# APPRENTISSAGE FIABLE PAR APPROCHE EM POUR MODELES A FACTEURS A UNE ÉQUATION STRUCTURELLE


Equipe :

Minjia FU
Sabrina MOCKBEL
Théophile ROUSSELLE

Encadrante :

Myriam TAMI

Référents :

Gurvan HERMANGE
Véronique LE CHEVALIER
Ioane MUNI TOKE

# Tables des Matières:

# I-INTRODUCTION

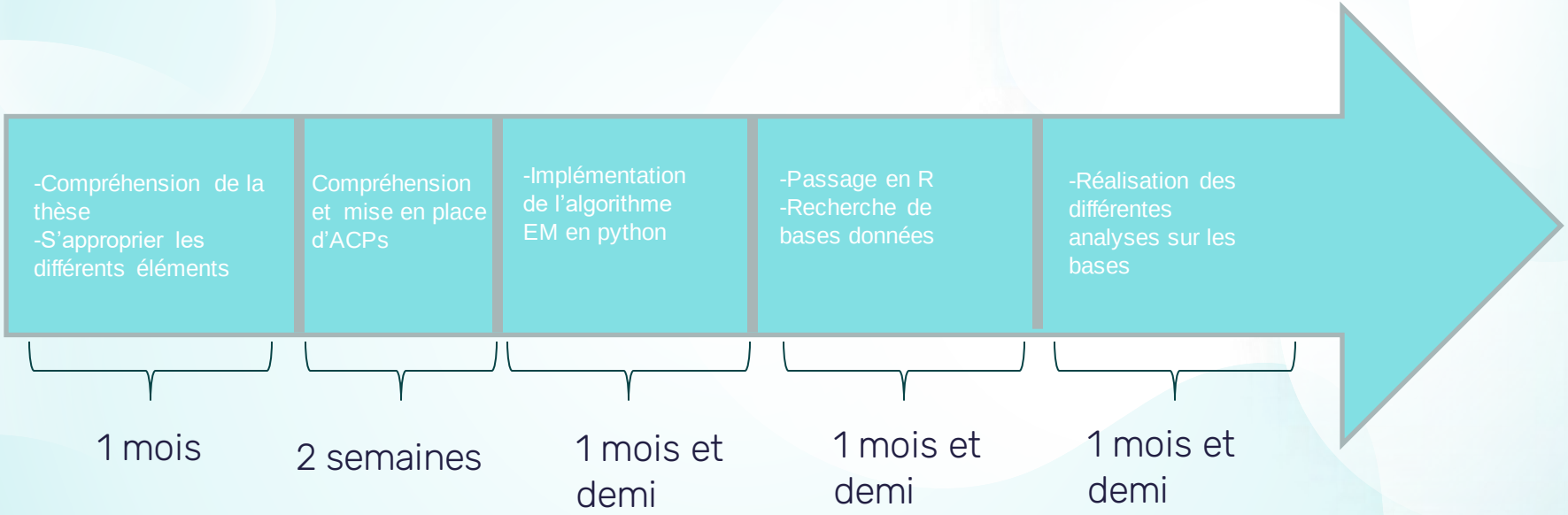
Contexte:

- Projet à CentraleSupélec dans le pôle Modélisation Mathématiques des Systèmes Complexes
- Travail sur la thèse Madame Myriam Tami: ***Approche EM pour modèle multi-blocs à facteurs à une équation structurelles***

Problématique:

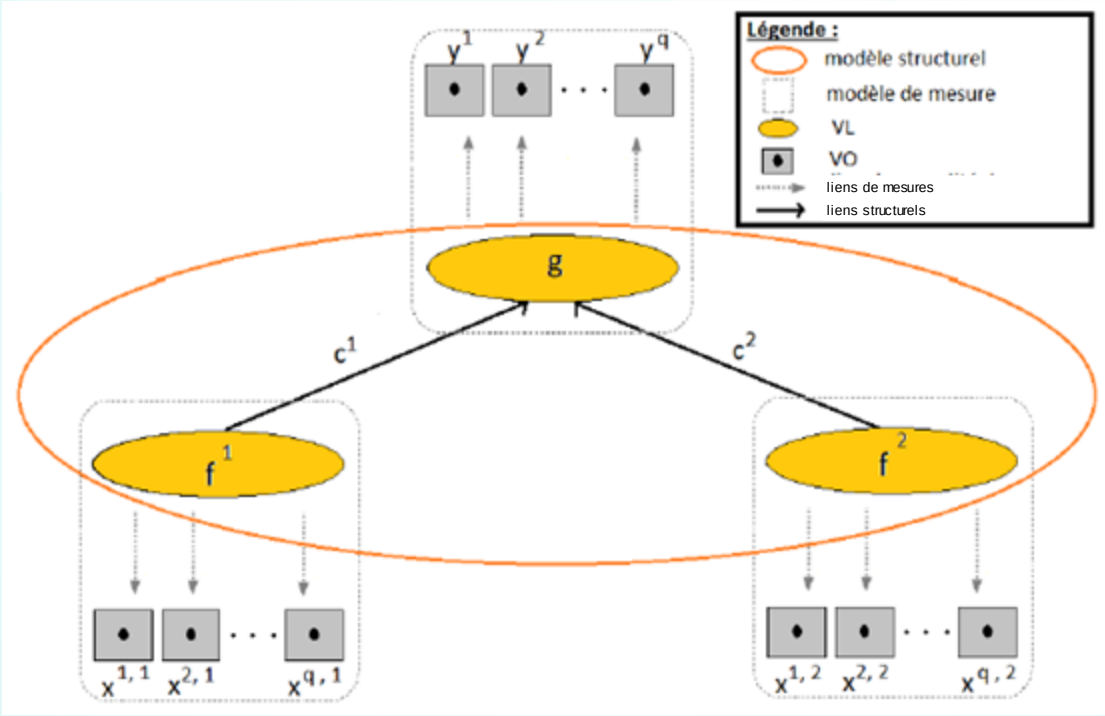
-S'approprier la méthode présentée dans la thèse afin d'en faire une application concrète.

# I-INTRODUCTION



# II-PRESENTATION DE LA THEORIE

## II.a Equations du Modèle



# II-PRESENTATION DE LA THEORIE

## II.a Equations du Modèle

Equation Structurelle:

$$g = c_1 f_1 + c_2 f_2 + \epsilon_g$$

Equations de mesures:

$$\begin{cases} Y = 1_n \mu_Y + gb + \epsilon_Y \\ X_1 = 1_n \mu_1 + f_1 a_1 + \epsilon_1 \\ X_2 = 1_n \mu_2 + f_2 a_2 + \epsilon_2 \end{cases}$$

# II-PRESENTATION DE LA THEORIE

## II.a Equations du Modèle

Hypothèses du modèle:

- $f_1 \sim \mathcal{N}(1, 0)$
- $f_2 \sim \mathcal{N}(1, 0)$
- $\epsilon_1^i \sim \mathcal{N}(0, \Psi_1)$ où $\Psi_1$ est la matrice diagonale composé des variances de chaques variables composant $X^1$
- $\epsilon_2^i \sim \mathcal{N}(0, \Psi_2)$ où $\Psi_2$ est la matrice diagonale composé des variances de chaques variables composant $X^2$
- $\epsilon_Y \sim \mathcal{N}(0, \Psi_Y)$ où $\Psi_Y$ est la matrice diagonale composé des variances de chaques variables composant $X^1$
- $\epsilon_g^i \sim \mathcal{N}(1, 0)$
- $g \sim \mathcal{N}(0, (c_1)^2 + (c_2)^2 + 1)$
- $\epsilon_g^i$ est indépedendant de $f_1$ et $f_2$ pour tout observation i
- $\epsilon_Y, \epsilon_1$et $\epsilon_2$ sont indépendants

Paramètres du modèle:

$$\theta = [\mu_Y, \mu_1, \mu_2, b, a_1, a_2, (\sigma_Y)^2, (\sigma_1)^2, (\sigma_2)^2, c_1, c_2]$$

# II-PRESENTATION DE LA THEORIE

## II.b Algorithme EM

### Phase E:

Calcul de différentes espérances des variables latentes sachant les variables observées pour un paramètre θ donnée :

$$\widetilde{\gamma_i} = \mathbb{E}_{z_i}^{h_i}[g_i^2] = \left(\mathbb{E}_{z_i}^{h_i}[g_i]\right)^2 + \mathbb{V}_{z_i}^{h_i}[g_i] = m_{1_i}^2 + \sigma_{11_i}\,; \qquad \widetilde{g_i} = \mathbb{E}_{z_i}^{h_i}[g_i] = m_{1i}\,;$$

$$\widetilde{\phi_i^1} = \mathbb{E}_{z_i}^{h_i}\left[\left(f_i^1\right)^2\right] = \left(\mathbb{E}_{z_i}^{h_i}[f_i^1]\right)^2 + \mathbb{V}_{z_i}^{h_i}[f_i^1] = m_{2_i}^2 + \sigma_{22_i}\,; \qquad \widetilde{f_i^1} = \mathbb{E}_{z_i}^{h_i}[f_i^1] = m_{2i}\,;$$

$$\widetilde{\phi_i^2} = \mathbb{E}_{z_i}^{h_i}\left[\left(f_i^2\right)^2\right] = \left(\mathbb{E}_{z_i}^{h_i}[f_i^2]\right)^2 + \mathbb{V}_{z_i}^{h_i}[f_i^2] = m_{3_i}^2 + \sigma_{33_i}\,; \qquad \widetilde{f_i^2} = \mathbb{E}_{z_i}^{h_i}[f_i^2] = m_{3i}.$$

### Phase M:

Réactualisation du paramètre théta en maximisant le log-vraisemblance

$$\frac{\partial}{\partial\theta}\mathcal{L}(\theta) = 0$$

# II-PRESENTATION DE LA THEORIE

## II.b Algorithme EM

2 critères d'arrêt:

-nombre maximum d'itérations

-Distance entre les θ faibles:

$$\sum_{k=1}^{K} \frac{\left|\theta^{t+1}[k] - \theta^{t}[k]\right|}{\left|\theta^{t+1}[k]\right|} < \epsilon$$

# III - Application à la base de données “genus”

## III.a Présentation de la base de donnée "genus"

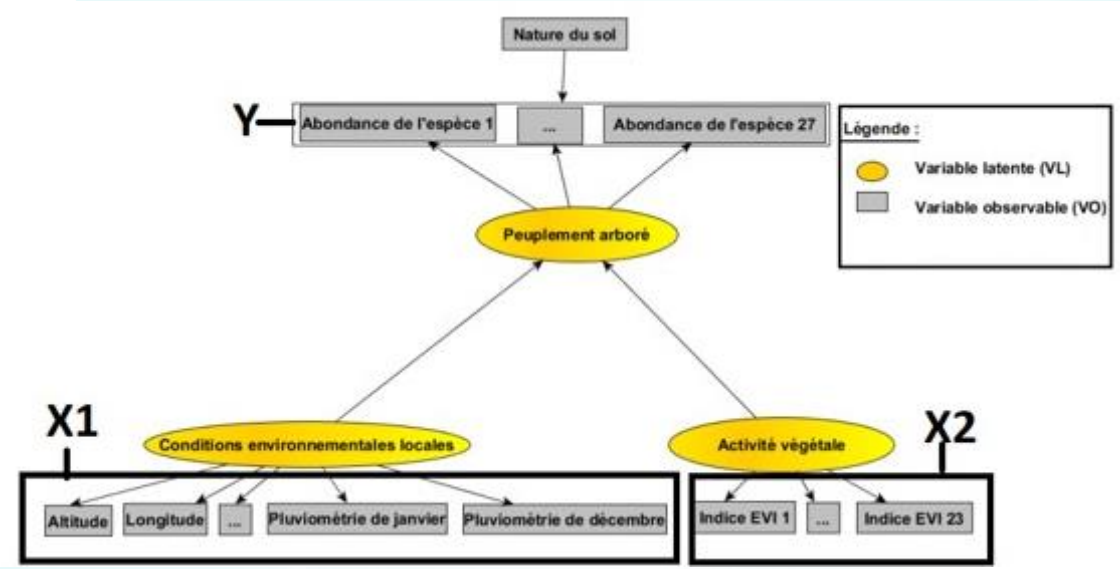



**Figure 1** – Diagramme structurel du peuplement arboré expliqué par deux blocs de variables observées liées aux conditions environnementales locales et à l’activité végétale

# III - Application à la base de données “genus”

## III.a Présentation de la base de donnée "genus"

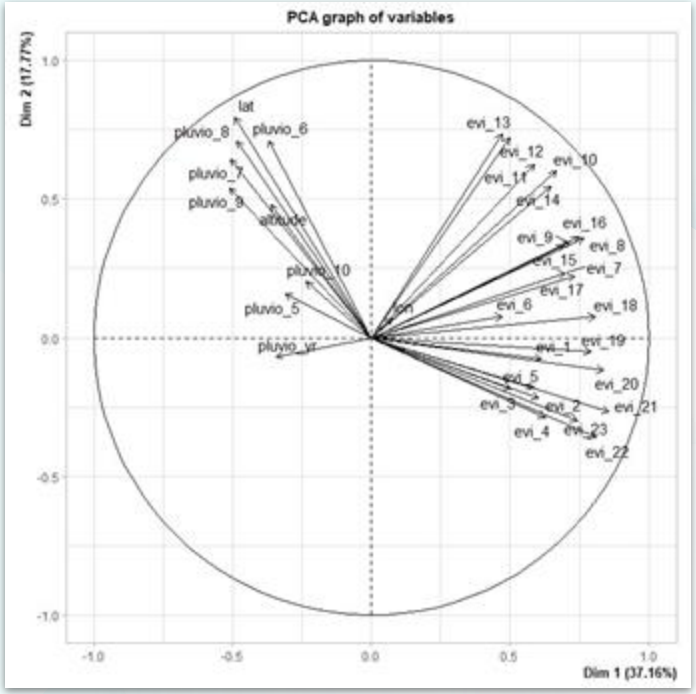
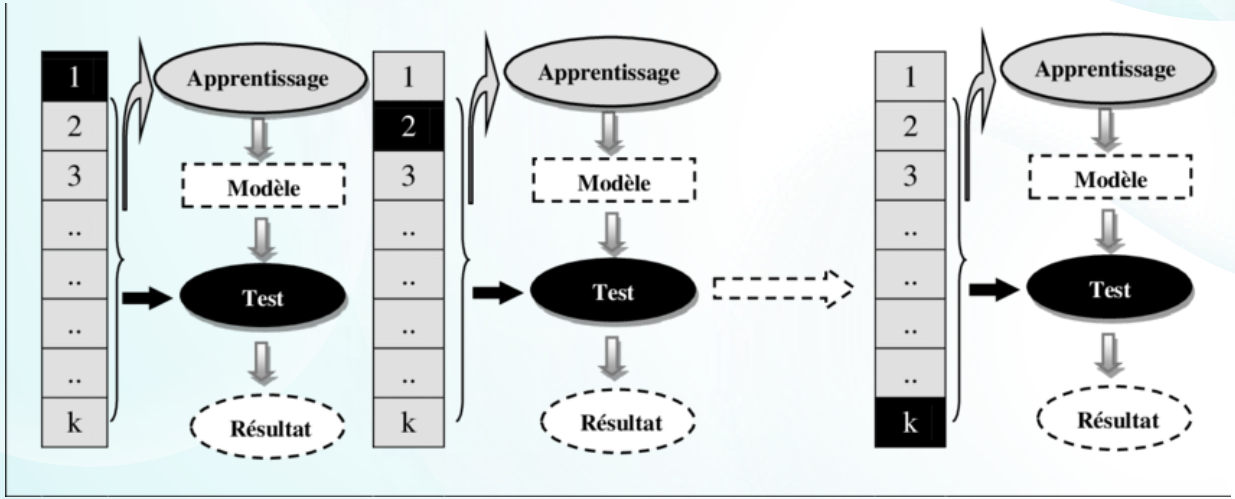



**Figure 2** – Analyse par composantes principales (ACP) de X1 et X2

# III - Application à la base de données “genus”

## III.b La Cross Validation

1. **Division des données en plis :** sous-ensembles d'observations et de variables.
2. **Entraînement et évaluation itératifs :** modèle entraîné sur k-1 plis et testé sur le pli restant.
3. **Mesure de la performance :** évaluation de la performance du modèle avec une métrique appropriée (MSE, Corrélations).
4. **Sélection du modèle :** choix du modèle avec la meilleure performance moyenne sur l'ensemble des plis.



# III - Application à la base de données “genus”

## III.c Analyse des résultats obtenus

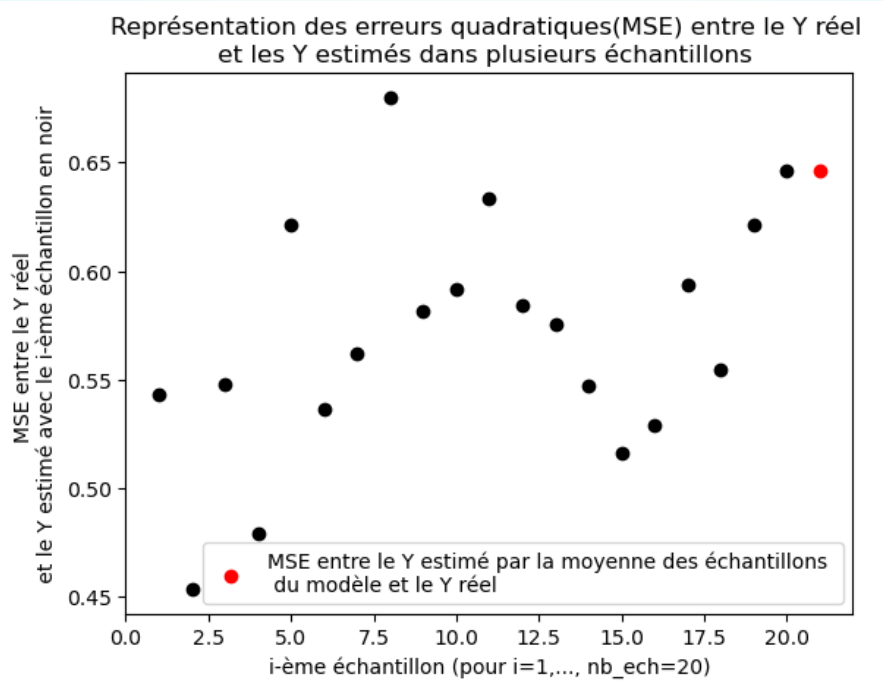



**Figure 3** – MSE entre les Y réel et les Y estimés pour différents échantillons

# III - Application à la base de données “genus”

## III.c Analyse des résultats obtenus

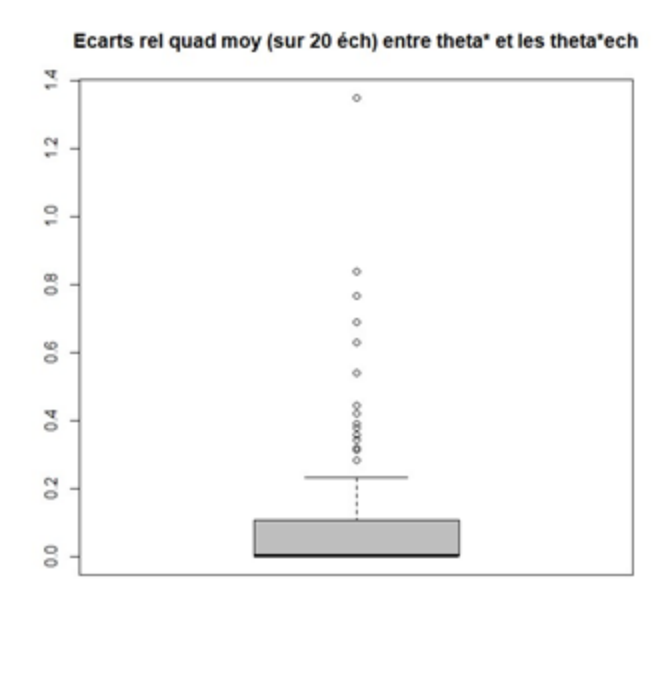



**Figure 4** – MSE entre les paramètres estimées pour les échantillons et les paramètres estimées par le modèle avec le data complet

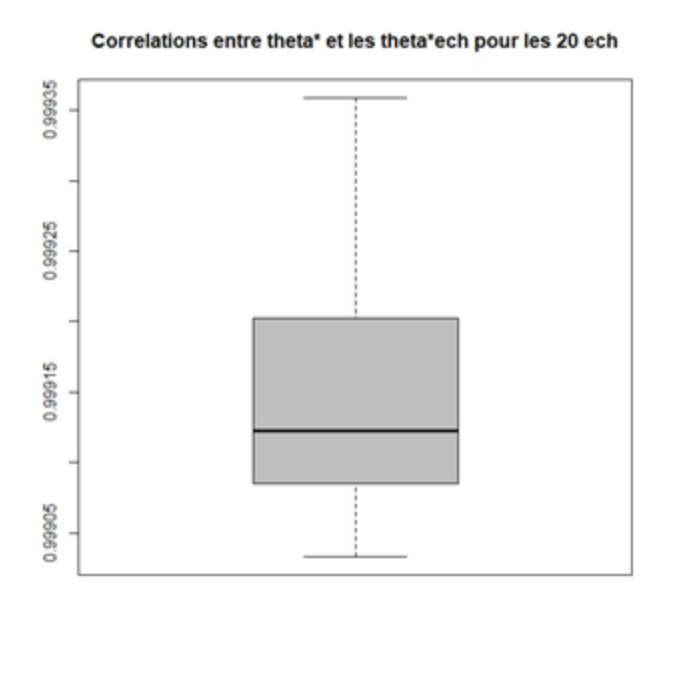



**Figure 5** – Corrélations entre les paramètres estimées pour les échantillons et les paramètres estimées par le modèle avec le data complet

# IV – Application of the algorithm on Penn World Table Dataset

# IV.a - Penn World Table Dataset

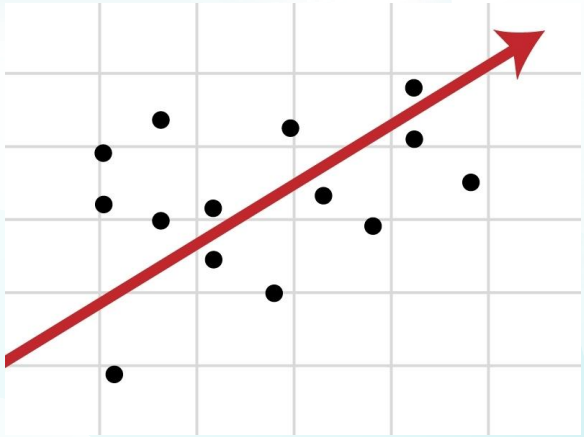
A comprehensive and widely recognized database that provides valuable economic information on countries around the world

A rich set of indicators, including GDP, population, trade, and productivity, enabling comparative analysis and cross-country studies.

An essential resource for understanding global economic trends and making informed decisions.

# IV.b - Our Objective

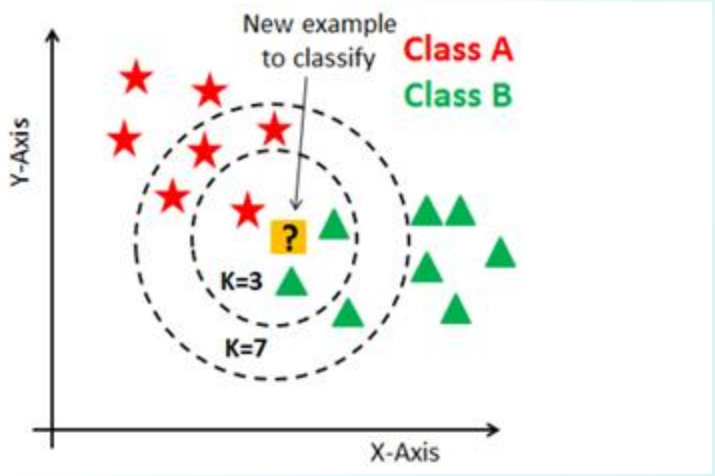
To create a model base on Global Economic Indicators and Sectoral Distribution to predict real GDP performance in different countries using EM algorithm.



# IV.c - Data Cleaning

## Problem?

Due to data availability and limitations in data collection processes across different countries, a substantial number of observations contain missing data




## Solution!

Implementation of the k-Nearest Neighbors (k-NN) imputation technique. We chose k=5 to better capture local patterns

## Why?

This approach ensures that the imputed values align closely with the observed data, thereby minimizing the potential impact on our overall analysis and prediction.

# IV.d - Separation of Variables

## Y1

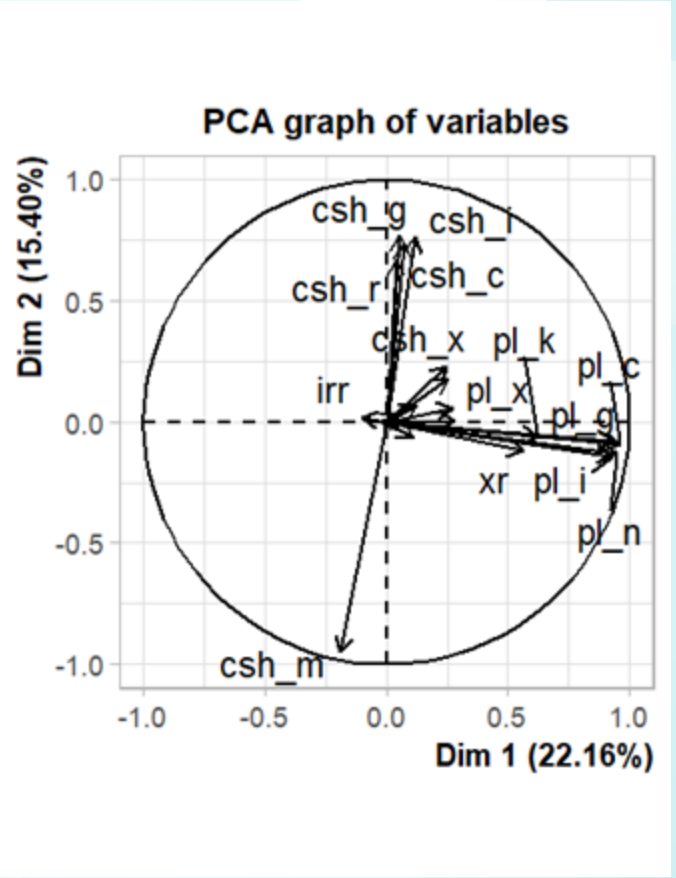
7 indicators of countries' GDP such as Expenditure-side real GDP at chained PPPs, Output-side real GDP at chained PPPs

## X1

Shares in constant GDP across the different sectors of economy

## X2

The Price levels, expenditure categories and capital and some other variables such as exchange rates.



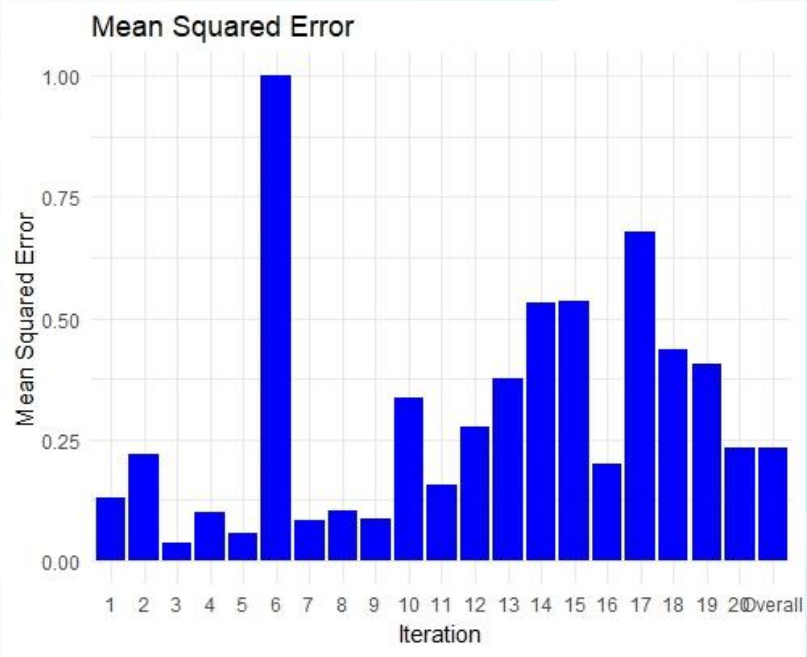
# IV.e - Results

| $c_1$ | $c_2$ | $\sigma_1^2$ | $\sigma_2^2$ | $\sigma_Y^2$ |
|---|---|---|---|---|
| -0.07350239 | 0.07397892 | 0.503451316 | 0.587434636 | 0.001068647 |

Price levels and shares in real GDP have limited impact on real GDP levels, resulting in only small changes based on the observed values. These factors do not significantly contribute to the overall performance of real GDP.

# IV.f - Further Discussions

- Evaluating the performance of our model by Cross validation.
- Significant fluctuations in the MSE values across the iterations.
- A relatively small average MSE
- The small average MSE demonstrates the model's accuracy in generating predictions by effectively capturing underlying patterns and relationships in the data.



# CONCLUSION

- Nous nous sommes approprié le modèle proposé par la thèse, avons réussi à l'implanter et à l'appliquer à différentes bases de données.
- Nous avons tous découvert et grandement progressé en R, un langage de programmation qui nous était tous inconnu.

Pour aller plus loin :

- Gagner en finesse en découpant les données en plus de blocs et en utilisant les matrices T, matrices d'adjonction de covariables permettant aux variables observables d'être déterminées par d'autre facteurs que les variables latentes

- Employer la méthode sur un jeu de données comportant plus d'observations

# MERCI POUR VOTRE ATTENTION

# AVEZ-VOUS DES QUESTIONS?

# COMPETENCE C4

**Repositionnement du Problème:**

- Comprendre une thèse

- Se l'approprier et l'exploiter

- Rendre la méthode rapidement exploitable et compréhensible

**Valeur ajoutée:**

- Analyse plus profonde de la base de données

- Code R prêt à l'emploi mise à jour

- Application sur une autre base de données

# ACP 3 dimensions

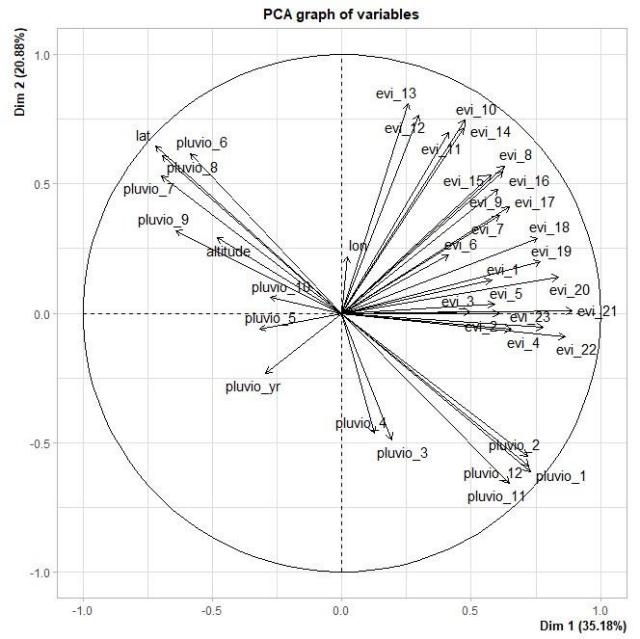



**Figure 6**– Analyse par composantes principales (ACP) de X1, X2 et X3