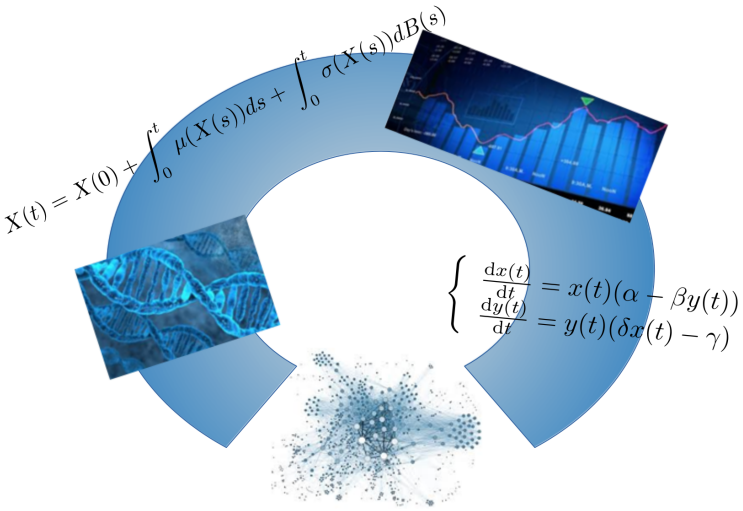
Apprentissage fiable

30/05/2023

Rapport final



Equipe n° 17 :
Minjia FU
Sabrina MOCKBEL
Théophile ROUSSELLE

Client :
Myriam TAMI

Référent :
Gurvan HERMANGE
Véronique LE CHEVALIER
Ioane MUNI TOKE

Pôle n° 15
Modélisation Mathématique des Systèmes Complexes

Table des matières

1 Introduction

Dans le cadre d'un projet à CentraleSupélec nous avons étudié la thèse de Madame Myriam Tami : *Approche EM pour modèle multi-blocs à facteurs à une équation structurelle* [1]. Cette thèse sera notre principale source tout au long du projet, nous n'utiliserons en plus que certains cours de mathématiques dispensé à CentraleSupélec.
L'objectif de ce projet est de compredre la méthodologie pour pouvoir se l'approprier et l'exploiter.
Nous allons donc présenter les points clefs de la thèse sur lesquels nous nous sommmes appuié, pour comprendre la théorie. Puis comment nous avons reproduit certains résultats de la thèse sur une base de données établie. Pour enfin appliqué la méthode à une base de données que nous avons choisi.

2 Élément théoriques

2.1 Présentation

Les modèles d'équations structurelles à variables latentes ont pour intérêt de permettre la modélisation des relations entre des variables observables et non observables. Leur formalisme mathématique se présente sous la forme d'un système d'équations organisé en deux parties distinctes. La première partie est composée d'équations qui modélisent les relations de causalité entre les variables latentes (qui ne sont pas directement observables), les équations structurelles. La deuxième partie est constituée d'équations qui décrivent les relations de causalité entre les variables latentes et les variables observables, les équations de mesure. Ces modèles semblent donc appropriés pour la modélisation et la quantification de systèmes de concepts complexes qui ne peuvent pas être mesurés directement.

2.2 Equation du modèle

2.2.1 Modèle comportant seulement 3 blocs de variables observées

Nous avons donc un jeux de donnée de n lignes avec différentes variables. Nous séparons donc nos varibles en 3 blocs Y,X^1 et X^2 dépendant respectivement des variables latentes g,f_1 et f_2.

On donne l'équation structurelle :

$$g = c_1 f_1 + c_2 f_2 + \epsilon_g \tag{1}$$

Et le système d'équations de mesures :

$$\begin{cases} Y = 1_n \mu_Y + gb + \epsilon_Y \\ X_1 = 1_n \mu_1 + f_1 a_1 + \epsilon_1 \\ X_2 = 1_n \mu_2 + f_2 a_2 + \epsilon_2 \end{cases} \tag{2}$$

Avec les matrices μ_Y,μ_1,μ_2 qui sont respecivement les matrices lignes différentes moyennes de variables composant Y,X^1 et X^2.

On fait ensuite les hypothèses suivantes :
-$f_1 \sim \mathcal{N}(1,\, 0)$
-$f_2 \sim \mathcal{N}(1,\, 0)$
-$\epsilon_1^i \sim \mathcal{N}(0,\, \Psi_1)$ où Ψ_1 est la matrice diagonale composé des variances de chaques variables composant X^1
-$\epsilon_2^i \sim \mathcal{N}(0,\, \Psi_2)$ où Ψ_2 est la matrice diagonale composé des variances de chaques variables composant X^2
-$\epsilon_Y \sim \mathcal{N}(0,\, \Psi_Y)$ où Ψ_Y est la matrice diagonale composé des variances de chaques variables composant X^1

-$\epsilon_g^i \sim \mathcal{N}(1,\, 0)$
-$g \sim \mathcal{N}(0,\, (c_1)^2 + (c_2)^2 + 1)$
-ϵ_g^i est indépedendant de f_1 et f_2 pour tout observation i
-ϵ_Y, ϵ_1et ϵ_2 sont indépendants

Le paramètres de notre modèle est donc $\theta = [\mu_Y, \mu_1, \mu_2, b, a_1, a_2, (\sigma_Y)^2, (\sigma_1)^2, (\sigma_2)^2, c_1, c_2]$ avec les simplification suivantes $\Psi_1 = (\sigma_1)^2 Id$, $\Psi_2 = (\sigma_2)^2 Id$ et $\Psi_Y = (\sigma_Y)^2 Id$

2.2.2 Modèle générale comportant p+1 blocs de variables observées

On peut dans un cas générale définir autant de bloques que l'on veut Y,X^1, X^2 jusqu'à X^p en suivant le système d'équation suivant :

L'équation structurelle :

$$g = c_1 f_1 + c_2 f_2 + ... + c_p f_p + \epsilon_g \tag{3}$$

Et le système d'équations de mesures :

$$\begin{cases} Y = 1_n \mu_Y + gb + \epsilon_Y \\ \forall m \in [[1; p]] X_m = 1_n \mu_m + f_m a_m + \epsilon_m \end{cases} \tag{4}$$

On fait ensuite les hypothèses suivantes :
-$f_m \sim \mathcal{N}(1,\, 0)$
pour tout m -$\epsilon_m^i \sim \mathcal{N}(0,\, \Psi_1)$ où Psi_m est la matrice diagonale composé des variances de chaques variables composant X^m
-$\epsilon_Y \sim \mathcal{N}(0,\, \Psi_Y)$ où Psi_Y est la matrice diagonale composé des variances de chaques variables composant X^1
-$\epsilon_g^i \sim \mathcal{N}(1,\, 0)$
-$g \sim \mathcal{N}(0,\, (c_1)^2 + (c_2)^2 + ... + (c_p)^2 + 1)$
-ϵ_g^i est indépedendant des f_m pour tout observation i
-ϵ_Y, ϵ_m et sont indépendants

2.3 Utilisation d'EM

EM, signifie Expection maximization, ce qui revient à maximiser le maximum de vraisemblance de notre modèle. Cette esimation nous permet :
-à partir d'un paramètre θ^t de calculer différentes espérances (voir la page 56 de la slide pour avoir l'intégralité des formules) et donc la vraisemblance.
-Puis on cherche le maximum selon le critère suivant :

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

qui permet d'optenir la valeur de θ^{t+1}.

Enfin on se donne comme critère d'arrêt :
-Soit un nombre passage maximum dans la boucle à ne pas atteindre que l'on prédéfini comme argument de la boucle.
-Soit : $\sum_{k=1}^{K} \frac{|\theta^{t+1}[k] - \theta^{t}[k]|}{|\theta^{t+1}[k]|} < \epsilon$ avec ϵ la précision désirée indiqué comme argument de l'algorithme

C'est donc cette méthode que nous allons implanter afin de fitter notre modèle sur différentes bases de données.

3 Applications à la base de données genus (données environnementales)

3.1 Introduction

Nous allons appliqué la théorie vu précédemment à la base de donnée genus présentée ci-dessous. Cette base a déjà été utilisée dans la thèse de référence pour y appliquer la théorie. Nous cherchons dans cette partie à reproduire les résultats de la thèse afin de nous assurer que notre méthodologie soit bonne. Ainsi, nous pourrons dans un seconde partie appliqué notre méthode à une autre base de données.

3.2 Présentation de la base de données genus

On applique donc le modèle à la base "genus" fournie dans la package R. Cette base ressence 27 espèces d'abres communes présentes dans la forêt tropical du Congo et la mesure de 39 autres varibales décrivant l'environnement. pour 1000 parcelles de terrains distincts.
Nous cherchons à analyser l'abondance des arbes en fonction des 39 variables environnementales restantes, nous avons donc identifié les 2 varibles composant le bloque Y. Nous réalisons une ACP sur les varibles restantes (Figures 2). De cette ACP on déduit deux bloques selon les deux directions orthogonales X1 et X2. X1 comportera toutes les mesures de chute de pluie ainsi que la localisation (lon, lat, alt) et X2 les autres mesures. On a représenté sur la figure ci-dessous le rôle des données dans genus et ce à quoi on veut parvenir.

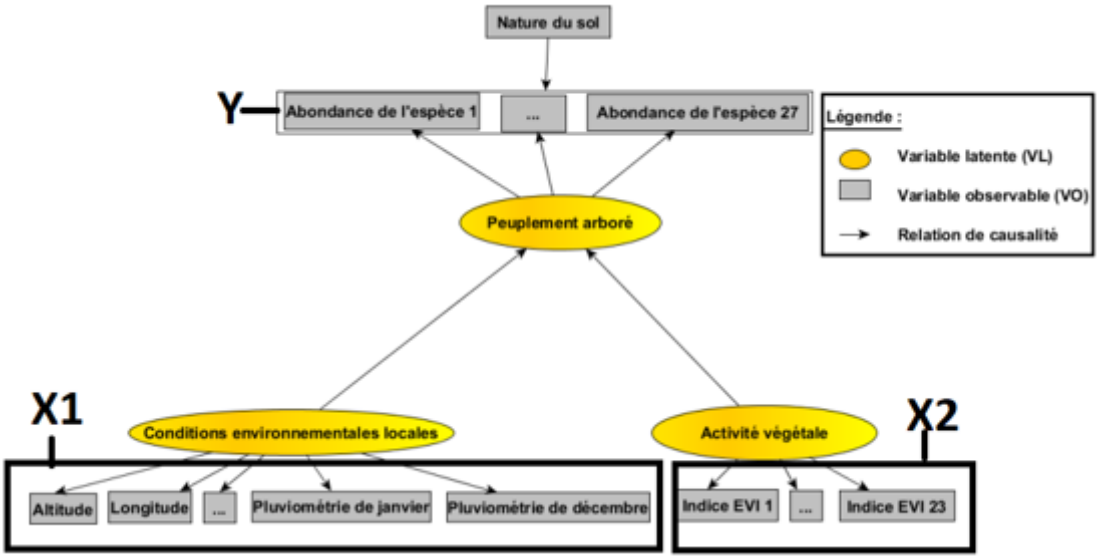


FIGURE 1 – Diagramme structurel du peuplement arboré expliqué par deux blocs de variables observées liées aux conditions environnementales locales et à l'activité végétale

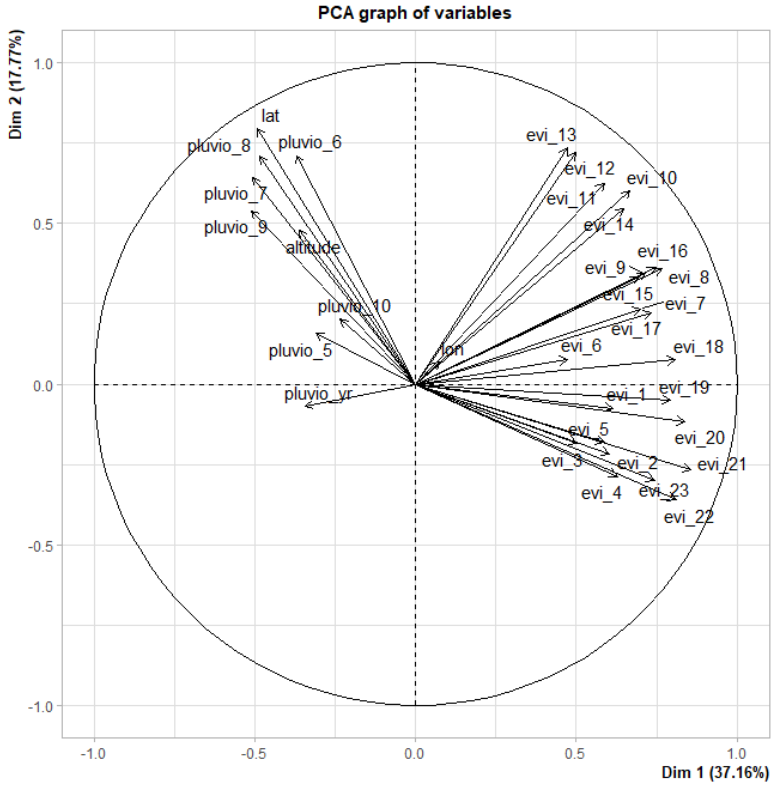


FIGURE 2 – Analyse par composantes principales (ACP) de X1 et X2

3.3 Méthodologie

Nous allons donc fitter notre modèle à l'aide de l'algorithme explicité précédemment. Pour juger de la qualité de notre fit nous utilisons la méthode de la cross validation. Avant d'appliquer l'algorithme nous standardisons les données afin que les valeurs prises restent raisonnables.

3.3.1 Explication générale de la cross validation

La cross-validation est une technique de Data Science utilisée pour évaluer la performance d'un modèle prédictif ou d'apprentissage automatique. Elle consiste à diviser l'ensemble des données disponibles en plusieurs sous-ensembles ou "plis", pour entraîner et tester le modèle de manière itérative .

Voici son fonctionnement :
- **Division des données :** Les données sont divisées en k plis de taille égale. Chaque pli est composé d'un sous-ensemble d'observations et de variables.
- **Entraînement et évaluation :** Le modèle est entraîné sur k-1 plis (appelés ensemble d'entraînement) et testé sur le pli restant (appelé ensemble de validation). Le processus est répété k fois, chaque pli étant utilisé une fois comme ensemble de validation.
- **Mesure de la performance :** Pour chaque itération, la performance du modèle est évaluée en utilisant une métrique appropriée, telle que l'exactitude (accuracy), l'erreur quadratique moyenne (mean squared error) ou l'aire sous la courbe ROC (area under the ROC curve). Les performances sur les k itérations sont ensuite agrégées pour obtenir une estimation globale de la performance du modèle.
- **Sélection du modèle :** Une fois la cross validation terminée, on peut choisir le modèle avec la meilleure performance moyenne sur l'ensemble des plis. Il est important de noter que cette évaluation est plus fiable que l'utilisation d'un seul ensemble d'entraînement et de validation, car elle utilise l'ensemble des données disponibles pour entraîner et tester le modèle de manière équitable.

En bref, la cross validation est une technique utilisée en Data Science qui divise les données en plusieurs plis, entraîne et teste le modèle de manière itérative, puis évalue sa performance globale. C'est un outil essentiel pour estimer la performance d'un modèle et prendre des décisions éclairées sur son utilisation.

3.4 Analyses des résultats obtenus

- Dans le code, on obtient une mesure de la corrélation entre les facteurs estimés et les facteurs réels. Des corrélations plus élevées indiquent une meilleure correspondance entre les facteurs estimés et les facteurs réels, ce qui doit correspondre à une meilleure performance du modèle.
- Les MSE calculées donnent une mesure de l'erreur moyenne entre les facteurs estimés pour chaque échantillon et les facteurs réels trouvés avec le data complet. Cela permet d'évaluer la qualité des estimations des facteurs obtenues à partir de chaque

échantillon par rapport aux facteurs réels du data complet. Une MSE plus faible indique une meilleure correspondance entre les facteurs estimés et les facteurs réels, ce qui doit correspondre à une meilleure performance du modèle.

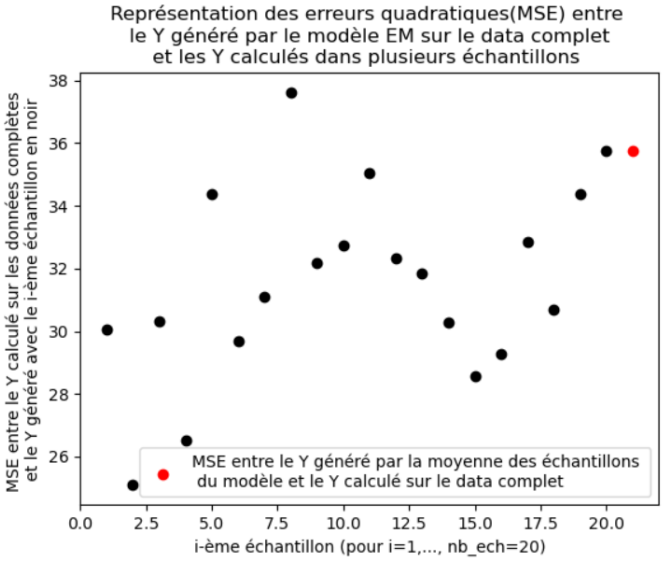


FIGURE 3 – MSE entre les Y générés par le modèle EM sur le data complet et les Y calculés pour différents échantillons

Sur la figure ci-dessus nous avons l'erreur quadratique moyenne obtenue pour différents fits. Les 20 premières erreurs sont celles de la cross validation. La dernière valeur est celle obtenue en utlisant comme paramêtre theta la moyenne de chaques thêtsa de la cross validation. Nous remarquons que l'erreur obtenue est grandement fonction de l'échantillon choisi et que prendre la moyenne des paramêtre n'améliore pas significativement l'erreur. Sur la figure ci desssus, on a représenté les MSE entre les différents paramètres échantillonés et le paramètre généré par le modèle pour le data complet. Ils sont relativement faible (de l'ordre de l'unité), ce qui suggère que le modèle est précis comme expliqué précédemment. Sur la figure ci desssus, on a représenté les corrélations les entre différents paramètres échantillonés et le paramètre généré par le modèle pour le data complet. Ils sont relativement proches, ce qui suggère que le modèle n'est pas précis, nous n'avons pas pu trouver de solutions à ce problème à temps.

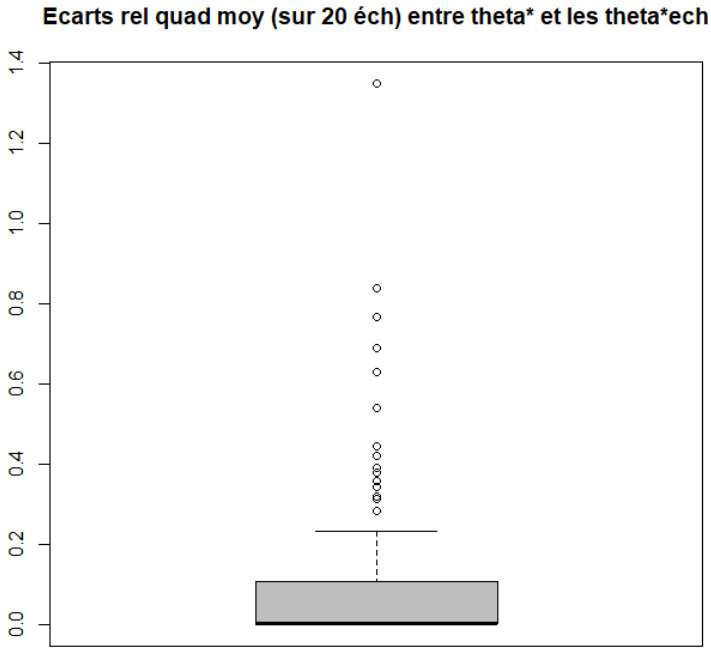


FIGURE 4 – MSE entre les paramètres échantillonnées et les paramètres générés par le modèle avec le data complet

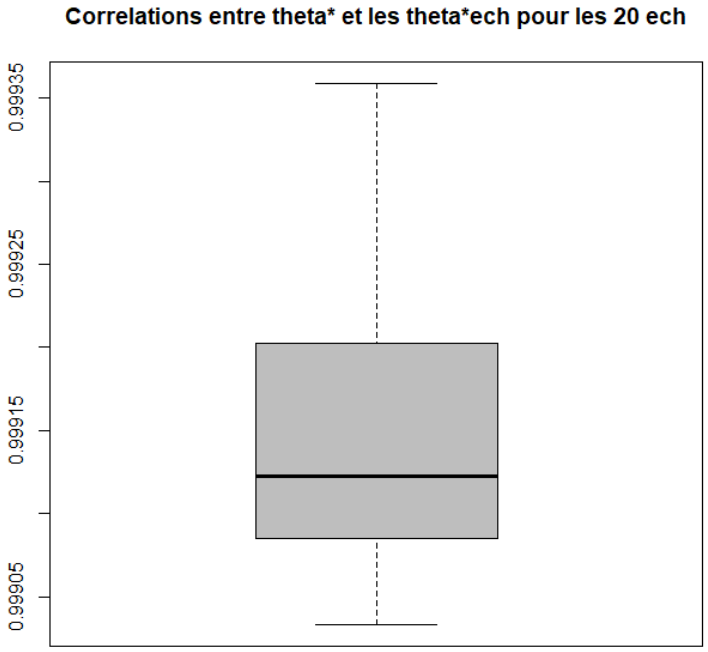


FIGURE 5 – MSE entre les paramètres échantillonnées et les paramètres générés par le modèle avec le data complet

4 Application of algorithm on the Penn World Table dataset

We have chosen the Penn World Table database to test the algorithm. The Penn World Table (PWT) is a renowned database that provides detailed economic information for countries worldwide. Developed by Robert Summers and Alan Heston at the University of Pennsylvania, the PWT offers macroeconomic indicators such as GDP, population, employment, and productivity measures. It employs purchasing power parity (PPP) to adjust for price level differences, enabling accurate cross-country comparisons. With its long-time series data, the PWT facilitates the study of economic trends, growth, and policy impacts.

4.1 Purpose

Our objective is to examine the connection between GDP in different countries and other economic indicators such as price levels and the distribution of Constant GDP at market prices among different sectors or components of the economy using an algorithmic approach.

4.2 Data Cleaning

In order to align our study with our specific research objectives, we have chosen to focus exclusively on data from the year 2000 onwards, disregarding the data from the previous century. This decision allows us to concentrate on more recent and relevant information for our analysis.

However, we have encountered a significant challenge in our dataset regarding missing values. Due to data availability and limitations in data collection processes across different countries, a substantial number of observations contain missing data. These gaps pose obstacles when attempting to apply our algorithm to the dataset effectively.

To address this issue, we have opted to employ the k-Nearest Neighbors (k-NN) imputation technique. By leveraging the inherent relationships between variables, k-NN imputation enables us to fill in the missing values using information from similar neighbouring observations. This approach ensures that the imputed values align closely with the observed data, thereby minimizing the potential impact on our overall analysis and findings.

4.3 Separation of the Variables

The dependent block of variables Y consists of variables which are the 7 indicators of countries' GDP ($q_Y = 7$) in this dataset such as Expenditure-side real GDP at chained PPPs, Output-side real GDP at chained PPPs. To have 2 explanatory groups of X, a PCA is performed on the variables.

The result of PCA gives us two variable bundles with almost orthogonal central directions. One of them X_1 consists of all the variables ($q_1 = 7$) about shares in

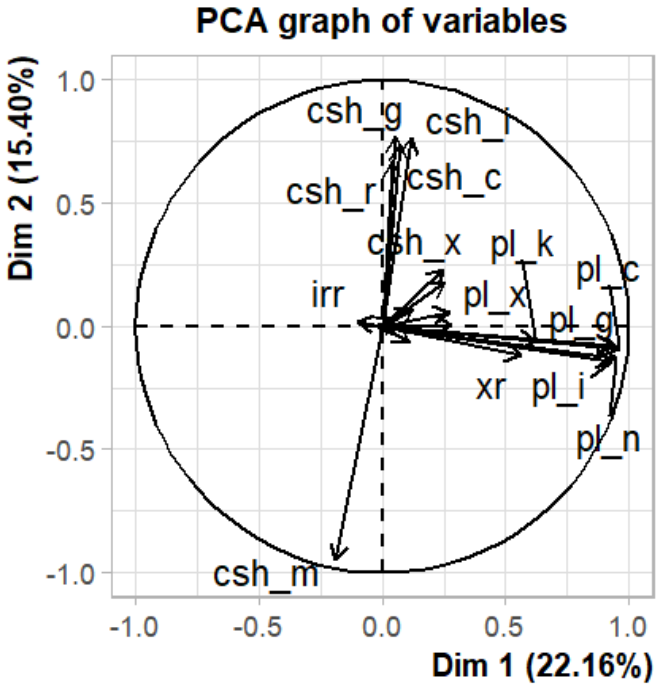


FIGURE 6 – Correlation-Scatterplot yielded by the PCA of the X1 and X2 variables

constant GDP in the market price across the different sectors of the economy and also the real internal rate of return. The other one X_2 ($q_2 = 14$) includes the Price levels, expenditure categories and capital and some other variables such as exchange rates.

In total, we have 3660 observations ($n = 3660$) which include 20 years of economic data for 183 countries.

4.4 Results

c_1	c_2	σ_1^2	σ_2^2	σ_Y^2
-0.07350239	0.07397892	0.503451316	0.587434636	0.001068647

Here we have our average estimates of c_1, c_2, σ_Y^2. We can observe that both price levels and shares in real GDP do not cause a significant impact on real GDP levels.

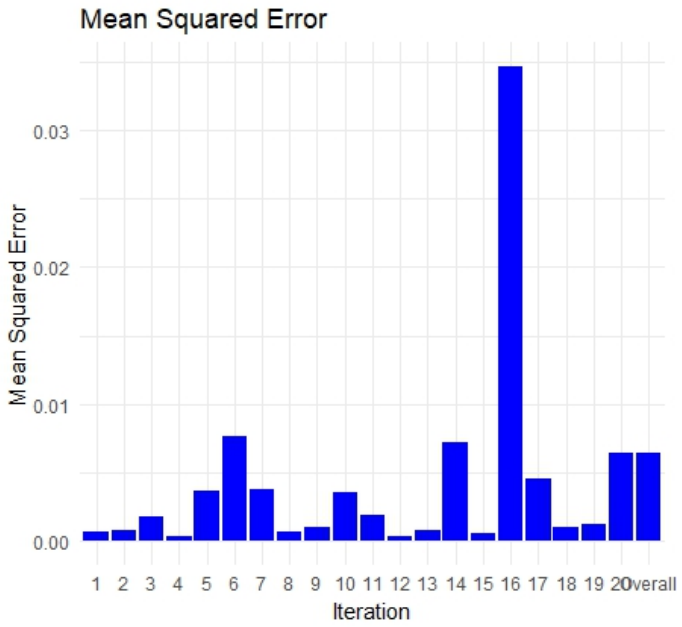


FIGURE 7 – Plot of mean squared error values using cross validation

4.5 Discussion

We conducted an analysis on the Penn World Table Dataset (Fig.7), comparing the results obtained from cross-validation and small-size sampling. We aimed to evaluate the performance of our predictive model.

During the cross-validation process, we performed random sampling multiple times, generating a total of 20 mean square error (MSE) values. We observed significant fluctuations in the MSE values across these iterations. The variability can be attributed to the different samples taken during each cross-validation round.

However, when we calculated the average MSE by taking the mean of all the sampling results (the 21st value), we obtained a relatively small average MSE. Despite the fluctuations, the overall average MSE appeared to be lower.

Given the considerable variability and the relatively small average MSE, we recommend conducting cross-validation multiple times before implementing it in the predictive model. Running cross-validation rounds multiple times would provide a more robust assessment of the model's performance and help mitigate the impact of sampling variations.

5 Conclusion

Ce projet nous a permis de nous initier à une méthode de modélisation innovante. Nous avons pu repoduire des résultat et les étendres à une nouvelle base de donnée.
Cependant, la modèle originale de la thèse donne un modèle plus complexes incorporant notemment des matrices T, adjonction de covariables (ou variables explicatives). Nous pourrions aussi avoir un modèle plus fin en incorporant plus de blocs. De plus nous n'avons appliqué notre méthode qu'à des bases de données comportants un faible nombre de lignes. Rien n'assure que l'algortihme tourne en un temps raisonnables pour un plus grand nombre de données.