP dans l'analyse des indicateurs de développement :

- Analyse des inégalités mondiales (Milanovic, 2016)
- Classification des pays selon leur niveau de développement humain (UNDP, 2020)
- Évaluation de l'efficacité de l'aide internationale (Easterly, 2006)

3 Données et variables

3.1 Source des données

Les données utilisées proviennent de la plateforme Kaggle, une source reconnue et largement utilisée par les professionnels et chercheurs en science des données. L'ensemble de données comprend 167 pays et 9 variables socio-économiques et sanitaires.

3.2 Description des variables

Le tableau 1 présente les 9 variables analysées :

TABLE 1 – Description des variables du dataset

Variable	**Description**	**Unité**
child_mort	Mortalité infantile (pour 1000 naissances)	Nombre
exports	Exportations en pourcentage du PIB	Pourcentage
health	Dépenses de santé en pourcentage du PIB	Pourcentage
imports	Importations en pourcentage du PIB	Pourcentage
income	Revenu par personne	USD
inflation	Taux d'inflation annuel	Pourcentage
life_expec	Espérance de vie à la naissance	Années
total_fer	Taux de fécondité total	Nombre
gdpp	PIB par habitant	USD

3.3 Renommage des variables

Pour améliorer la lisibilité et l'interprétation, les variables ont été renommées en français :

TABLE 2 – Correspondance entre noms originaux et noms français

Nom original	Nom français
child_mort	Décès_enfants
exports	Exportations_PIB
health	Santé_PIB
imports	Importations_PIB
income	Revenu_par_personne
inflation	Inflation
life_expec	Espérance_de_vie
total_fer	Fécondité_totale
gdpp	PIB_par_habitant

4 Méthodologie

4.1 Préparation des données

4.1.1 Importation et structuration

Les données ont été importées à partir du fichier CSV et structurées de manière appropriée pour l'analyse. Le code R suivant illustre cette étape :

Listing 1 – Importation des données

```
# Importer les donn es
country_data <- read.csv("~/Country-data.csv", head=TRUE, sep = ",")
rownames(country_data) <- country_data$country
country_data <- country_data[, -1]
```

4.1.2 Analyse des valeurs manquantes

Une fonction spécifique a été développée pour analyser les valeurs manquantes :

Listing 2 – Analyse des valeurs manquantes

```
proportion_valeurs_manquantes <- function(data) {
  nb_valeurs_manquantes <- sapply(data, function(x) sum(is.na(x)))
  proportion_manquantes <- nb_valeurs_manquantes / nrow(data)
  data.frame(Nombre = nb_valeurs_manquantes,
             Proportion = proportion_manquantes)
}
```

4.2 Analyse exploratoire

4.2.1 Statistiques descriptives

L'analyse descriptive permet de comprendre la distribution de chaque variable et d'identifier les éventuelles anomalies. Les fonctions R `str()` et `summary()` ont été utilisées pour obtenir une vue d'ensemble des données.

4.2.2 Visualisation des distributions

Des histogrammes ont été générés pour chaque variable quantitative afin de visualiser leur distribution :

Listing 3 – Génération des histogrammes

```
for (var in names(country_data)[vars_quantitatives]) {
  print(ggplot(country_data, aes_string(x = var)) +
          geom_histogram(bins = 30, fill = "blue", color = "black") +
          theme_minimal() +
          labs(title = paste("Histogramme de", var),
               x = var, y = "Fr quence"))
}
```

4.3 Analyse des corrélations

4.3.1 Matrice de corrélation

La matrice de corrélation permet d'identifier les relations linéaires entre les variables :

Listing 4 – Calcul de la matrice de corrélation

```
donnees_quantitatives <- country_data[, vars_quantitatives]
matrice_correlation <- cor(donnees_quantitatives, use = "complete.obs")
```

4.3.2 Visualisation des corrélations

Le package `corrplot` a été utilisé pour visualiser la matrice de corrélation :

Listing 5 – Visualisation de la matrice de corrélation

```
corrplot(matrice_correlation, method = "color", type = "upper",
         order = "hclust", tl.col = "black", tl.srt = 45,
         addCoef.col = "black")
```

4.4 Analyse en Composantes Principales

4.4.1 Fondements mathématiques

L'ACP repose sur plusieurs étapes mathématiques fondamentales :

Centrage et réduction des données

$$M_{scale} = \frac{x_i - \text{Moyenne estimée}}{\text{Écart-type estimé}} \tag{1}$$

Calcul de la matrice de corrélation

$$\text{Matrice de corrélation} = \frac{M_{scale}^T \cdot M_{scale}}{n} \tag{2}$$

Décomposition en valeurs propres

$$\text{Matrice ACP} = M_{\text{Vecteur propre}} \times M_{\text{Valeur propre}} \times M_{\text{Vecteur propre}}^T \tag{3}$$

4.4.2 Implémentation

Listing 6 – Centrage et réduction des données

Centrage et réduction

```
donnees_centrees_reduites <- scale(donnees_quantitatives,
                                   center = TRUE, scale = TRUE)
```

Listing 7 – Réalisation de l'ACP

Réalisation de l'ACP

```
resultat_acp <- PCA(donnees_centrees_reduites, graph = FALSE)
```

5 Résultats

5.1 Analyse des valeurs manquantes

L'analyse des valeurs manquantes révèle que le dataset est relativement complet, avec peu de valeurs manquantes. Cette caractéristique permet une analyse robuste sans nécessiter d'imputation complexe.

5.2 Statistiques descriptives

Les statistiques descriptives montrent une grande variabilité entre les pays pour tous les indicateurs analysés, confirmant la pertinence de l'approche par ACP pour identifier les structures sous-jacentes.

5.3 Matrice de corrélation

La matrice de corrélation révèle plusieurs relations importantes :
— Corrélation positive forte entre PIB par habitant et espérance de vie
— Corrélation négative entre mortalité infantile et développement économique
— Relations complexes entre les variables commerciales (exportations/importations)

5.4 Résultats de l'ACP

5.4.1 Valeurs propres et variance expliquée

L'analyse des valeurs propres permet de déterminer le nombre de composantes principales à retenir. Le critère de Kaiser (valeurs propres > 1) et l'analyse du coude suggèrent de retenir les 2 ou 3 premières composantes.

5.4.2 Cercle de corrélation

Le cercle de corrélation (Figure **??**) illustre les relations entre les variables dans l'espace des composantes principales. Les variables proches du cercle unité sont bien représentées, tandis que leur position angulaire indique leurs corrélations.

5.4.3 Classification des pays

Le biplot permet de visualiser simultanément les pays et les variables dans l'espace des composantes principales, facilitant l'identification des groupes de pays aux caractéristiques similaires.

6 Interprétation et discussion

6.1 Interprétation des composantes principales

6.1.1 Première composante principale

La première composante semble capturer une dimension de "développement économique et sanitaire", opposant les pays développés (fort PIB, haute espérance de vie) aux pays en développement (forte mortalité infantile, faible revenu).

6.1.2 Deuxième composante principale

La deuxième composante pourrait représenter l'orientation économique, distinguant les pays selon leur degré d'ouverture commerciale et leurs spécialisations économiques.

6.2 Classification des pays

6.2.1 Pays prioritaires

Les pays situés dans le quadrant correspondant à un faible développement économique et sanitaire constituent les candidats prioritaires pour l'aide humanitaire.

6.2.2 Pays intermédiaires

Les pays en position intermédiaire nécessitent une analyse plus fine pour déterminer leurs besoins spécifiques.

6.2.3 Pays développés

Les pays développés, bien que n'étant pas prioritaires pour l'aide humanitaire, peuvent constituer des partenaires potentiels pour le financement.

7 Recommandations stratégiques

7.1 Allocation des ressources

Basé sur les résultats de l'ACP, nous recommandons :

1. Allocation de 60% du budget aux pays du premier quartile (priorité maximale)
2. Allocation de 30% aux pays du deuxième quartile (priorité élevée)
3. Allocation de 10% pour les interventions d'urgence et la coordination

7.2 Indicateurs de suivi

Les composantes principales identifiées peuvent servir d'indicateurs de suivi pour évaluer l'efficacité des interventions.

7.3 Partenariats stratégiques

L'analyse suggère des opportunités de partenariat avec les pays développés pour maximiser l'impact des interventions.

8 Limites et perspectives

8.1 Limites de l'étude

8.1.1 Limitations méthodologiques

— L'ACP assume des relations linéaires entre variables
— La méthode est sensible aux valeurs aberrantes
— L'interprétation des composantes reste subjective

8.1.2 Limitations des données

— Données statiques ne reflétant pas l'évolution temporelle
— Possible biais dans la collecte des données par pays
— Absence de certains indicateurs qualitatifs importants

8.2 Perspectives d'amélioration

8.2.1 Enrichissement des données

— Intégration de données temporelles pour une analyse longitudinale
— Ajout d'indicateurs de gouvernance et de stabilité politique
— Inclusion de données sur les catastrophes naturelles et conflits

8.2.2 Méthodologies complémentaires

— Analyse de classification automatique (clustering)
— Modélisation prédictive pour l'évolution des indicateurs
— Analyse de sensibilité pour tester la robustesse des résultats

9 Conclusion

Cette étude démontre l'efficacité de l'Analyse en Composantes Principales pour identifier les pays prioritaires dans le cadre de l'allocation d'aide humanitaire. L'approche développée offre plusieurs avantages :

1. **Objectivité** : La méthode statistique élimine les biais subjectifs dans la sélection des pays
2. **Transparence** : Les critères de classification sont explicites et reproductibles

3. **Efficacité** : L'analyse multivariée capture les interactions complexes entre indicateurs
4. **Adaptabilité** : La méthode peut être actualisée avec de nouvelles données

L'identification des deux principales dimensions du développement (économique-sanitaire et orientation commerciale) fournit un cadre conceptuel robuste pour la prise de décision stratégique. La classification des 167 pays selon ces dimensions permet une allocation optimisée des 10 millions de dollars disponibles.

Les résultats suggèrent qu'une approche différenciée, concentrant les efforts sur les pays les plus vulnérables tout en développant des partenariats stratégiques, maximiserait l'impact de l'aide humanitaire.

Cette méthodologie peut être étendue et adaptée à d'autres contextes de développement international, offrant un outil précieux pour les décideurs et les organisations humanitaires dans leur mission d'amélioration des conditions de vie mondiale.

10 Références bibliographiques

Références

[1] Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11), 559-572.

[2] Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24(6), 417-441.

[3] Jolliffe, I. T. (1986). *Principal Component Analysis*. New York : Springer-Verlag.

[4] Milanovic, B. (2016). *Global Inequality : A New Approach for the Age of Globalization*. Cambridge, MA : Harvard University Press.

[5] United Nations Development Programme. (2020). *Human Development Report 2020*. New York : UNDP.

[6] Easterly, W. (2006). *The White Man's Burden : Why the West's Efforts to Aid the Rest Have Done So Much Ill and So Little Good*. New York : Penguin Press.

[7] Kaggle. (2024). Clustering PCA Assignment Dataset. Récupéré de https ://www.kaggle.com/datasets/vipulgohel/clustering-pca-assignment/data

[8] R Core Team. (2024). R : A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

[9] Lê, S., Josse, J., & Husson, F. (2008). FactoMineR : An R package for multivariate exploratory data analysis and data mining. *Journal of Statistical Software*, 25(1), 1-18.

[10] Wickham, H. (2016). *ggplot2 : Elegant Graphics for Data Analysis*. New York : Springer-Verlag.

11 Annexes

11.1 Annexe A : Code R complet

Listing 8 – Code R complet pour l'analyse ACP

```
# Chargement des biblioth ques n cessaires
if (!require("VIM")) install.packages("VIM")
if (!require("ggplot2")) install.packages("ggplot2")
if (!require("corrplot")) install.packages("corrplot")
if (!require("FactoMineR")) install.packages("FactoMineR")
if (!require("factoextra")) install.packages("factoextra")

library(VIM)
library(ggplot2)
library(corrplot)
library(FactoMineR)
library(factoextra)

# Importer les donn es
country_data <- read.csv("~/Country-data.csv", head=TRUE, sep = ",")
rownames(country_data) <- country_data$country
country_data <- country_data[, -1]

# Renommer les colonnes
colnames(country_data) <- c(
  "D  c  s_enfants",
  "Exportations_PIB",
  "Sant _PIB",
  "Importations_PIB",
  "Revenu_par_personne",
  "Inflation",
  "Esp rance_de_vie",
  "F condit _totale",
  "PIB_par_habitant"
)

# Analyse des valeurs manquantes
proportion_valeurs_manquantes <- function(data) {
  nb_valeurs_manquantes <- sapply(data, function(x) sum(is.na(x)))
  proportion_manquantes <- nb_valeurs_manquantes / nrow(data)
  data.frame(Nombre = nb_valeurs_manquantes,
             Proportion = proportion_manquantes)
}

# Analyse des corr lations
vars_quantitatives <- sapply(country_data, is.numeric)
donnees_quantitatives <- country_data[, vars_quantitatives]
matrice_correlation <- cor(donnees_quantitatives, use = "complete.obs")

# Centrage et r duction des donn es
donnees_centrees_reduites <- scale(donnees_quantitatives,
                                   center = TRUE, scale = TRUE)

# R alisation d'une ACP
resultat_acp <- PCA(donnees_centrees_reduites, graph = FALSE)
```

```
# Visualisations
# Cercle de corr lation
fviz_pca_var(resultat_acp, col.var = "cos2",
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE)

# Biplot
fviz_pca_biplot(resultat_acp,
                select.ind = list(cos2 = 0.2),
                select.var = list(cos2 = 0.2),
                repel = TRUE,
                col.ind = "blue", col.var = "red",
                labelsize = 3,
                pointsize = 1)
```

11.2 Annexe B : Statistiques descriptives détaillées

[Ici seraient insérées les statistiques descriptives complètes pour chaque variable]

11.3 Annexe C : Tableaux des résultats ACP

[Ici seraient insérés les tableaux détaillés des valeurs propres, contributions, etc.]