ROYAUME DU MAROC
-*-*-*-
HAUT COMMISSARIAT AU PLAN
-*-*-*-*-*-*-

INSTITUT NATIONAL
DE STATISTIQUE ET D'ECONOMIE APPLIQUEE

INSEA

Projet de Fin d'Etudes
****

Identification des facteurs Macro et Micro qui influencent la qualité de l'éducation

(Structural Equation Modelling, Partial Least Square)

Préparé par : ***SIDI Alakè Aïchath***

Sous la direction de : ***Mr SAID NSIRI*** **(INSEA)**
Mr Pierre VARLY **(Varly-project)**

Soutenu publiquement comme exigence partielle en vue de l'obtention du

Diplôme d'Ingénieur d'Etat

Option : Statistique

Devant le jury composé de :

- ***Mr SAID NSIRI*** (INSEA)
- ***Mr F. EL ABDI*** (INSEA)
- ***Mr PIERRE VARLY*** (Varly-Project)

****Juin 2012****

Résumé

"*L'éducation est pour l'enfance ce qu'est l'eau pour une plante.*"
(La Rochefoucauld-Doudeauville)

L'éducation est de nos jours une nécessité. C'est pourquoi l'un des Objectifs du Millénaire pour le Développement est d'atteindre d'ici à 2015 l'éducation primaire universelle. Il importe donc d'identifier les facteurs qui améliorent la qualité de l'éducation.

Bien que de nombreuses études aient été menées, il reste des facteurs qui ne sont pas pris en compte et qui ont tout de même un effet non négligeable sur les rendements scolaires. En effet, les facteurs contextuels propres aux pays en développement tels que la diversité linguistique et la croissance démographique constituent des entraves à la qualité de l'éducation, tandis que l'accès à l'information (comme Internet) peut faciliter la réussite des élèves. Aussi, le Partenariat Mondial pour l'Education, programme multilatéral d'appui technique et financier peut constituer un apport positif à la qualité de l'éducation pour les pays participant, apport que nous tenterons de quantifier.

Notre étude a donc pour but de fournir un modèle assez complet sur la scolarisation des enfants. En plus de cela, nous utilisons une méthode modélisation, celle des moindres carrés partiels (*Partial Least Squares*), qui permet de mieux présenter les résultats et de contourner le problème de colinéarité dont souffrent souvent les variables de l'éducation.

Cette étude a également permis d'obtenir un bon modèle qui permet de prévoir les taux d'achèvement et pourront donc aider les décideurs à améliorer leurs politiques éducatives, plus particulièrement ceux du Burundi grâce à l'analyse de données sur la lecture tirées d'une enquête EGRA (Early Grade Reading Assessment).

Mots clés : Variables contextuelles – Variables de politiques éducatives, Partenariat Mondial pour l'Education, Education primaire universelle, Moindres carrés partiels, Burundi, lecture.

Abstract

"*Education for children is what water is to a plant*"

(La Rochefoucauld-Doudeauville)

Education is a necessity these days. This is why one of the Millennium Development Goals is to achieve by 2015 universal primary education. It is therefore important to identify factors that improve the quality of education.

Although several studies have been conducted, there are still factors that are not taken into account and still have a significant effect on school performance. Indeed, the contextual factors specific to developing countries such as linguistic diversity and population growth is an impediment to quality education, while access to information (such as the Internet) can facilitate student success. Also, the Global Partnership for Education, multilateral program of technical and financial support can make a positive contribution to the quality of education for the countries involved, contribution we try to quantify.

Our study therefore aims to provide a relatively complete model of education for children. In addition, we use a modeling method, the partial least squares (Partial Least Squares), to better present the results and to circumvent the collinearity problem that often plagues the education variables.

This study also yielded a good model that predicts completion rates and can therefore help policy makers improve their education policies, especially those of Burundi through the analysis of data from the reading of an investigation EGRA (Early Grade Reading Assessment).

Keywords: Contextual Variables - Variables of educational policies, Global Partnership for Education, Universal Primary Education, Partial Least Squares, Burundi, reading.

Dédicace

A mes très chers parents,

Je ne serai jamais assez reconnaissance pour tout l'amour et l'assistance que vous m'avez témoignée durant cette période assez déterminante pour moi.

A mes chères frères et sœurs,

Vous êtes pour moi une source de paix et un refuge permanent

A tous mes amis,

Loin des miens, vous avez été d'une aide précieuse et d'un soutient morale que je n'oublierai jamais

A Monsieur Pierre VARLY et à tous ceux qui luttent pour rendre efficace le domaine de l'éducation

A Monsieur Said NSIRI et à tous les professeurs de l'INSEA

A tous ceux qui m'aiment …

Remerciements

Au terme de la rédaction de ce mémoire, je tiens à adresser ma reconnaissance à tous ceux qui de près ou de loin ont participé à son l'élaboration.

Je tiens à remercier Monsieur Pierre VARLY, Directeur de la société Varlyproject, qui a accepté de me recevoir durant toute cette période et qui n'a ménagé aucun effort à m'encadrer efficacement.

Je remercie particulièrement l'équipe d'Education et Territoire Magrheb qui m'ont témoigné un accueil chaleureux durant la durée de mon stage.

Ma sincère gratitude va également à l'endroit de Mr Saïd NSIRI, professeur à l'INSEA qui a accepté d'être mon encadrant, pour ses conseils pertinents et ses mises en garde.

Sommaire

2ème Partie : Analyse Microéconomique (Burundi)

Liste des abréviations

EGRA : Early Grade Reading Assessment

PASEC : Programme d'Analyse des Systèmes Educatifs de la CONFEMEN

CONFEMEN : Conférence des Ministres de l'Éducation des pays ayant le français en partage

OCDE : Organisation de Coopération et de Développement Economique

NER : taux net de scolarisation

ACP : Analyse en composantes principales

PLS : Partial least square (moindres carrés partiels)

SOFRECO : Société française leader dans le conseil et l'assistance technique au développement économique et social durable.

PARSEB : Projet d'appui à la reconstruction du Système Educatif Burundais piloté par la Banque Mondiale.

PIB : Produit intérieur but

GDP : Gross Domestic Product

OCDE : Organisation de Coopération et de Développement Economique

PME : Partenariat mondial pour l'éducation

PNUD : Programme des Nations Unies pour le développement

UNESCO : Organisation des Etats Unis pour l'Education, la Science et la Colature

IREDU : Institute for Research in the Sociology and Economics of Education

CIA : Central Intelligence Agency

EMIS : Education Management Information System

IDA : Association Internationale de Développement (Banque Mondiale

NFHS : National Family Health Survey

SACMEQ : Southern an Eastern Africa Consortium for Monitoring Educational Quality

PISA : Programme for International Student Assessment

ANOVA : Analyse Of Variance

VIF : Inflation of Variance Factor

Liste des tableaux et graphiques

Introduction

Avec le temps, le champ de l'*éducation* s'est bien élargi. Auparavant, l'homme comprenait l'*éducation* comme une simple relation entre éducateur et éduqué. Or, notre époque conçoit l'*éducation* d'une manière plus ample que cette relation binaire. En effet, plusieurs facteurs qui ont été souvent négligés entrent en jeu et influent sur la capacité de l'enfant à profiter pleinement de l'Education.

Si autrefois, le rôle de l'éducation dans la société et son importance pour la croissance économique ont été discutés, aujourd'hui, les chercheurs se concentrent plus sur les moyens de perfectionnement de l'éducation. En effet, l'essor des pays d'Asie a été une conséquence des différentes découvertes en technologie rendues possibles par l'accumulation du capital humain à travers l'éducation.

Il est donc important d'identifier les réels facteurs, autres fois négligés qui influenceraient directement ou indirectement l'éducation. Il y a eu certes, plusieurs études qui ne convergent pas en termes de résultats à cause des limites de certaines méthodes statistiques, vu la complexité des variables éducatives et plus généralement de l'étude des comportements humains.

Depuis que les pays se sont engagés sur la voie de la scolarisation primaire universelle en 2000, les systèmes statistiques nationaux ont été renforcés avec l'aide des bailleurs de fonds et les données sur l'éducation se sont multipliées dans les pays du Sud. Il est donc possible désormais d'introduire des variables relatives à ces pays (comme la diversité linguistique) et de développer des modèles qui leur sont propres. En effet, ces variables contextuelles n'ont pas été prises de manière systématique ans les modèles expliquant la qualité de l'éducation alors qu'elles ont potentiellement de grandes influences.

L'objectif de notre projet de fin d'études est d'élargir la gamme des variables déjà prises en compte comme influant sur l'éducation et de voir leur impact. D'autre part, nous présentons ici une méthode de 'modélisation douce' Partial Least Square, fiable et adaptée aux variables liées à l'éducation et expliquons chacune des étapes suivies pour aboutir à nos résultats.

Nous allons partir d'une vue macroéconomique, pour voir globalement les traits communs aux pays qui réussissent et identifier les facteurs de freins pour d'autre. Nous allons débuter notre analyse macroéconomique par une étude préliminaire où nous allons identifier les variables qui sont corrélées entre elles et identifier les pays qui n'ont pas un bon taux d'achèvement. Nous ferons ensuite une analyse en composantes principales (ACP) pour voir si les variables considérées peuvent être réduites en composantes. Nous allons ajuster un modèle par les moindres carrés ordinaires et finir par la modélisation des moindres carrés partiels (PLS).

Etant donné que les facteurs macroéconomiques à eux seuls ne peuvent expliquer l'aboutissement des élèves du primaire, nous allons voir ce qu'il en ait pour le Burundi en nous intéressant aux données microéconomiques. Nos individus seront donc dans ce cas les

élèves, nous allons expliquer leur réussite en lecture en 2012 par leur niveau initial mesuré en 2011, leur condition de vie, les conditions de l'école, les méthodes de l'enseignant et des observations en classe. Nous partirons d'une étude préliminaire où nous allons voir les variables qui sont corrélées au nombre de mots lus par minute, pris comme indicateur des compétences en lecture. Nous ferons ensuite un modèle avec les moindres carrés ordinaires et nous finirons par l'application des moindres carrés partiels.

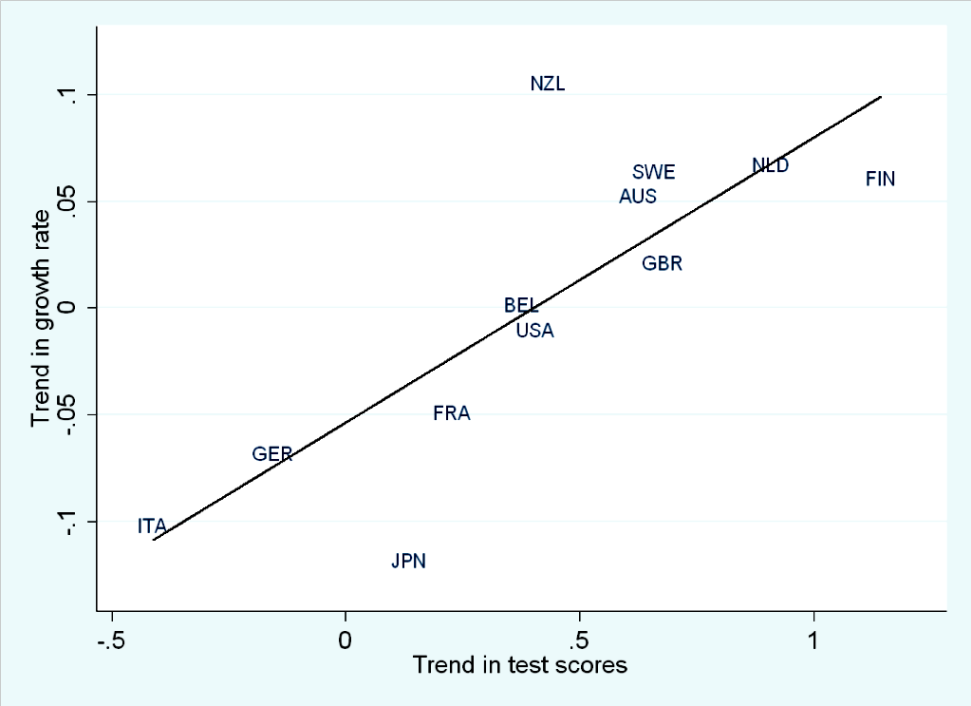
A tout ceci suivra une conclusion générale où ne ferons le point sur les deux études et en tirerons des conclusions opérationnelles et jugerons de l'intérêt des méthodes PLS pour analyser les données sur l'éducation.

Première partie :

I- Problématique étudiée :

Plusieurs études sur l'éducation l'ont considérée comme un facteur qui faciliterait la croissance économique, mais le lien de causalité entre l'éducation et la croissance économique n'est pas univoque (dans un seul sens). L'éducation a un impact positif sur la croissance économique et la croissance économique est aussi un facteur favorisant la bonne qualité de l'éducation. Cette relation est positive, comme nous pouvons le voir à travers ce graphe qui nous montre pour quelques pays l'évolution de la croissance économique en fonction des résultats aux tests de compétences des élèves (mesure de la qualité de l'éducation).

Graphique1 : Evolution de la croissance économique en fonction des résultats aux tests



Source : Hanushek & Woessmann(2010).

Toutefois, selon Doudjidingao (2011), le modèle le plus bénéfique est d'étudier l'effet de la croissance économique sur l'éducation. La croissance en effet permet d'augmenter les moyens mis pour améliorer la qualité de l'éducation, qui à sont tour impacte la croissance par la main d'œuvre qualifiée qu'elle fournit. Le but de notre étude est d'expliquer les facteurs macros et micro agissants sur l'éducation. Nous avons deux types de facteurs, les facteurs de contexte et les facteurs sur lesquelles les autorités en charge de l'éducation peuvent agir (facteurs de politique éducative). Notre objectif est donc de cadrer les politiques

d'amélioration de la qualité de l'éducation en montrant l'importance de certaines variables dans l'explication des différences de taux de scolarisation ou niveaux de connaissances d'un pays à un autre, d'une tranche de revenu à une autre. Certains pays ont également bénéficié d'une aide technique et financière à l'éducation en participant au Partenariat Mondial pour l'Education. Nous allons aussi voir l'impact de cette participation sur l'éducation à travers une variable qui est la durée de participation des pays à ce partenariat afin de répondre à la demande du Conseil d'Administration du Partenariat Mondial pour l'Education.

D'après Hanushek et Kimko (2000), au niveau micro, les facteurs éducatifs (formation des enseignants, taille des classes) n'arrivent pas à contrebalancer les facteurs contextuels, liés à la pauvreté. Au niveau macro, il existe un écart considérable de niveaux d'acquisitions scolaires entre économies développées et économie en développement, qui peut être estimé à quatre années de scolarisation. Selon Ross (2009), un enfant de 6ème année, par exemple en Afrique Australe, a un niveau de connaissance semblable à un enfant de 2ème année dans un pays d'Europe ou d'Amérique du Nord. Afin de comparer ce qui est comparable, nous allons nous restreindre à l'étude des pays en développement ou plus précisément à bas revenus ou à moyen revenus de la tranche supérieure où la variation de niveau des élèves est également importante. On cherchera donc à expliquer la variance de la qualité de l'éducation entre économies en développement, en prenant soin d'élargir la gamme de facteurs contextuels mesurée traditionnellement dans la littérature.

Dans les études précédentes, plusieurs variables n'ont pas été prises en compte et qui peuvent avoir un impact important sur la qualité de l'éducation, du moins au niveau micro. Il s'agit de la diversité des langues, de l'accès à l'internet, de la gouvernance (ou niveau de corruption) mais également de la quantité de données éducatives fournies, en soit un indicateur de transparence et de gouvernance. Il faudra donc mesurer l'importance de ces variables. Nous allons également tester l'apport du Partenariat Mondial d'aide à l'Education.

Cette étude a été menée dans les locaux de VARLYPROJECT et vise à apporter de la valeur ajoutée aux rapports produits par la société dans le cadre d'un contrat avec le Partenariat Mondial d'aide à l'éducation et le Burundi.

II- Présentation de l'organisme

VARLYPROJECT, qui m'a accueillie durant quatre mois de stage, est une SARL spécialisée dans l'éducation et travaille principalement pour l'Afrique. Elle est installée à Rabat (Maroc) depuis le 21 juin 2010. Varlyproject s'engage dans la recherche et le développement à travers un blog et la production d'articles pour le compte d'organisations non gouvernementales comme One Laptop Per Child France. La société est dirigée par Pierre VARLY, statisticien, titulaire d'un Master en Econométrie et ayant plus de 10 ans d'expérience dans l'analyse de données sur l'éducation. En tant que coordonnateur du Programme d'Analyse des Systèmes Educatifs de la CONFEMEN (PASEC), à Dakar de 2005 à 2009, Monsieur VARLY a encadré une équipe de six jeunes statisticiens appuyée ponctuellement par des contractuels et des stagiaires. Le PASEC a pris régulièrement en stage deux stagiaires de l'école de statistique de Dakar, mais également accueilli des élèves de l'Ecole Polytechnique par le passé. Un des stagiaires, Makan Doumbouya, a reçu un prix international de Statistique grâce à un article tiré de son stage.

Varlyproject est donc particulièrement impliquée dans la formation des jeunes des pays du Sud, la clé du développement.

Les domaines de compétences de VARLYPROJECT sont les suivants :

- L'évaluation des acquis scolaires

- L'analyse des données d'enquêtes

- L'économie de l'éducation

- La planification et l'appui à l'élaboration de politiques éducatives

Elle exerce ses activités pour le compte de gouvernements ou d'organismes privées, où la société agit comme sous-contractant ou en contrat direct. VARLYPROJECT a obtenu deux contrats :

- Un pour une analyse des données sur la qualité des pays participant au partenariat Mondial pour l'Education
- L'autre pour une évaluation des compétences fondamentales en lecture au (EGRA) au Burundi, pour le compte de la SOFRECO[1] et du PARSEB[2].

Ces deux études permettant d'obtenir deux jeux de données originales. Dans un des rapports présentés au Comité directeur pour le Partenariat Mondial pour l'Education, Proman (2011) suggère de mettre en œuvre des méthodes de *partial least square* pour analyser l'impact du partenariat :

"To further analyse the determinants of educational trends a secondary statistical analysis of existing data sets is recommended. Whereas the quality "rigorized" impact evaluations focus on the effects of causes for the secondary statistical analysis the causes of effects are brought

[1] Société française leader dans le conseil et l'assistance technique au développement économique et social durable.
[2] Projet d'appui à la reconstruction du Système Educatif Burundais piloté par la Banque Mondiale.

*to the fore. Caused by several limiting aspects it is recommended to work with a structural equation model based on **partial least squares** analysis as this approach seems to be feasible even for non-experts."*

Nous allons donc mettre en œuvre ces techniques sur des données macro puis voir comment les employer sur des données micro-économiques.

1ère Partie : Analyse Macroéconomique

I- Construction des bases de données et des indicateurs

1-1) Revue de la littérature

Le rôle de l'éducation dans la croissance globale d'un pays étant, comme mentionné plus haut, crucial, plusieurs chercheurs se sont attelés à identifier les facteurs les plus déterminants. Plusieurs chercheurs ce sont contredits par rapport à l'influence de certains facteurs, surtout du côté des dépenses dans l'éducation. Ce qui est aussi dû au manque de fiabilité de certaines méthodes statistiques ou du non respect des contraintes d'hypothèses. Nous verrons plus loin les limites de certaines méthodes, à cause du choix de la méthode utilisée. Nous présentons ici quelques études principales qui ont été faites autant sur le plan macro que micro, et leurs résultats.

1-1-1) Modèles Testés dans la littérature

Hanushek & Kimko (2000) ont travaillé sur un échantillon de 39 pays, principalement des pays à haut revenu par habitant. Ils ont combiné les notes des tests en math et en science pour calculer les scores des élèves. Ensuite, ils ont régressé les scores obtenus après évaluation des élèves sur les variables qui étaient les mesures classiques de l'éducation : la taille de classe et les dépenses de l'Etat dans l'éducation.

Lee & Barro (2001) quant à eux ont utilisé une base de données panel et ont considéré les résultats en maths, sciences et lecture pour des élèves de différents âges nés de 1964 à 1991. Ils ont ensuite effectué une régression des scores obtenus en prenant comme variables explicatives la taille de classe, les dépenses de l'Etat dans l'éducation, le salaire des enseignants et le taux de répétition.

Enfin, Altinok (2010) a testé trois modèles avec comme variable dépendante les scores obtenus à différents tests sur un échantillon de pays incluant de nombreux pays en développement. Premièrement, il a régressé les scores obtenus avec les variables Produit Intérieur Brut (PIB) par personne, l'alphabétisation des adultes, le salaire des parents, la dépense dans l'éducation et le nombre d'élèves par enseignants (toutes sont en logarithme sauf la dernière). Ensuite, a ajouté à ce modèle le carré du nombre d'élèves par enseignant pour tenir compte d'éventuels effets de seuil. Enfin, dans le troisième modèle il a remplacé le carré du nombre d'élèves par enseignant par la moyenne du taux de croissance économique annuel. Il a ensuite testé les mêmes modèles par la méthode des effets fixes. Le travail Altinok constitue un bon point de départ pour notre étude car :

- La base inclus des pays en développement
- De nombreuses variables sont introduites dans le modèle
- Plusieurs spécifications des modèles sont testées

1-1-2) Résultats des modèles

Hanushek & Kimko (2000) ont conclus que les conditions qu'offrent l'école (la taille de classe au primaire et les dépenses de l'Etat dans l'éducation), qui étaient les mesures classiques de l'éducation (Altinok, 2010) n'ont pas d'effet significatif sur la qualité de l'éducation.

Lee & Barro (2001) quant à eux ont montré que les conditions scolaires considérées par Hanushek et Kimko, en plus du salaire des enseignants, soit la taille de classe au primaire, les dépenses de l'Etat dans l'éducation, et le salaire des enseignants ont un impact positif sur les acquis des élèves. Aussi, le taux de répétition est positivement corrélé à la taille de classe. Verhoeven & Tiongson (1999) ont montré l'importance de distinguer les résultats selon les capacités économiques des pays.

Altinok (2010) a trouvé que l'influence des parents mesurée par le taux d'alphabétisation des adultes a un effet positif et significatif sur la qualité de l'éducation et ce surtout pour l'école primaire. Un résultat surprenant est que l'augmentation du salaire des enseignants a un effet négatif sur la qualité de l'éducation surtout au primaire. Le ratio élève-enseignant a un effet insignifiant. En effet, lorsqu'un pays dépense beaucoup pour payer ses enseignants, il a moins de budget pour des dépenses telles que l'achat de matériel didactique et la formation des enseignants.

La critique que l'on peut adresser à ces travaux est souvent une inférence trop vite conclue, alors que certains facteurs contextuels ont peut être été omis. De plus, les modèles comportent plusieurs variables logiquement liées entre elles et les équations structurelles doivent nous permettre de modéliser ces relations. Cette remarque est particulièrement valable avec le taux d'alphabétisation de la population adulte et les taux de scolarisation au primaire actuels.

Ces résultats changent lorsqu'on considère les pays selon leur classification, par exemple l'influence des parents ou le taux d'alphabétisation des adultes perd son effet significatif pour les pays à haut revenu, devient négatif pour les pays à moyen revenu et positif et très signifiant pour les pays à faible revenu. Pour ce qui du salaire des enseignants l'effet demeure négatif, mais uniquement significatif pour les pays à moyens et à faible revenus.

Pour le secondaire, seule la dépense par étudiant a un effet positif et significatif. Mais lorsqu'on utilise le taux net de scolarisation comme moyen de mesure de la qualité de l'éducation, on note un effet positif et significatif du niveau de scolarisation des adultes. Le salaire des enseignants a toujours un effet négatif et significatif. Il en est de même pour la taille des classes et le taux de redoublement.

Pour mesurer l'impact de la capacité économique, Altinok (2010) a distingué entre les pays de l'Organisation de Coopération et de Développement Economique (OCDE) et les

autres. L'alphabétisation des adultes n'a pas n'ont plus un effet significatif lorsqu'on opère cette distinction, ainsi que la dépense en éducation. Le salaire des enseignants a un effet négatif pour ce qui est des pays développés et en développement. La taille des classes a un effet positif pour les pays en développement et négatif pour les pays développés. Le taux de redoublement a un effet négatif significatif pour les pays développés. En utilisant le taux net de scolarisation[3], le niveau d'alphabétisation des adultes est positif et significatif surtout pour les pays développés.

Pour conclure, les différentes considérations effectuées n'ont pu déterminer un effet significatif de l'augmentation des ressources allouées à l'école, selon les auteurs les effets des variables ne sont pas les mêmes et les effets varient selon la tranche de revenu. Nous pouvons émettre trois hypothèses : soit le modèle est biaisé à cause de l'harmonisation des scores de différents pays (hypothèse difficilement vérifiable) ou il n'y a véritablement pas d'effet de l'augmentation de ces ressources ou le modèle est mal spécifié (omission de certaines variables, problème de choix du type de modèle).

Les études passées sur la fonction de production de l'éducation (Altinok, Hanushek) ont en général essayé d'expliquer le lien entre les résultats scolaires et l'évolution de l'économie (Lee-Barro), ou d'identifier les déterminants de résultats scolaires. Ces études ne prennent pas en compte la qualité de l'information et les indices de gouvernance, ni certains facteurs contextuels pourtant identifiés dans les études micro. Cela crée donc un biais, que l'on peut caractériser de biais d'omission et qui n'est pas mesuré. De plus, ces études se focalisent sur les pays qui ont fourni des données, et il y a donc un biais de sélection, car les pays qui ne fournissent pas de données ont possiblement des caractéristiques différentes des autres.

Nous pouvons récapituler ces résultats dans le tableau suivant :

[4]Tableau1 : Effet des variables testées dans la littérature

Variables	Description	Auteurs	Effet
PTR	Ratio élèves par enseignant	Hanushek & Kimko	NS
		Lee & Barro	-
		Altinok	NS
Expend	Dépenses d'éducation	Hanushek & Kimko	NS
		Lee & Barro	+
		Altinok	NS
Teacher_Pay	Salaire des enseignants	Lee & Barro	+

[3] Taux net de scolarisation : rapport du nombre d'inscrits ayant l'âge officiel de scolarisation au nombre total d'enfants ayant l'âge officiel de scolarisation.

		Altinok	-
Literacy	Taux d'alphabétisation	Altinok	+
repet	Taux de répétition	Lee & Barro	-
		Altinok	-

Source : auteur à partir de la revue de littérature

1-2) Nouvelles variables introduites

Les modèles énumérés précédemment nous donnent donc un point de départ et une base théorique, mais nous pouvons ajouter d'autres variables :

- Langues : nous pensons que la diversité de langues dans un pays peut nuire à la qualité de l'éducation. En effet, si les élèves d'une même classe ne partagent pas la même langue maternelle, il leur sera impossible de leur inculqué des notions dans la langue qu'ils comprennent le plus (leur langue maternelle) surtout que notre étude porte uniquement sur l'école primaire. Il serait aussi difficile à l'Etat de permettre l'enseignement primaire dans les langues nationales quand elles sont nombreuses.
- Internet ou le taux d'internaute : comme proxy de l'accès à l'information, plusieurs sites Internet facilitent non seulement l'épanouissement des jeunes mais aussi des apprentissages de diverses sortes, nous pensons donc que cette variable peut avoir un effet positif sur la qualité de l'éducation. En règle générale, en économie les questions d'accès à l'information et d'ouverture sont centrales et incluent dans des modèles de croissance. Nous pensons qu'en éducation un tel raisonnement peut être appliqué. C'est aussi un indicateur d'investissement dans les nouvelles technologies et donc de développement.
- Corruption : l'indice de Transparancy International nous renseigne sur le degré de corruption de chaque pays ou de transparence. En effet, plus un pays est corrompu, plus les dépenses prévues par l'Etat pour l'éducation ne seront pas toutes investies réellement. Nous pensons donc ajuster la variable de 'dépense dans l'éducation' par un indice de corruption sorte de proxy du taux de déperdition entre dépenses prévues et dépenses touchant véritablement les écoles. Ces taux ont été parfois estimés dans des enquêtes spécifiques dites *Public Expenditure Tracking Survey*, mais sur trop peu de pays pour être pris en compte dans nos analyses.
- Population en 2009 : Il s'agit de tester si la taille de la population a un effet significatif sur la qualité de l'éducation. En effet, en dehors des Etats fédéraux (comme l'Inde ou le Nigéria), la taille de l'administration de l'éducation n'est pas véritablement proportionnelle à la taille du pays. On peut faire l'hypothèse que les pays les plus peuplés (à taille de l'administration centrale équivalente) sont plus difficilement gérables que les moins peuplés.
- Croissance démographique : c'est la moyenne du taux de croissance de la population de 2004 à 2009. Nous pensons que les pays à forte croissance démographique ont plus

de mal a avoir une bonne qualité de l'éducation que les autres. Cette relation est avérée mécaniquement lorsque l'on considère le taux d'achèvement comme variable réponse. En effet, ce taux est calculé comme étant les nouveaux entrants en fin de cycle primaire divisé par la population en âge d'achever le cycle. Plus un pays a un taux de croissance démographique, plus il doit faire d'effort pour scolariser les élèves.

- Fastrack : c'est une variable indicatrice qui permet de mesure la différence entre les pays qui ont reçu une aide à l'éducation par rapport à ceux qui n'en ont pas reçu, à travers la participation au Partenariat Mondial pour l'Education.
- Dur_gpe : c'est la durée en année de participation au Partenariat Mondial pour l'Education.

Pour récapituler, les variables présentes dans le modèle sont :

Tableau2 : Description des variables prises en compte dans le modèle

Nom Variable	**Descriptif variable**	**Sources**	**Concept mesuré**	**Effet potentiel**
Completion rate (Complet)	Le taux d'achèvement du primaire	Module Edstats + Partenariat Mondial pour l'Education		
Primary score (primary)	Les scores des élèves aux différents tests internationaux.	Base de données Altinok		
Adult Literacy (Literacy)	Le taux de scolarisation des adultes	Module Edstats		Les adultes ou parents alphabétisés ont tendance à mettre leurs enfants à l'école et à les encadrer.
Internet users (Internet)	Le nombre d'internautes pour 100 habitants	Module Edstats	Facilité de l'accès à l'information des parents et des élèves	Les questions d'accès à l'information et d'ouverture sont souvent inclus dans les modèles de croissance.
Log GDP per Capita (loggdp)	Le log du PIB par habitant est une mesure inclus dans l'indice du PNUD	Banque Mondiale	Revenu des parents, niveau de développement économique	Si les parents ont un revenu élevé, ils auront donc les moyens de scolariser leurs enfants et de les faire encadrer, c'est aussi un indicateur macroéconomique de la richesse des pays.
female	Le taux de femmes	Module Edstats	Indique le degré de parité genre dans	Plus il y a de femmes enseignantes, plus les parents scolarisent leurs

	enseignantes		l'enseignement	filles.
ptr	Le nombre moyen d'élèves par enseignant	Module Edstats	La taille de classe dans les pays.	Plus il y a d'élèves dans une classe, moins la qualité de l'enseignement est bonne.
repet	Le taux de redoublement	Module Edstats	Efficacité interne du système éducatif	Un grand nombre de redoublants indique une mauvaise qualité de l'éducation
popgrowt	Croissance annuelle démographique.	Nations Unies		Une forte croissance entraine des difficultés d'accès à l'éducation (surcharge des classes, coûts).
Log population (logpop)	Le log de la population en 2009	Nations Unies	Taille de la population et difficulté à gérer le pays	Les Etats ayant une forte population ont surement plus de difficultés à assuer une bonne qualité de l'éducation.
Corruption (corrupt)	Indice de mesure de la corruption		Mesure du degré de corruption (ou de transparence) dans le pays.	Plus l'Etat est corrompu moins les dépenses prévues dans l'éducation sont effectivement réalisées.
private	La part des écoles privées présentes dans le pays	Module Edstats	Mesure la dominance du privé dans le pays.	Les écoles privées forment parfois mieux que les écoles publiques.
expend	La part de la dépense publique en éducation dans l'enseignement primaire	Module Edstats	Mesure de l'investissement de l'Etat dans l'éducation.	Plus l'Etat investit dans l'éducation, plus on devrait avoir de bons résultats.
Language	Le taux de diversité de langues du pays	Statsilk à partir d'Ethnologue (SIL)	Mesure de la diversité linguistique	les enfants qui n'apprennent pas dans leur langue maternelle peuvent avoir des difficultés
response	Score pays montrant le taux de données fournies		Mesure de la transparence dans la fourniture des données.	Les pays ayant une bonne qualité de l'éducation fourniraient davantage de données.

Dur_gpe	La durée de participation au Partenariat Mondial pour l'Education	Site du Partenariat Mondial pour l'Education	Mesure de la durée de participation au Partenariat Mondial pour l'Education	La durée de participation à cette initiative devrais avoir un effet positif su la qualité de l'éducation (plus financements, plus d'appuis techniques).
fastrack	Variable indicatrice, notant l'appartenance ou pas à l'initiative d'aide à l'éducation	Site du Partenariat Mondial pour l'Education	Appartenance au Partenariat Mondial pour l'Education	Effet de l'initiative d'aide à l'éducation

Fait par les auteurs

1-3) Mode de collecte des données

La plupart des statistiques proviennent de l'enquête annuelle réalisée par l'Institut de Statistiques de l'UNESCO (Union) auprès des Etats membres qui sont publiées sur le module Edstats de la Banque Mondiale. Les pays membres de l'initiative Site du Partenariat Mondial pour l'Education (PME) fournissent certains indicateurs lors des revues sectorielles, notamment des données financières.

- Module Edstats accédé en février 2012, à partir des statistiques de l'Institut de Statistiques UNESCO
- Site de la Banque Mondiale
- Site web du Global Partnership for Education février 2012
- Institut de Statistiques de l'UNESCO (2011), Le financement de l'éducation en Afrique Subsaharienne, relever les défis de l'expansion, de l'équité et de la qualité, UNESCO BREDA, UNESCO IIPE, UNESCO ISU, Montréal.
- Institut de Statistiques de l'UNESCO (2011), Global education digest 2008, UNESCO- ISU, Montréal.
- Indice de perception de la corruption : Transparancy International
- Diversité linguistique : Statsilk Ethnologue.
- Population et croissance : Les données sur la population en 2009 et le taux de croissance ont été prise sur le site des Nations Unies. Dans cette base, il n'y avait pas de données sur les pays suivants: Gambie, Tanzanie, Bolivie, Côte d'Ivoire, Congo, République Démographique du Congo, Egypte, Kyrgyzstan, Laos, Moldavie, Vietnam, Cisjordanie et le Yémen. Pour pallier à ce manque de données nous avons pris sur le site de l'Agence Centrale de Renseignement (*CIA World Factbook)* la population et le taux de croissance en 2012. Nous avons donc calculé la population de ces pays en 2009 en considérant comme fixe le taux de croissance.

$$Pop_{2009} = Pop_{2012}*(1\text{-taux})^{-3}$$

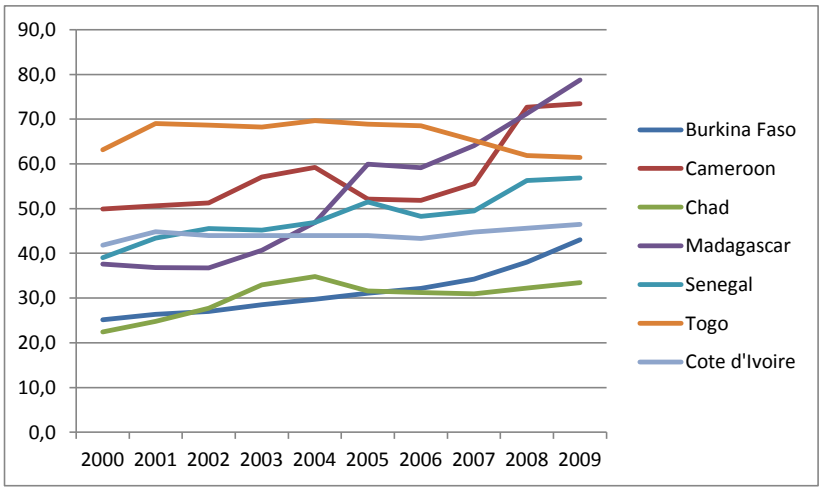
Nous avons également pris pour ces pays le taux de croissance de 2012 sachant qu'il varie légèrement selon les années.

1-4) Note sur la qualité des données sur l'éducation

L'utilisation de modèles de régression implique d'avoir suffisamment d'observations. Si pour un pays une variable est manquante, ce pays ne sera pas dans les observations. Etant donné qu'on se focalise sur les pays à faibles et moyens revenus, ils n'ont pas toujours des systèmes d'informations statistiques (en anglais EMIS *Education Management Information System*) et ne fournissent pas toujours des informations aux agences internationales telles que l'UNESCO ou la Banque Mondiale.

Certaines données fournies sont aussi peu fiables notamment lorsque l'on considère des tendances dans le temps. L'illustration en est fournie par le taux d'achèvement du cycle primaire pour quelques pays d'Afrique Francophone qui subit des variations importantes d'une année à l'autre.

Graphique2 : Taux d'achèvement pour quelques pays d'Afrique Francophone



Note: Les données manquantes ont été extrapolées
Source: Edstats accessed February 2012

On voit par exemple que la courbe du Cameroun suit une trajectoire pour le moins étrange. Certains pays s'abstiennent de fournir des données pour des raisons politiques, notamment lorsque la tendance n'est pas favorable.

Considérant que les pays répondent à des enquêtes internationales (comme celles mises en œuvre par l'Institut de Statistiques de l'UNESCO), la non-réponse peut être liée à

une volonté de masquer une situation. L'hypothèse est donc que les pays non répondants n'ont pas les mêmes caractéristiques que les pays répondants en termes de taux de scolarisation et de dépenses pour l'éducation

Il est donc crucial de faire le point sur la qualité des données fournies à travers les taux de réponse. Les données ont été collectées à travers le portail Edstats, l'outil StatSilk et complétées par des informations fournies par le Secrétariat GPE (profils pays en ligne).

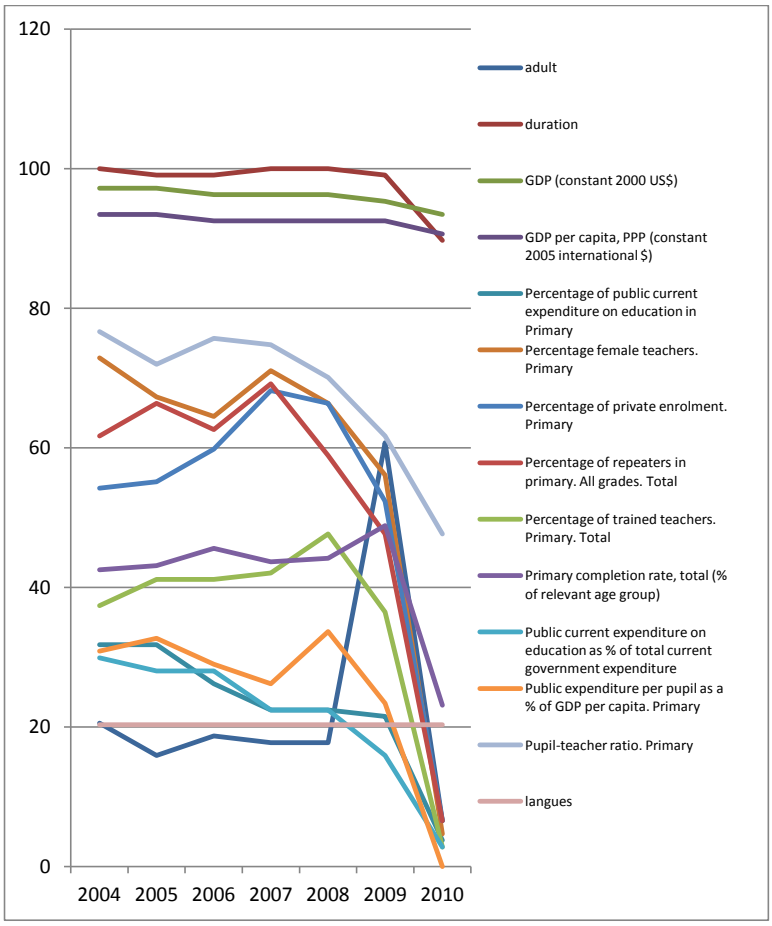
On considère dans les enquêtes sur l'éducation qu'en dessous de 80% de non réponse, on ne doit pas utiliser la variable, entre 80% et 95% tester différentes méthodes d'imputation et à plus de 95%, on peut utiliser des méthodes simples sans trop de risques d'erreurs.

On veut savoir quels sont les pays qui ne fournissent pas de données, ceux qui fournissent le plus de données, l'évolution dans la fourniture des données dans le temps et comparer la performance des pays GPE versus les pays non GPE. Dans le cadre de l'initiative GPE, certains pays ont reçu des appuis techniques et financiers sur la fourniture des données statistiques et l'on cherche à savoir si ces appuis ont été efficaces, i.e. il y a plus de données réellement.

On va donc calculer :

1-4-1) Les taux de réponse par variable, par année

Graphique3 : Taux de réponse par variable, par année



[5]

*En 2009 des estimations ont été produites sur la scolarisation des adultes.

L'année 2010 n'a pas été prise en compte car les données ne sont pas encore toutes disponibles auprès de l'UNESCO.

Les variables les plus documentées sont donc la durée du cycle primaire (duration), et le PIB (gdp). Mis à part ces variables, les autres ont un taux faible de données présentes qui a même tendance à se dégrader pour la plus part d'elles. Les variables les plus mal renseignées sont : la part accordée à l'éducation des dépenses publiques, les dépenses publiques sur chaque élève du primaire, la part accordée à l'éducation primaire des dépenses accordées à l'éducation, le niveau de scolarisation des adultes et le nombre d'élèves par enseignants. Ces différentes variables sont étroitement liées à la politique économique de l'Etat et refléteraient directement la performance de cette politique. La raison de l'absence de ces données seraient peut être politiques, et un manque de transparence. Nous avons supprimé la variable duration of primary (durée du primaire) parce qu'elle est constante pour plusieurs pays et ne mesure

[5] Fait par l'auteur à partir de Edstats UNESCO

pas vraiment la qualité de l'enseignement. Les variables PIB et part des dépenses faites dans le primaire par rapport aux dépenses dans l'éducation ont été également supprimées car elles manquaient de données et fournissent la même information que les variables PIB par personne et part de la dépenses dans l'éducation dans les dépenses de l'Etat. Enfin, la variable *trained teacher* (part des enseignants formés) a été également supprimée car elle est imprécise. En effet le niveau de formation n'est pas uniforme dans tous les pays. Nous allons voir les pays qui manquent d'information en calculant leur taux de réponse aux variables.

1-4-2) Les taux de réponse par pays

Le taux de réponse par pays que nous avons appelé *Taux_de_réponse* mesure le taux de réponse à toutes les variables de 2004 à 2009. Nous marquons dans un premier temps la présence ou l'absence d'au moins une information sur chaque variable de 2004 à 2009 (1 si il y a au moins une information et 0 sinon). Ensuite nous faisons la moyenne de toutes les variables pour obtenir le taux de réponse par pays.

Tableau3 : Taux de réponse

Obs	PAYS	Taux_de_r_ponse	Variables sans information
1	Kiribati	63,6 %	Expend, Literacy, private, Repet.
2	West Bank and Gaza	72,7%	Expend, Language, Corrupt.
3	Cuba	81,8%	Private, GDP.
4	Tajikistan	81,8%	Expend, Private.
5	Belize	81,8%	Literacy, Corrupt.
6	Fiji	81,8%	Literacy, Corrupt.
7	Iraq	81,8%	Expend, Private.
8	Uzbekistan	81,8%	Expend, Private.
9	Albania	90,9%	Expend.
10	Azerbaijan	90,9%	Private.
11	Jordan	90,9%	Female.
12	Kyrgyz Republic	90,9%	Expend.
13	Nepal	90,9%	Complet.
14	Sierra Leone	90,9%	Private
15	Angola	90,9%	Female.
16	Cape Verde	90,9%	Private.
17	Congo, Rep.	90,9%	Literacy.
18	Djibouti	90,9%	Literacy.
19	Egypt, Arab Rep.	90,9%	Expend.
20	Guyana	90,9%	Literacy.
21	Honduras	90,9%	Expend.
22	India	90,9%	Private.
23	Moldova	90,9%	Internet.
24	Pakistan	90,9%	Expend.
25	Sri Lanka	90,9%	Expend.
26	Sudan	90,9%	Expend.

27	Swaziland	90,9%	Private.
28	Timor-Leste	90,9%	Internet
29	Ukraine	90,9%	Expend
30	Yemen, Rep.	90,9%	Expend.

Fait par les auteurs

1-4-3) Les scores de réponse par pays

Nous avons calculé un taux de réponse par pays en considérant toutes les variables et toutes les années afin de calculer un Score pays. Ces scores sont classés par ordre croissant :

Tableau4 : Score de réponse

Num	PAYS	SCORE	Num	PAYS	SCORE	Num	PAYS	SCORE
1	Korea, Dem. Rep	8,8%	**34**	Uzbekistan	48,4%	**67**	Cuba	64,8%
2	Somalia	9,9%	**35**	Egypt	49,5%	**68**	Djibouti	64,8%
3	Zimbabwe	16,5%	**36**	Syrian Arab Republic	49,5%	**69**	Kyrgyzstan	64,8%
4	Tuvalu	19,8%	**37**	Vietnam	50,6%	**70**	Nicaragua	64,8%
5	Marshall Islands	22,0%	**38**	Côte d'Ivoire	51,7%	**71**	Mauritania	65,9%
6	Turkmenistan	24,2%	**39**	Sri Lanka	51,7%	**72**	Mongolia	65,9%
7	Micronesia	25,3%	**40**	West Bank and Gaza	51,7%	**73**	Senegal	65,9%
8	Solomon Islands	25,3%	**41**	Bhutan	52,8%	**74**	Bangladesh	67,0%
9	Angola	27,5%	**42**	Paraguay	52,8%	**75**	Central African	67,0%
10	Gabon	28,6%	**43**	Vanuatu	52,8%	**76**	Eritrea	67,0%
11	Iraq	31,9%	**44**	Ethiopia	53,9%	**77**	Maldives	67,0%
12	Tonga	35,2%	**45**	Low income	53,9%	**78**	Rwanda	68,1%
13	Armenia	35,2%	**46**	Ecuador	55,0%	**79**	Burundi	69,2%
14	Liberia	38,5%	**47**	Sudan	55,0%	**80**	Ghana	69,2%
15	Kiribati	39,6%	**48**	Swaziland	56,0%	**81**	Pakistan	69,2%
16	Timor-Leste	39,6%	**49**	Ukraine	56,0%	**82**	Benin	70,3%
17	Thailand	39,8%	**50**	Botswana	57,1%	**83**	Cambodia	70,3%
18	Malawi	40,7%	**51**	Philippines	58,2%	**84**	Uganda	70,3%
19	Myanmar (Burma)	40,7%	**52**	Guatemala	59,3%	**85**	Belize	71,4%
20	Samoa	40,7%	**53**	Latvia	59,3%	**86**	El Salvador	71,4%
21	Sierra Leone	40,7%	**54**	Kenya	60,4%	**87**	Lesotho	71,4%
22	Bolivia	41,8%	**55**	Nepal	60,4%	**88**	Guyana	72,5%
23	Dem. Rep. of Congo	41,8%	**56**	Azerbaijan	61,5%	**89**	Gambia	73,6%
24	Albania	42,9%	**57**	Indonesia	61,5%	**90**	Morocco	73,6%
25	India	42,9%	**58**	Lao	61,5%	**91**	Mozambique	73,6%
26	Nigeria	44,0%	**59**	Tunisia	61,5%	**92**	Cameroon	75,8%

27	Comoros	45,1%	**60**	Congo	62,6%	**93**	Madagascar	75,8%
28	Yemen	45,1%	**61**	Moldova	62,6%	**94**	Cape Verde	76,9%
29	Honduras	46,2%	**62**	Tanzania	62,6%	**95**	Mali	76,9%
30	Jordan	46,2%	**63**	Zambia	62,6%	**96**	Togo	76,9%
31	Fiji	47,3%	**64**	Chad	63,7%	**97**	Mauritius	78,0%
32	Georgia	48,4%	**65**	Guinea	63,7%	**98**	South Africa	78,0%
33	Tajikistan	48,4%	**66**	Namibia	63,7%	**99**	Burkina Faso	79,1%
						100	Niger	80,2%

Fait par les auteurs

Ce tableau nous présente des scores en données présentes pour chaque pays. Nous pouvons remarquer que la capacité à fournir des données ne dépend pas de la capacité économique. Le Niger par exemple est le pays ayant fournit le plus de données tandis que Gabon qui par contraste fait partie des pays les plus pauvres en données. Le Gabon n'est pas éligible aux crédits IDA (de la Banque Mondiale), il a donc moins d'incitations à fournir des données, alors que dans les pays recevant des financements, une condition (indicateur déclencheur) est la fourniture des données en temps voulu. Les pays IDA reçoivent davantage d'aide technique en matière statistique, de même que les pays GPE et fournissent donc plus d'informations. Les trois pays qui fournissent le moins de données (Corée du Nord, Somalie et Zimbabwe) souffrent de problèmes de gouvernance très important, qui suggère un lien de causalité entre fourniture des données et gouvernance.

1-4-4) Les scores moyens par tranche de revenu (comparaison de moyennes)

Tableau5 : score moyen par tranche de revenus

Pays à bas revenus	Pays à revenus moyens de la tranche inférieure	Pays de référence
0,58	0,53	0,54

Fait par les auteurs

Nous avons également calculé les scores moyens par tranches de revenus ; remarquons que les pays à plus faible revenu sont ceux qui produisent le plus de données, ce qui peut être contre-intuitif car les pays à plus grande capacité économique devrait avoir une plus grande capacité à fournir des données. Cela s'explique encore une fois par les appuis techniques reçus et les incitations à produire des statistiques.

1-1-5) Les scores moyens pour les pays GPE et non GPE (comparaison de moyennes)

Tableau6 : Score moyen

Pays GPE	Pays non-GPE
0,62	**0,50**

Fait par les auteurs

Tableau7 : Comparaison de moyenne

Méthode	Variances	DDL	Valeur du test t	Pr > |t|
Pooled	Equal	1389	-5.38	<.0001
Satterthwaite	Unequal	905.82	-5.60	<.0001

Fait par les auteurs(SAS)

Le premier tableau présente des scores moyens calculés pour la totalité des pays GPE et ceux qui ne le sont pas. Nous remarquons que les pays GPE fournissent plus de données que ceux qui ne le sont pas et cette différence est significative d'après le test de Student dont les résultats se trouvent dans le deuxième tableau. Cet avantage des pays GPE est surement dû au fait qu'ils doivent fournir des informations sur la gestion des investissements qu'ils reçoivent en vue d'encourager les bailleurs de fonds.

Il est tout à fait notable la bonne performance du Niger. En effet, en 2006, des scandales financiers ont éclaté dans le secteur de l'éducation, et la Banque Mondiale a identifié des détournements d'argent dans le secteur de l'éducation. Par la suite, les bailleurs de fond ont demandé au Niger la production régulière d'informations financières. Il n'y a donc pas de relation univoque entre fourniture d'informations sur l'éducation et gouvernance. Si la fourniture de données est sans doute un indicateur de transparence et de bonne gouvernance, d'autres facteurs entrent en jeu telles que les demandes des bailleurs de fond, le montant de l'aide reçue, la participation à certaines initiatives telles que GPE.

Conclusion

Les pays qui ne donnent pas des informations n'ont pas les mêmes caractéristiques que les pays qui les pays qui en fournissent. A priori on pourrait s'attendre à ce que les pays pauvres soit les moins fournisseurs d'information justement par manque de moyens financier et de ressources personnelles et techniques qualifiées, mais tel n'est pas le cas. Il est donc possible que la carence en information soit liée à une stratégie politique.

Nous avons également examiné un indicateur de la corruption, qui est sensiblement basé sur le même principe que le taux de réponse étudié ci-dessus, c'est une mesure de la transparence.

Il est calculé à partir de différentes sources d'information sur la gouvernance et les risques financiers dans un pays, complétées par des études (assez rares) sur la perception de la corruption, ou d'enquêtes spécifiques à l'éducation. D'autres enquêtes telles que le *Public Expenditure Tracking Survey* mesurent la part de l'argent destiné aux écoles qui est perdu, mais cet indicateur n'est pas disponible pour beaucoup de pays.

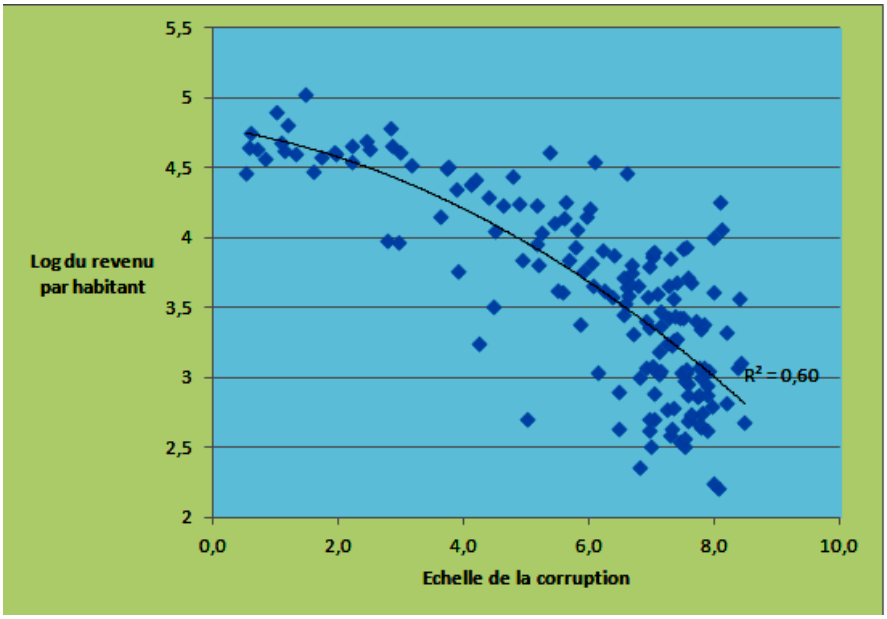
Les pays tel que la Somalie, la Koreé, le Myanmar, le Yemen, le Zimbabwe, le Turkmenistan, l'Angola et l'Afghanistan se trouvent être les pays les plus corrompus sur base de ce indicateur et sont en même temps les moins bons fournisseurs d'information sur l'éducation. L'indicateur de Transparency International possède donc une forte de validité

corrélationnelle, il mesure sensiblement la même chose que la fourniture des données sur l'éducation, mais élargi à tous les domaines de la vie économique et politique.

Par la suite, on procèdera à des imputations mais devra donc intégrer le score sur la fourniture des données et l'indice de Transparency International comme des variables de gouvernance qui peuvent être liés à la qualité de l'éducation. Le graphique ci-dessous montre que l'indice de Transparency est aussi lié au revenu par habitant.

On voit déjà apparaître des corrélations entre différents indicateurs tels que le taux de réponse, l'indice de corruption, le revenu par habitant et la participation au Partenariat Mondial pour l'Education dont nous devons tenir compte au moment de la modélisation. Cela justifie l'utilisation de modèles adaptés aux cas de multicolinéarités.

Graphique4 : Log du revenu par habitant en fonction de l'échelle de corruption



Source : Varly (2012), La corruption dans l'éducation, blog sur l'éducation dans les pays du Sud.

Les variables qui seront introduites dans les modèles sont très corrélées entre elles, et il faudra donc prendre soin de contrôler les multicolinéarités dans les modèles, voir à créer plusieurs modèles alternatifs.

On notera que quelques pays ne fournissent quasiment aucune donnée et sont exclues des analyses :

Afghanistan, Haiti, Sao Tome & Principe, Guinée Bissau, Papouasie Nouvelle Guinée.

Ces pays ne fournissent pas non plus de données sur les acquis scolaires.

La base constituée contient des pays dit *benchmark* soit qui viennent de passer à la tranche de revenu supérieur (*Upper middle income*), soit leurs systèmes éducatifs ont des caractéristiques intéressantes ou font figure de modèles (Cuba) de par leur réussite dans les enquêtes internationales.

1-5) Méthodes d'imputation des données manquantes

Face aux données manquantes évoquées plus haut, nous avons exploré plusieurs méthodes d'imputation des données :

1-5-1) Moyenne et voisin le plus proche

Ces deux méthodes considèrent une ressemblance entre les observations selon plusieurs variables ou selon des groupes donnés. En effet, la méthode d'imputation par la moyenne propose d'imputer la valeur de la moyenne d'un groupe donné aux observations manquantes. On suppose donc que la valeur manquante est l'observation moyenne des autres. Cette méthode réduit la variance qui devait normalement observée si l'on avait toutes les observations. Dans notre cas, nous pouvons faire une imputation selon les différentes tranches de revenus.

La méthode du plus proche voisin quant à elle consiste à affecter la valeur du voisin le plus proche, c'est-à-dire celui qui a les plus proches valeurs pour toutes la variables observées, à la valeur manquante. Cette méthode suppose que les pays qui se ressemblent pour toutes les variables, se ressembleraient aussi pour les valeurs manquantes. Elle n'est donc pas très appropriée ici, car le but même de notre étude est de voir l'impact des variables dans chaque pays, ce qui nous permettra de former un modèle fiable. De plus, il n'est pas évident de trouver un voisin vraiment proche lorsqu'il y a beaucoup de variables. Elle réduit aussi la variance observée sans valeur manquante.

1-5-2) Méthode d'imputation par la régression

Elle consiste à détecter des variables qui sont potentiellement liées à la variable qui possède des valeurs manquantes, faire une régression en tenant compte des valeurs observées. Après l'obtention des coefficients de régression, on peut donc prévoir les valeurs manquantes. Cette méthode requiert d'avoir des données monotones. C'est-à-dire que lorsque pour un pays donné, une variable n'est pas fournie, alors les autres variables qui suivent (dans l'ordre de régression) aussi doivent manquer de données. Notre base ne vérifie pas cette propriété, et nous ne pouvons donc appliquer cette méthode.

1-5-3) Méthode d'imputation MIMIC

Cette méthode est utilisée pour générer des nombres pseudo aléatoires. Premièrement, elle simule des valeurs pour chaque vide dans la base grâce à une estimation de la moyenne et de la matrice de covariance. Ensuite, elle simule la moyenne et la matrice de covariance à postériori grâce à la base complète estimée. Avec ces nouvelles valeurs elle reprend l'étape une et ainsi de suite. Il est préférable pour une meilleure qualité d'imputation de ne prendre dans le modèle d'imputation que des variables qui sont liées à la variable qui manque de

valeurs. Le nombre d'observation (90 pays) ne permet pas d'assurer la validité de cette méthode.

Nous avons imputé les valeurs manquantes par la moyenne par groupe selon la classification en tranches de revenus, ce qui a l'avantage de ne pas créer artificiellement de variance.

1-6) Base de données finale

Comme souligné plus haut, les indicateurs de l'éducation sont un peu rares pour un grand nombre de pays. Ces indicateurs ne varient pas souvent d'une année à l'autre. Par exemple les taux d'alphabétisation sont mesurés à travers les recensements, les enquêtes ménages ou d'enquêtes spécifiques, comme le projet LAMP de l'UNESCO. Les données sont généralement déclaratives (Savez-vous lire et écrire ?). Les taux d'alphabétisation ne varient pas beaucoup d'une année à l'autre et on a peu de données en tendance. En 2009, davantage de données sont fournies, correspondant à des estimations de l'Institut de Statistiques de l'UNESCO. Le taux d'alphabétisation est à la fois une cause et une conséquence de taux de scolarisation élevés et potentiellement d'une meilleure qualité de l'éducation. Au niveau micro, les ménages alphabétisés ont des enfants qui réussissent mieux que les autres. Aussi, la pratique du redoublement est très dépendant des systèmes éducatifs, Bernard (2005) montre que les pays francophones et lusophones font plus redoubler que les pays anglophones en Afrique. Cela tient à un certain « héritage » colonial dans les pratiques pédagogiques, la France et le Portugal ont des taux de redoublement plus élevés qu'en Angleterre. Il en est de même pour le nombre d'élèves par enseignants, la proportion des femmes enseignantes…

Nous avons donc fait une moyenne de 2004 à 2009 pour toutes les variables. Lorsque l'on veut étudier la relation entre économie et éducation (ou capital humain), il existe plusieurs types de modèle, des modèles dits de flux (ou on raisonne en termes de taux d'accumulation du capital) soit des modèles de stock (ou on raisonne en coupes). Pour schématiser, pour certains auteurs c'est la croissance économique qui incite à investir dans l'éducation, pour d'autres c'est le niveau de revenu par habitant qui détermine la capacité à investir dans l'éducation. Enfin, en termes de causalité, on ne sait pas si ce sont les taux de scolarisation qui déterminent les revenus ou la croissance ou si c'est l'inverse. Néanmoins, Doudjidingao (2011) nous indique que le développement économique est un préalable à des taux de scolarisation et préfère utiliser des modèles de stock.

Ici, nous avons des problèmes de données qui ne sont pas disponibles pour toutes les années et nous pouvons difficilement utiliser des modèles de flux. La littérature semble indiquer que nous pouvons nous contenter de modèles de stock. Néanmoins, nous allons tenter de modéliser à la fois le niveau de scolarisation mais également la croissance des taux de scolarisation dans le temps. De plus, pour certaines variables telles que la démographie, nous allons inclure dans le modèle à la fois des variables de stock et des variables de flux, mais gardons à l'esprit que nous utilisons davantage les variables selon les données disponibles plutôt que sur la base de théorie de flux et de stock. Nos modèles allient donc variables de stock et variables de flux dans la mesure du possible.

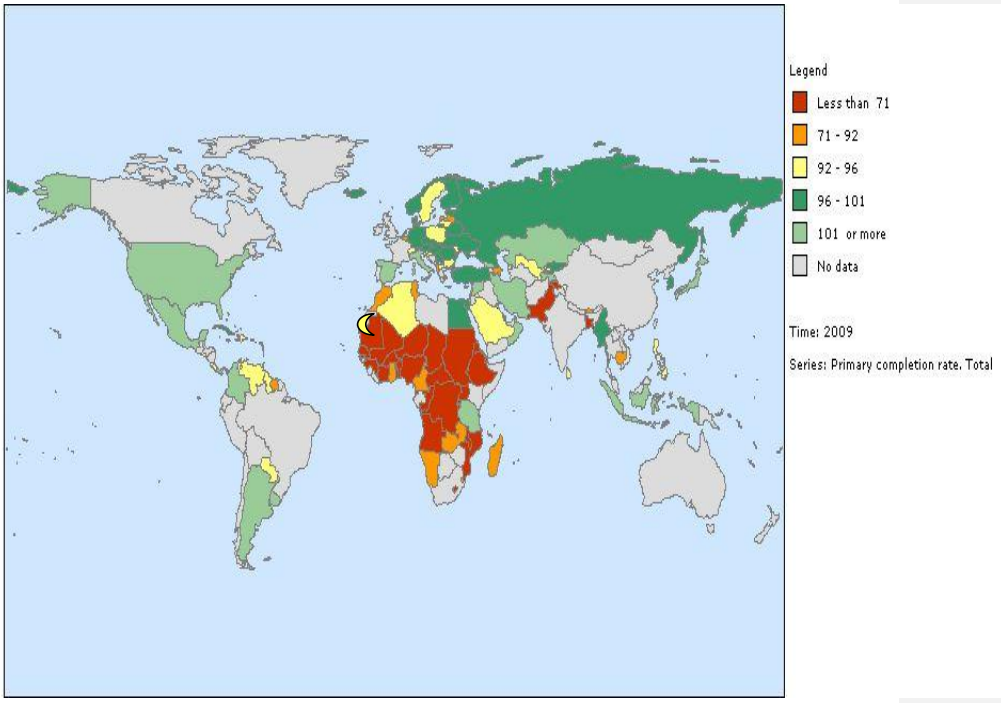
II- Etude Préliminaire

2-1) Etude descriptive

Le but de notre étude est de voir à l'aide des données macro, les facteurs qui déterminent une bonne qualité de l'éducation. Nous cherchons donc à identifier d'une part les variables qui ont le plus d'effet sur les résultats scolaires et d'autres part les pays performants en éducation et leurs politiques éducatives.

Pour un premier aperçu, cette carte nous présente le taux d'achèvement de tout les pays en 2009 :

Graphique5 : Répartition du taux d'achèvement (carte)



Source : Edstats

Nous pouvons constater comme souligné plus haut que ce sont les pays d'Afrique qui fournissent le plus de données, contrairement aux pays de l'Amérique Latine et de l'Europe. Aussi les pays de l'Afrique de l'ouest et d'Afrique Centrale, pour la plus part francophones, sont dans la plus basse catégorie du taux d'achèvement, c'est-à-dire moins de 71%. Les pays de l'Afrique anglophone sont mieux classés. Les meilleurs pays en éducation en Afrique sont les pays arabes.

Nous avons effectués ici quelques analyses descriptives pour appréhender les données :

Tableau8 : Statistiques descriptives des variables

Variable	Observations	Obs. avec données manquantes	Minimum	Maximum	Moyenne	Ecart-type
Complet	90	0	31,8	100	77,6	22,1
Primary	90	43	24,4	61	42,5	8,2
Literacy	90	0	26,1	99,83	73,9	20,5
Internet	90	0	0,2	53,56	7,2	7,8
Loggdp	90	0	5,66	9,54	7,8	0,9
popgrowt	90	0	-1,01	4,4	1,8	1,0
Female	90	0	12,1	99,2	55,7	22,6
Logpop	90	0	4,6	14	8,7	1,8
Ptr	90	0	8,6	90,3	34,6	15,6
Repet	90	0	0,6	31,06	9,37	7,7
expend	90	0	2,9	40,7	14,23	7,7
dur_gpe	90	0	0	10	2,93	3,6
Corrupt	90	0	3,9	9,5	7,08	0,9
language	90	0	0	21,3	0,81	2,2
response	90	0	0,3	0,8	0,58	0,1
Private	90	0	0,5	98,9	14,1	18,3

Fait par les auteurs(XLSTAT)

La qualité de l'éducation dépend également des systèmes éducatifs, que l'on a classés en groupes selon l'appartenance géographique et les modes d'organisation politiques. Le tableau ci-dessous nous montre les moyennes et écarts types des différentes variables selon différents groupes de systèmes éducatifs.

Tableau9 : Moyennes par systèmes éducatifs

Système	Complet	Primary	literacy	internet	Loggdp	popgrowt	Corrupt	langues	logpop	dur_gpe	Expend	Female	ptr	repet	response	private
ANGLO	72,6	35,9	70,5	4 ,75	7,5	2,3	6,6	0,7	9,2	3,2	13,5	50,6	42,6	9,6	0,6	9,8
	(15,3)	(4,5)	(17,9)	(4,5)	(1,07)	(1,02)	(1,02)	(0,24)	(1,5)	(3,45)	(5,8)	(17,7)	(15,9)	(5,7)	(0,13)	(8,7)
ARAB	81,8	40,1	74,9	11,4	8,2	2,1	7,4	0,6	9,9	1,0	15,5	57,2	24,3	5,9	0,5	9,3
	(17,4)	(7,5)	(13,4)	(7,8)	(1,06)	(0,86)	(1,2)	(0,27)	(0,7)	(3)	(2,6)	(14,5)	(5,5)	(4,4)	(0,12)	(9,2)
ASIA	83,8	47,4	74,6	6,6	7,8	1,5	7,1	0,7	10,4	2,7	13,6	56,3	32,6	7,8	0,6	12,0
	(16,2)	(6,6)	(18,5)	(5,9)	(0,6)	(0,5)	(0,9)	(0,24)	(2,3)	(3,2)	(5,8)	(21)	(10)	(6,7)	(0,1)	(12,4)
FRANC_LUS	53,3	38,7	57,7	2,5	7,2	2,6	7,3	0,6	8,8	4,3	13,4	35,4	47,6	17,6	0,7	18,2
	(15)	(5,2)	(18,5)	(2,8)	(0,8)	(0,6)	(0,9)	(0,3)	(1,2)	(3,9)	(8,2)	(15,8)	(14,1)	(7,2)	(0,14)	(18,1)
ISLE	95,8		88,9	7,5	8,2	1,2	6,9	0,4	5,4	0,0	15,5	65,6	23,6	5,6	0,5	27,9
	(7,1)		(10,7)	(4,7)	(0,25)	(0,8)	(0,5)	(0,4)	(0,8)	0	(4,4)	(9,3)	(4,8)	(4,3)	(0,12)	(35,8)
LATIN	91,0	45,6	82,9	10,1	8,3	1,5	7,3	0,3	8,3	3,3	10,4	72,1	28,3	6,2	0,6	20,8
	(11,3)	(1,3)	(6,4)	(5,8)	(0,37)	(0,8)	(0,33)	(0,3)	(1,4)	(5)	(3,6)	(7,3)	(5,5)	(4,2)	(0,11)	(25,9)
SOVIET	95,5	54,1	99,2	15,6	8,4	0,3	7,1	0,6	8,9	2,8	19,2	87,8	16,7	1,2	0,5	6,5
	(3,2)	(6,7)	(1,13)	(13,6)	(0,68)	(0,8)	(0,9)	(0,3)	(1)	(3,28)	(10,9)	(11)	(5,4)	(0,6)	(0,1)	(6,2)
Moyenne	82,0	43,6	78,4	8,4	7,9	1,6	7,1	0,8	8,7	2,5	14,4	60,7	30,8	7,7	0,6	14,9

Fait par les auteurs

Anglo désigne les pays de l'Afrique anglophone, *arab* :les pays arabes, *asia* :les pays de l'Asie, *Franc_lus* : l'Afrique francophone et lusophone, *isle* les îles. Enfin, *Latin* et *soviet* désignent respectivement les pays latins et les pays de l'ex bloc soviétique.

Cette classification pour l'Afrique est différente de celle utilisée par Doudjidingao qui a tendance à mettre certains pays lusphones avec les pays anglophones. De même, Djibouti ou le Soudan sont parfois considérés comme des pays arabes dans les classifications de l'Unesco.

Les pays de l'Afrique francophone et lusophone ont en moyenne le plus bas taux d'achèvement ainsi que les scores indiquant la qualité de l'éducation. Pour la plus part, ils n'ont de bons résultats ni pour le taux d'achèvement ni pour les scores obtenus lors des différents tests. Les exceptions sont essentiellement le Gabon, le Cape vert, le Congo et les Comores. Ensuite il y a les pays de l'Afrique anglophone qui ont des résultats acceptable mais avec une forte variabilité. En effet, l'Erithrée, l'Ethiopie, le Malawi, et l'Ouganda n'ont pas de bons résultats. Les pays arabes quant à eux ont également de bons résultats et présente une forte variabilité. Cela est dû au Soudan et au Yémen qui ont des résultats carrément bas par rapport aux autres pays arabes. Nous remarquons aussi que les pays de l'Europe et de l'Asie centrale en plus de Cuba (SOVIET) ont de très bons résultats avec une variance faible, qui nous montre l'homogénéité de leurs résultats scolaires.

Pour ce qui est de la variable literacy (le taux d'alphabétisation des adultes), les pays du groupe de L'Afrique francophone et Lusophone (FRANC_LUS) sont toujours les moins performants, avec en moyenne un taux de 57,74% mais qui présente une grande variabilité. Il y a certains pays qui ont un taux assez bas d'alphabétisation des adultes comme le Burkina Faso et le Mali (26%), le Niger (28%) et la Guinée (39%).

Les pays de l'Asie (ASIA) ont en moyenne un taux assez élevé d'alphabétisation des adultes, mais avec une très forte variabilité. En effet il y a l'Indonésie, Sri Lanka, la Mongolie, les Phillipines et la Thaïlande qui ont taux d'alphabétisation supérieur à 90%, et d'autre part Timor-Leste, Pakistan, Népal, Bhutan et Bangladesh qui ont un taux de 50% environ.

Cet effet est également observé chez les pays de l'Afrique anglophone (ANGLO) et arabes (ARAB). Parmi les pays anglophone on a par exemple la Sierra Leone qui a un taux d'alphabétisation des adultes de 37,87% et l'Ethiopie qui a un taux de 32,86%. Concernant les pays arabes, on a le Yémen et le Maroc qui ont respectivement 58,6% et 54,51%, les taux les plus bas de ce groupe. Le Soudan a quant à lui un taux d'alphabétisation de 70,2%. Les pays soviet présentent un très fort taux d'alphabétisation (99,15%) des adultes avec une variabilité très faible (1,13) ce qui prouve qu'ils ont atteint la scolarisation universelle il y a déjà quelques années. Ils ont également le plus grand nombre de femmes enseignantes. Les pays latins ont en moyenne plus de femmes enseignantes (72,13%) que les îles (65,56%). A part cette exception, les variables *Literacy* et *female* sont réparties de la même manière dans les groupes.

Le nombre d'internautes est faible dans les pays de l'Afrique francophone et Lusophone et meilleure dans les pays soviet, notamment Latvia (53,56) et Azerbaijan (21,5).

Pour ce qui est du revenu par habitant, les pays que nous qualifierons de plus pauvres sont premièrement ceux de l'Afrique francophone et Lusophone, ensuite ceux de l'Afrique anglophone suivi des pays de l'Asie. Ces résultats sont généraux car les variances sont très faibles. De plus par construction, nous nous sommes restreints aux pays à faible ou moyen revenus.

Nous remarquons que les pays de l'Afrique en général connaissent une grande croissance démographique et sont également plus peuplés. Les pays les moins peuplés sont ceux de l'ex block soviétique, les îles (par définition), les pays d'Amérique Latine et ceux de l'Asie. Ils ne connaissent pas non plus de forte croissance démographique.

Les pays les moins corrompus sont les îles et les pays de l'Afrique anglophone.

Quant à la diversité des langues, nous le remarquons surtout dans les pays africains et plus chez les anglophones.

Les pays participants au partenariat mondial pour l'éducation (nommés fastrack) sont en général ceux de l'Afrique francophone et Lusophone, il y a également les pays latins, ceux de l'Afrique anglophone et de l'ex block sovétique. Les îles ne participent pas à ce programme.

Pour ce qui est des dépenses dans l'éducation, les premiers sont les pays de l'ex block soviétique, mis à part Azerbaijan qui place seulement dans l'éducation 5,61% de ses dépenses dans l'éducation. Il y a ensuite les pays arabes et les îles. Les pays d'Amérique Latine sont ceux qui dépensent le moins dans l'éducation, par rapport à leur dépense publique.

Les variable ptr (nombre d'élèves par enseignants) et repet (taux de répétition) se comportent de la même manière pour tous les groupes. Les pays de l'Afrique ont en général un taux de répétition et un nombre d'élèves par enseignant élevé, sauf le Cap Vert qui a 25 élèves par enseignant contrairement à la Centrafrique qui présente en moyenne 90 élèves par enseignants, un ratio peu compatible avec une bonne qualité de l'éducation.

En ce qui concerne le taux de réponse, nous remarquons que les pays qui ne fournissent pas de données sont pour la plus part les pays de l'ex block soviétique, les pays arabes et les îles. Mais cette variable a une variance élevée et varie au sein de chaque groupe.

Enfin, nous remarquons qu'il y a une forte dominance du privé dans les pays latins, et les pays francophones et lusophones. Ces derniers présentent une forte variabilité car contrairement au Mozambique (1,8%), le Rwanda (1,7%), le Burundi (1,2%) et le Niger (3,9%) qui une faible scolarisation dans le privé, le Congo Démocratique (82,51%), le Gabon (29%)et le Mali (38%).Il y a une forte présence du privé à Fiji (98,95%) mais pour les autres îles, le taux le plus élevé est de 26%.

2-2) Etude empirique des corrélations entre variables explicatives

Les variables explicatives de par leur définition risquent d'être fortement corrélées entre elles. Nous ne pensons pas qu'il y ait de colinéarité exacte, mais empiriquement nous pouvons nous attendre à une corrélation positive par exemple entre le taux d'alphabétisation des adultes et le taux de femmes enseignantes. En effet, plus les adultes sont alphabétisés plus il y a une forte probabilité que les femmes le soient et donc nous aurons plus de femmes enseignantes. Le taux d'alphabétisation peut être également corrélé négativement avec le taux de répétition car facilitant une bonne réussite des élèves. La variable internet peut être liée au PIB par habitant car c'est aussi une mesure du développement du pays. Elle peut être également liée au taux de répétition car l'accès à internet peut faciliter la formation des élèves. Le nombre d'élèves par enseignants sera fortement lié au taux de répétition car plus les élèves redoublent, plus ils rempliront les salles de classes. Le taux de répétition peut être lié à l'indice de corruption, car plus le pays est corrompu, plus il y aura des facilités à faire passer des élèves non pas à cause de leur acquis scolaire mais plutôt de l'influence de leurs parents. La dépense dans l'éducation peut être liée à la prédominance du privé dans le pays. En effet, plus il y a d'écoles privées, moins l'Etat sera obligé de construire des écoles publiques et donc la dépense dans l'éducation sera moindre. Notons toutefois que l'Etat subventionne également certaines écoles privées.

La dépense dans l'éducation peut être également liée au nombre d'élèves par enseignants. Cette corrélation est négative car plus l'Etat réduit le nombre d'élèves par classe, plus il lui faudra embaucher des enseignants et créer des salles de classes, il augmente ainsi sa dépense dans l'éducation. Le taux de réponse quant à lui peut être négativement lié à la corruption dans le pays car les pays mal gouvernés ont tendance à être moins transparents sur les performances du secteur public. Dans la constitution de notre base de données nous avons supprimé les pays qui n'avaient pas assez de données comme l'Afghanistan et Haiti ; ces pays ont également un fort taux de corruption. Le taux de réponse est positivement corrélé à la durée de participation à l'initiative fastrack car les pays qui y participent sont tenus de fournir les informations sur l'évolution de leurs résultats scolaires. A partir de ce canevas théorique, nous avons identifié certaines relations entre les variables puis les tester empiriquement à travers l'examen de la matrice des corrélations.

3- Analyse de la matrice de corrélation :

Tableau10 : Matrice de corrélation

Variables	literacy	internet	loggdp	popgrowt	Corrupt	language	logpop	dur_gpe	expend	female	ptr	repet	response	private	Complet	Primary
literacy	**1,00**															
internet	0,46	**1,00**														
loggdp	0,63	0,59	**1,00**													
popgrowt	-0,64	-0,53	-0,61	**1,00**												
Corrupt	-0,16	-0,24	-0,37	0,25	**1,00**											
language	0,09	-0,02	-0,02	-0,25	0,01	**1,00**										
logpop	-0,24	-0,10	-0,14	0,15	0,22	-0,09	**1,00**									
dur_gpe	-0,37	-0,21	-0,44	0,16	0,10	0,12	-0,01	**1,00**								
expend	0,11	0,23	0,03	-0,26	-0,10	0,32	-0,12	0,10	**1,00**							
female	0,81	0,53	0,63	-0,67	-0,17	0,15	-0,13	-0,23	0,17	**1,00**						
ptr	-0,61	-0,55	-0,64	0,54	0,19	-0,09	0,22	0,26	-0,30	-0,61	**1,00**					
repet	-0,42	-0,43	-0,41	0,47	0,17	-0,11	-0,02	-0,02	-0,15	-0,54	0,58	**1,00**				
response	-0,28	-0,01	-0,22	0,04	-0,14	0,04	0,11	0,27	0,10	-0,17	0,25	0,15	**1,00**			
private	-0,12	-0,09	-0,05	0,13	0,05	-0,06	-0,11	-0,19	-0,15	-0,22	-0,03	0,09	0,06	**1,00**		
Complet	0,76	0,55	0,69	-0,64	-0,24	0,08	-0,19	-0,32	0,08	0,73	-0,71	-0,64	-0,27	-0,01	**1,00**	
Primary	0,57	0,55	0,45	-0,80	0,05	0,36	-0,01	-0,27	0,31	0,63	-0,54	-0,34	-0,24	-0,08	0,49	**1,00**

Fait par les auteurs (XLSTAT)

La matrice de corrélation confirme la présence de fortes corrélations entre variables explicatives, ce qui nous incite à prendre en compte les problèmes de multicolinéarités dans les modèles. Nous pouvons également identifier les variables qui sont les plus corrélées avec le taux d'achèvement et les scores des évaluations, il s'agit du taux d'alphabétisation des adultes, de la variable internet, du PIB par habitant, de taux de femmes enseignantes, du nombre d'élèves par enseignants, du taux de redoublement, de la croissance de la population et de la variable mesurant la durée de participation d'aide à l'éducation. La diversité des langues est corrélée avec les scores mesurant la qualité de l'éducation, ce qui est logique, plus il y a de langues dans un pays, plus les élèves devront fournir le double effort d'apprendre une nouvelle langue et de se former. La dépense dans l'éducation n'est pas fortement corrélée avec la qualité de l'éducation contrairement aux résultats de Lee et Barro (2001) ; ils ont démontrés que les ressources de l'école ont un effet positif sur la qualité, mais cette relation n'est malheureusement pas prouvée par l'ensemble des études.

III- Etude Approfondie: Analyse en Composantes Principales

A partir de la matrice de corrélation observée plus haut, nous pouvons nous attendre à avoir une bonne factorisation des variables. Le tableau suivant nous montre les différents axes extraits et le pourcentage de la variance expliquée :

3-1) Résultats de l'ACP

Tableau11 : Inertie expliquée des axes de l'ACP

Total Variance Explained						
	Initial Eigenvalues			Extraction Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	5,576	37,174	37,174	5,576	37,174	37,174
2	1,696	11,308	48,482	1,696	11,308	48,482
3	1,252	8,349	56,831			
4	1,156	7,709	64,54			
5	0,968	6,453	70,994			
6	0,873	5,823	76,817			
7	0,778	5,184	82			
8	0,691	4,608	86,608			
9	0,472	3,147	89,756			
10	0,373	2,488	92,244			
11	0,32	2,135	94,379			
12	0,292	1,944	96,323			
13	0,215	1,433	97,756			
14	0,194	1,29	99,046			
15	0,143	0,954	100			
Extraction Method: Principal Component Analysis.						

Fait par les auteurs (XLSTAT)

Nous voyons que le premier axe explique 37,17% de la variance et le deuxième axe en explique 11,3%. La projection dans le premier plan factoriel explique donc environ 50% de la variance. Le tableau ci-dessous présente les coordonnées des variables dans les deux axes :

Tableau12 : Coordonnées des variables

Coordonnées des variables	1er axe	2ème axe
Complet	0,888	-0,134

literacy	0,844	-0,134
internet	0,702	0,162
gdp	0,822	-0,177
popgrowt	-0,787	-0,208
Corrupt	-0,326	-0,217
pop	-0,247	-0,045
ptr	-0,818	0,014
repet	-0,66	-0,157
female	0,855	0,069
dur_gpe	-0,365	0,628
expend	0,237	0,657
response	-0,263	0,451
language	0,335	0,583
private	-0,120	-0,377

Fait par les auteurs (XLSTAT)

3-2) Analyse du premier axe

Les variables : complet, literacy, female, internet et gdp ont de grandes valeurs positives sur la première composante, comme le montre le tableau suivant :

Tableau13 : variables corrélées positivement au premier axe de l'ACP

Variables	Component1
Complet	0,888
female	0,855
literacy	0,844
gdp	0,819
internet	0,702

Fait par les auteurs

Les variables complet ou taux d'achèvement, literacy qui est le taux d'alphabétisation des adultes et female qui la proportion de femmes enseignantes sont des indicateurs de la qualité de l'éducation et les variables internet, qui mesure l'accès à l'information des populations, et gdp, qui mesure le revenu moyen des parents, sont des facteurs qui facilitent l'éducation.

Le premier axe oppose ces variables aux variables ptr, repet et popgrowth qui ont quant à elles de grandes valeurs négatives comme représenté dans le tableau suivant :

Tableau14 : variables corrélées négativement au premier axe de l'ACP

Variables	Component1
ptr	-0,818
popgrowt	-0,787
repet	-0,66

Fait par les auteurs

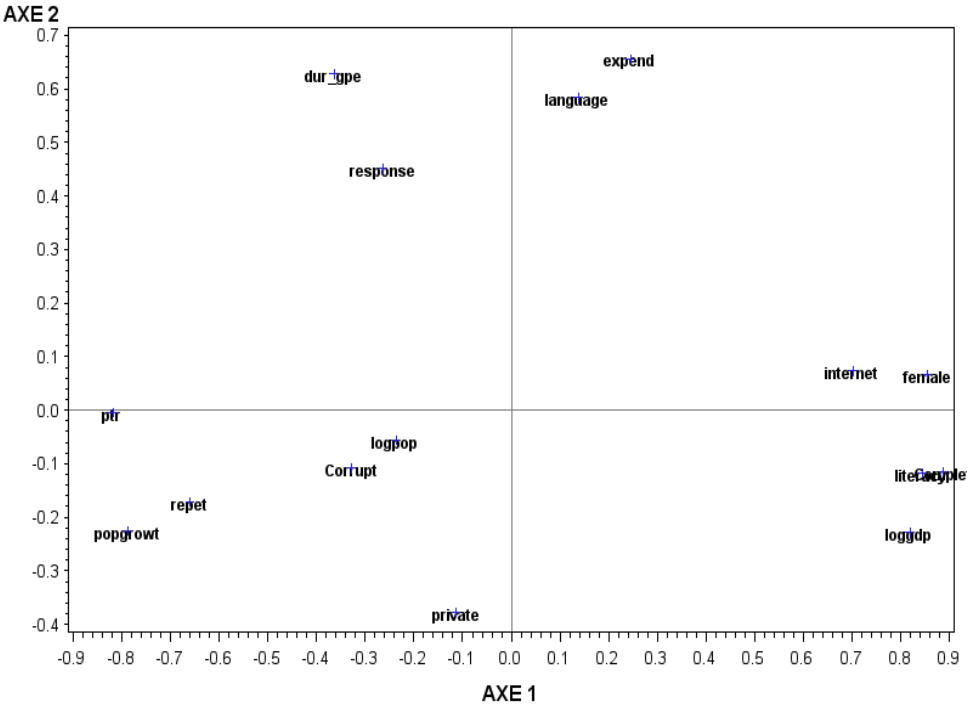
Or, ces variables : ptr (nombre d'élèves par enseignant) et repet, le taux de répétition, sont des indicateurs d'une mauvaise qualité de l'éducation. De plus, la variable popgrowth (croissance de la population) est un handicap à une bonne qualité de l'éducation.

Le premier axe peut donc être considéré comme une mesure la qualité de l'éducation, car il possède, comme nous l'avons montré plus haut, de grandes valeurs pour les variables indiquant une bonne qualité de l'éducation et de grandes valeurs négatives pour les variables indiquant une mauvaise qualité de l'éducation. Comme expliqué dans la présentation de la méthode[6], les variables dont les supports forment un petit angle entre eux et qui vont dans le même sens sont corrélées positivement, celles qui vont dans le sens contraire sont corrélées négativement et celles qui forment un angle de 90° ne sont pas corrélées.

Les variables dont les supports forment un angle de 90° avec le premier axe sont celles qui n'expliquent pas vraiment la qualité de l'éducation. Le graphique suivant nous montre la représentation dans le premier plan factoriel des variables :

[6] Voir page : présentation de la méthode.

Graphique6 : Représentation des variables dans le premier plan



Fait par les auteurs

En plus des variables identifiées dans la littérature comme indiquant une bonne qualité de l'éducation, à savoir le taux d'alphabétisation des parents (literacy), le taux de femmes enseignantes et le revenu par habitant, nous remarquons que le taux d'internautes par pays est également un bon indicateur de la qualité de l'éducation. Il en est de même pour la corruption et le taux de croissance démographique de la population qui sont des indicateurs d'une mauvaise qualité de l'éducation et qui n'ont pas été pris en compte dans la littérature.

3-3) **Analyse du deuxième axe**

Le tableau suivant nous montre les variables qui ont de grandes valeurs positives pour le deuxième axe :

Tableau15 : Variables corrélées positivement au deuxième axe de l'ACP

Variables	Component2
expend	0,657
dur_gpe	0,628
language	0,583
response	0,451

Fait par les auteurs

La variable qui a la plus grande valeur est expend ; c'est la dépense de l'Etat dans l'éducation. Ensuite nous avons les variables : dur_gpe qui est la durée de participation à l'initiative d'aide à l'éducation, language qui mesure la diversité linguistique des pays et response qui est le taux de réponse des pays. Ces variables sont également des indicateurs des dépenses des pays[7].

Ces variables sont opposées à la variable private qui est une mesure de la dominance du privé, et qui est un facteur de réduction des dépenses de l'Etat[8]à moins qu'il ne subventionne ces écoles :

Tableau16 : variables corrélées négativement au deuxième axe de l'ACP

Variables	Component2
private	-0,377

Fait par les auteurs

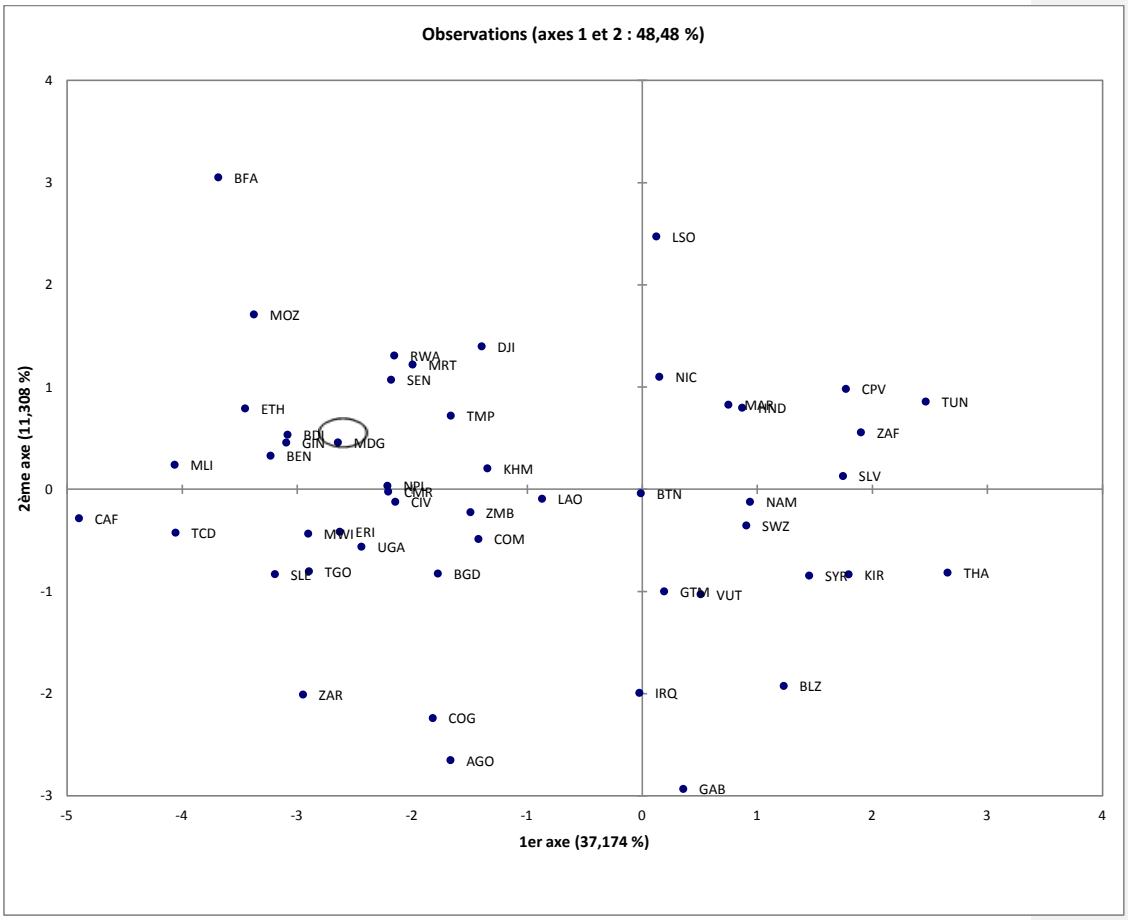
Nous pouvons donc conclure que le deuxième axe mesure le niveau de dépense et de gouvernance. Ces variables ne sont pas corrélées au 1er axe et donc ne sont pas extrêmement importantes pour déterminer la qualité de l'éducation.

3-4) Analyse individus

La représentation des individus dans le premier plan factoriel nous montre d'une part la qualité d'éducation des pays à travers leur position sur le premier axe et nous renseigne également sur leur niveau de gouvernance et de dépenses dans l'éducation. Les données ayant été centrées, le centre du plan représente le centre de gravité du nuage de point, ou encore le pays moyen.

[7] Voir lien empirique entre les variables (page)
[8] Voir lien empirique entre les variables (page)

Graphique7 : Représentation des observations dans le premier plan



Fait par les auteurs (XLSTAT)

Un premier constat est que le nuage de points individus est bien allongé autour du premier axe, contrairement au deuxième axe. Ceci explique le fait que l'inertie du premier axe soit égale à plus du triple de celle du deuxième axe.

Comme vu dans la présentation de la méthode[9], ce plan est le même que celui de représentation des variables. Donc les pays les plus performants se trouvent complètement à droite du plan et les moins performants en éducation se trouvent à gauche. Nous avons par exemple la Lettonie qui a la meilleure qualité d'éducation dans notre échantillon, en effet elle possède un taux d'achèvement de 93,713 et un taux d'alphabétisation de 99,8284. Nous avons ensuite Cuba qui possède le taux le plus élevé d'alphabétisation des adultes(99,8284). Comme

[9] Voir présentation de la méthode : page

autres pays performants en éducation nous retrouvons tous les pays ayant un taux d'achèvement de 100% à savoir Equateur, Ukraine, Tonga, Maldives, Mongolie, Thaïlande, Vietnam, Indonésie, Sri Lanka, Syrie, Belize, Kiribati et Guyana.

D'autre part, les pays les moins performants sont du côté négatif de l'axe. Ce sont pour la plus part ceux de l'Afrique francophone, notamment la Centrafrique qui est très bien représentée (0,710) et qui est a la plus basse qualité de l'éducation (coordonnée sur le premier axe). Cela s'explique par le fait qu'il possède le plus bas taux d'achèvement (31,825%). On a ensuite le Togo, le Mali, le Niger…

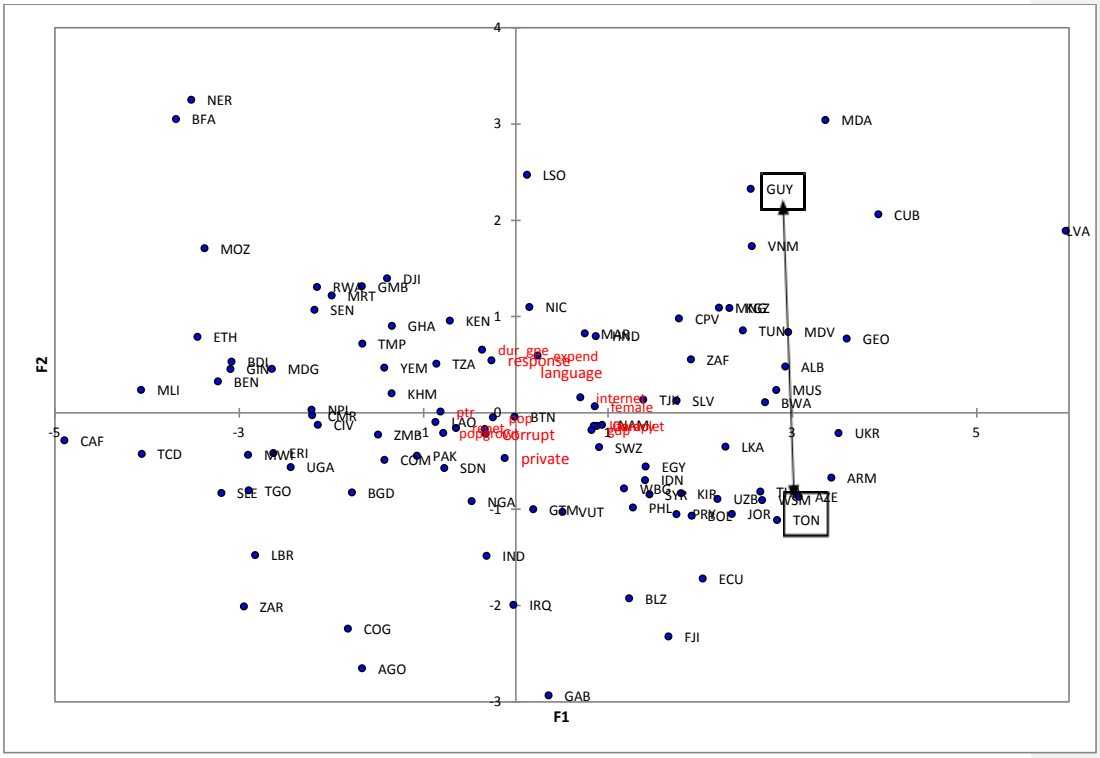
Le Burundi quant à lui, comme indiqué sur le schéma fait partir des pays les moins performants de l'Afrique.

Le Maroc fait parti des pays assez performants, cela se justifie par son taux d'achèvement élevé (80,73%) et par son taux d'alphabétisation des adultes (54,51%).

Les pays qui ont les mêmes coordonnées sur le premier axe ont la même qualité d'éducation, mais ils diffèrent par leurs valeurs au niveau du deuxième axe, c'est à dire par leurs dépenses dans l'éducation et ont de fortes valeurs pour les variables dont ils sont proches. Concrètement, nous pouvons comparer Guyana (GUY) et Tonga (TON) qui ont respectivement 0,296 et 0,423[10]comme qualité de représentation, et ont à presque les mêmes qualités de l'éducation :

[10] Tableau des qualités de représentation voir ANNEXE

Graphique8 : Comparaison des pays (Guyana-Tonga)



Fait par les auteurs (XLSTAT)

Toutes fois Guyana doit plus dépenser dans l'éducation que Tonga et aussi avoir une plus grande diversité linguistique. Ce dernier doit avoir plus d'écoles privées. Aussi, Guyana doit avoir plus d'internautes et de femmes enseignantes que Tonga, et celui-ci doit avoir de plus grands taux d'achèvement, d'alphabétisation et de revenus par habitants. Nous pouvons voir à travers le tableau suivant tiré de notre base que ces comparaisons sont vérifiées :

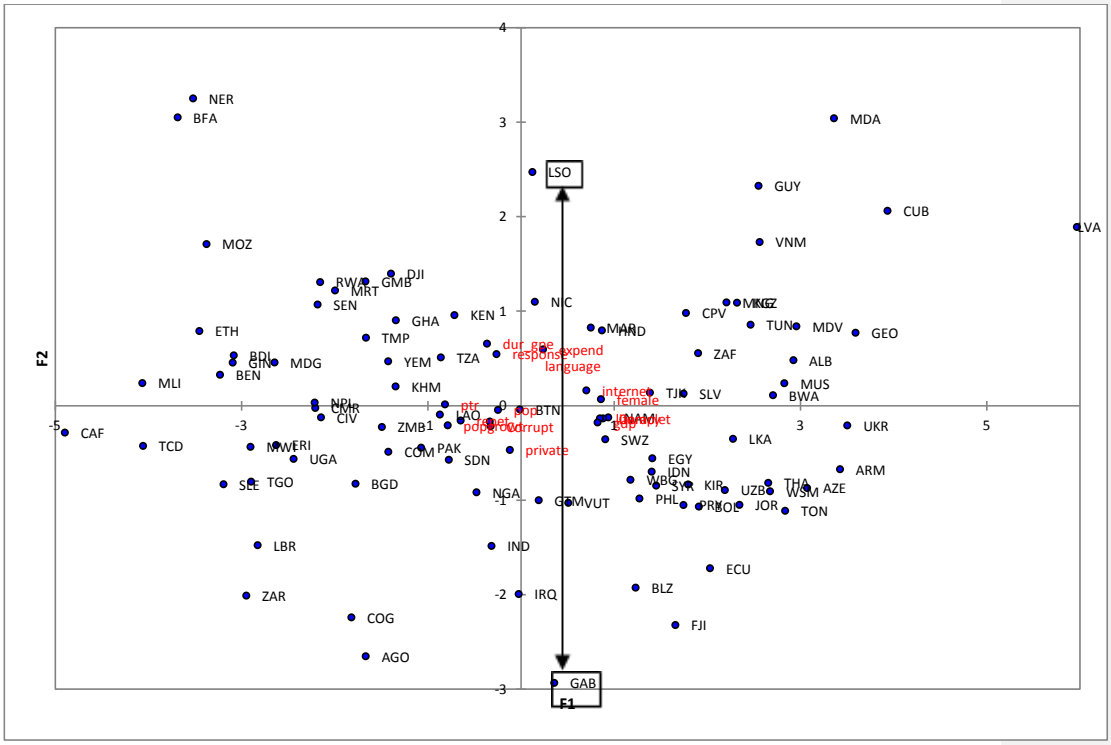
Tableau17 : Comparaison des pays (Guyana- Tonga)

PAYS	abr	Complet	literacy	internet	Gdp	female	expend	language	private
Guyana	GUY	100,00	78,63	24,36	7,88	87,29	11,16	0,10	2,29
Tonga	TON	100,00	99,02	6,39	8,33	62,99	9,75	0,00	8,79

Fait par les auteurs

Nous pouvons aussi comparer Lesotho(LSO) et la Gabon(GAB) qui ont presque les mêmes qualités de l'éducation. Ils ont également représentés presqu'au même pourcentage : 0,1% et 0,7% respectivement.

Graphique9 : Comparaison des pays (Lesotho-Gabon)



Fait par les auteurs (XLSTAT)

Le Lesotho possède de plus grandes valeurs que le Gabon pour toutes les variables qui se trouvent du côté positif du 2ème axe. Il s'agit de la durée de participation au Partenariat Mondial pour l'Education à l'éducation (dur_gpe), de la diversité linguistique (language), de la dépense dans l'éducation (expend) et du taux de réponse (response). Le Gabon quant à lui a un plus grand taux d'écoles privées que le Lesotho. En ce qui concerne les variables du premier axe, la grandeur des valeurs dépend de la proximité avec chaque variable. Nous pouvons voir cela dans ce tableau tiré de la base de données.

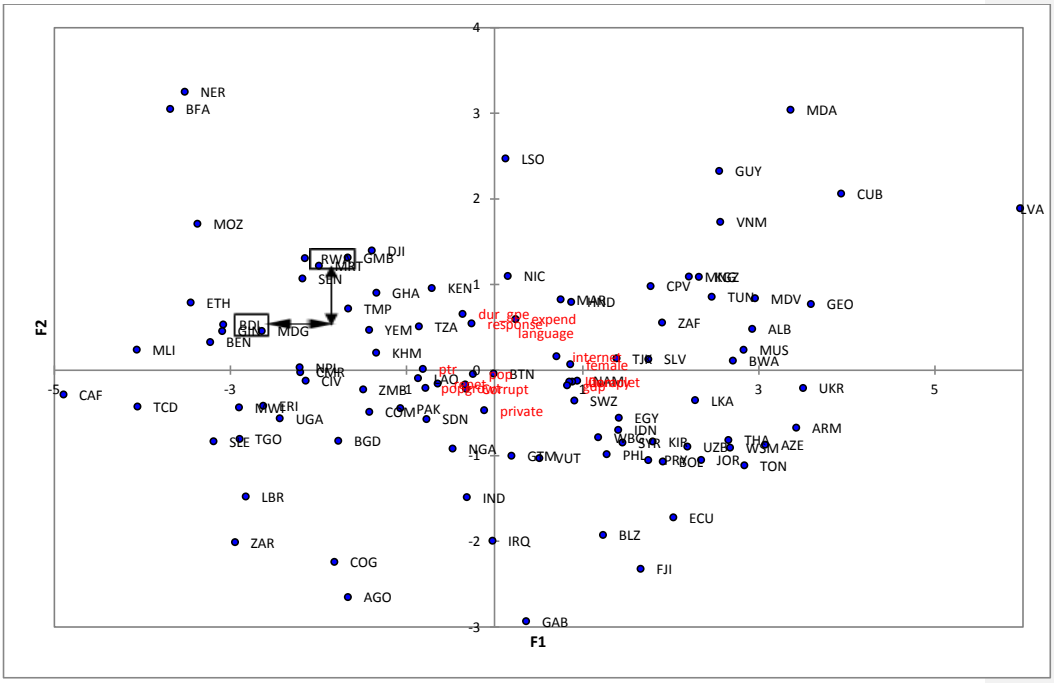
Tableau18 : Comparaison des pays (Lesotho-Gabon)

PAYS	abr	popgrowt	pop	ptr	repet	dur_gpe	Corrupt	expend	language	response	private
Lesotho	LSO	0,9	7,6	37,9	20,4	7	6,4	27,1	0,3	0,7	0,6
Gabon	GAB	1,9	7,3	36	24	0	7,0	4	0,3	0,2	29,3

Fait par les auteurs

Voyons ce qu'il en est aussi du Burundi et du Rwanda, pays voisins qui ont des géographies similaires et une langue commune. Ils ont respectivement comme qualité de représentation 0,357 et 0,220 et sont donc comparables :

Graphique10 : Comparaison des pays (Rwanda- Burundi)



Fait par les auteurs (XLSTAT)

Le Rwanda a de plus grandes valeurs que le Burundi pour les deux axes. Il possède donc une meilleure qualité de l'éducation ; sa position par rapport au deuxième axe nous montre qu'il a également une meilleure stratégie de gouvernance. En effet nous pouvons voir que les variables qui sont liées à la qualité de l'éducation ont une plus grande valeur au Rwanda qu'au Burundi :

Tableau19 : Comparaison des pays (Rwanda- Burundi)

Pays	Complet	literacy	internet	gdp
Burundi	41,9	66,6	0,6	5,8
Rwanda	46,8	70,6	2,0	6,8

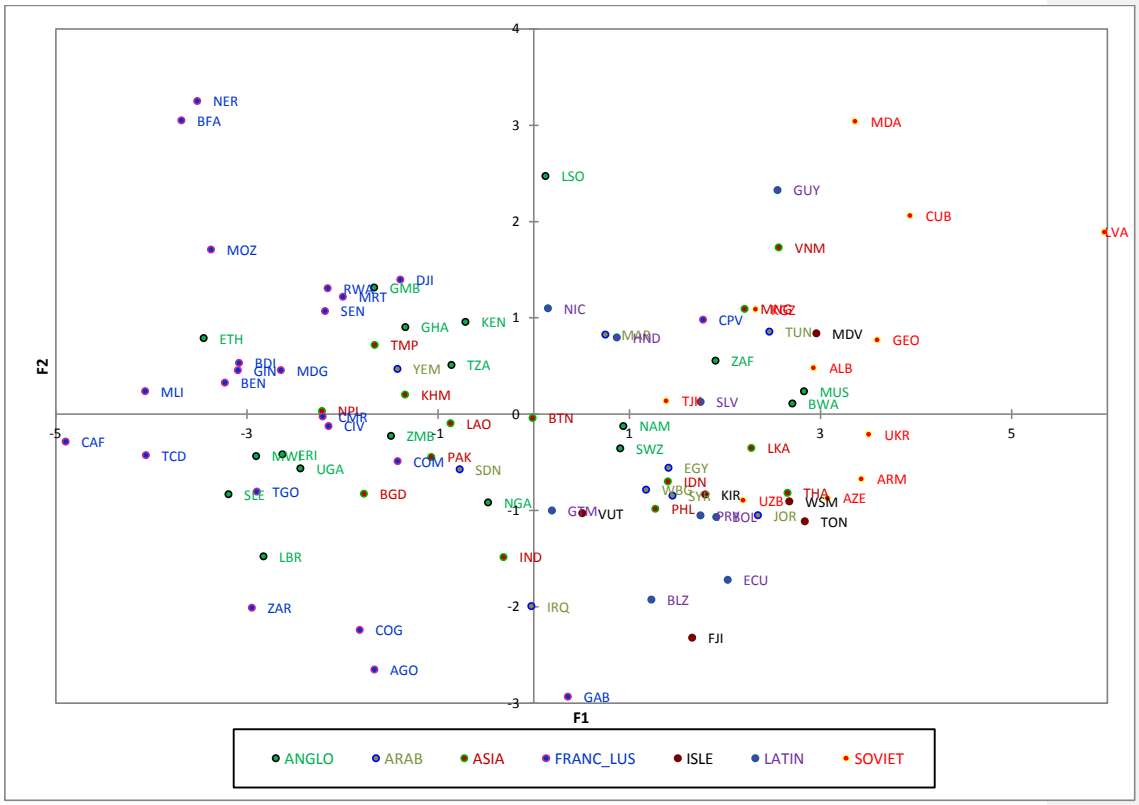
Fait par les auteurs

Le Burundi dépense dans l'éducation 19,4% de son PIB tandis que le Rwanda n'y consacré que 8,4%. Malgré cela, la position du Burundi par rapport au premier axe est dû au fait qu'il possède un très fort taux de redoublement, le plus élevé de tous les pays considérés, de 31,% contre 15,8% au Rwanda. Ce fort taux de redoublement serait donc le principal facteur affectant la qualité de l'éducation.

Nous remarquons également que la plus part des pays de l'Afrique francophone se trouvent à l'extrémité gauche du graphique et que les pays de l'ex block soviétique se trouvent à l'extrémité droite. Il est donc possible qu'il y ait un regroupement des différents

systèmes éducatifs. Cela nous permet donc de distinguer selon ces systèmes afin de voir leur effet sur la qualité de l'éducation. Faisons donc la représentation dans le premier plan factoriel en différenciant les systèmes éducatifs :

Graphique11 : Représentation des pays dans le premier plan avec distinction des systèmes



Fait par les auteurs (XLSTAT)

SOVIET : Rouge, ILES : Noir, ARAB : Olive, LATINS : Navy, FRANC_LUS : Bleu, ANGLO : Vert, ASIE : Marron.

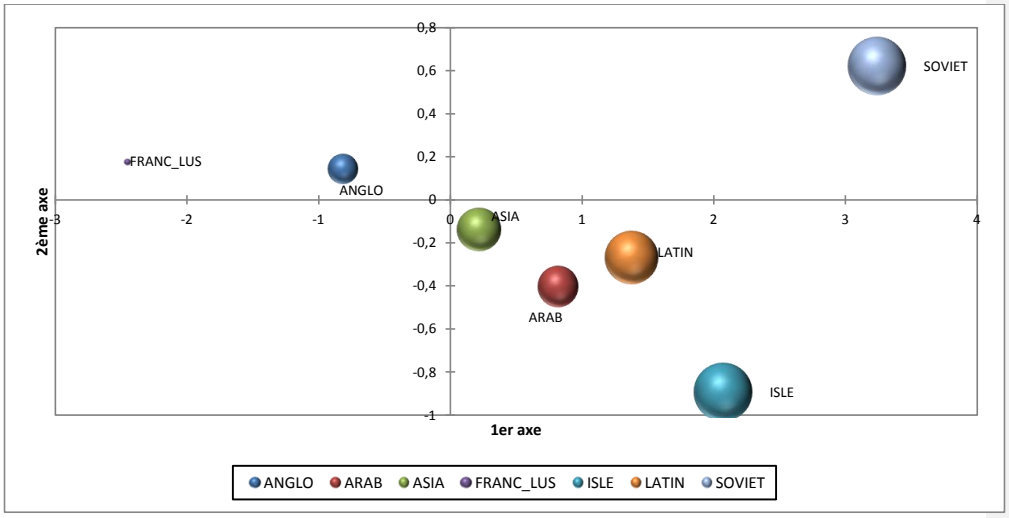
Nous remarquons qu'il ya certains systèmes qui sont meilleurs, quelque soit le pays. En effet, les pays de l'ex bloc soviétique ou ayant eu des économies centralisées (qui sont en rouge) c'est-à-dire les pays de l'Europe et de l'Asie centrale en plus du Cuba ont les meilleurs résultats d'éducation primaire. Parmi ces pays, il y a Lettonie et Cuba qui représente les valeurs extrêmes de l'analyse, ce sont les pays les plus performants de l'analyse. Cela confirme donc notre étude préliminaire où nous avions remarqué que les pays du bloc SOVIET avec les meilleures valeurs et une variance faible pour la variable complet (taux d'achèvement) et aussi pour toutes les variables qui lui sont liés positivement. Nous avions aussi remarqué qu'ils avaient de faibles valeurs pour les variables qui ont influence négative sur l'éducation, notamment le nombre d'élèves par enseignants et le taux de redoublement.

Ensuite nous voyons les Îles (en noir) qui sont pour la plus part meilleures. Nous avions noté également dans notre analyse descriptive que ce sont des pays qui avaient une faible population démographique et investissaient assez dans l'éducation. Les pays arabes suivent (en olive), ils ont pour la plus de bons résultats à part le Yemen qui est du côté négatif du premier axe. Ils sont en général les plus performants en Afrique. Nous avons ensuite les pays latins (en Navy) qui ont des résultats positifs. Ce sont des pays que nous qualifierons de moyens dans notre analyse, ils ont après les îles une faible population et une croissance faible. Les pays les moins performants en éducation primaire sont ceux de l'Afrique Francophone et lusophone (qui sont en bleus) mis à part quelques pays comme le Gabon qui n'est pas très mal placé. En effet il a un taux d'achèvement de 70% et un taux d'alphabétisation de 58,8%, et d'autre part il a en moyenne 36 élèves par enseignants et un taux de répétition de 24%. Ce groupe avait en moyenne de faibles valeurs pour les variables indiquant la qualité de l'éducation et de fortes valeurs pour celles qui nuisent à l'éducation. Ils sont suivis de ceux de l'Afrique anglophone (en vert) qui sont plus performants, parmi ces pays il y en a qui se démarquent comme la Namibie et l'Afrique du Sud. Pour ce qui des pays de l'Asie (en marron), ils sont dispersés sur le premier axe, il y en qui ont une très bonne qualité et d'autres qui se retrouvent dans la catégorie de l'Afrique francophone. En effet ces pays présentaient également une forte variance pour les différentes variables.

Pour ce qui est du deuxième axe, nous remarquons que les systèmes éducatifs ne se différencient pas vraiment. Comme nous l'avons vu plus haut, les pays qui dépensent plus dans l'éducation sont ceux de l'ex block soviétique. De plus, Moldova possède un taux assez faible de privé comme la plus part des pays de l'ex block soviétique. Ensuite nous pouvons remarquer le Niger et le Burkina Faso qui ont également de grandes valeurs pour le deuxième axe. En effet, ils font partie des plus anciens pays élus au partenariat mondiale pour l'éducation (10ans) et possèdent également une grande diversité linguistique. Il y a pour finir le Gabon qui possède la valeur la plus basse pour le deuxième axe, malgré qu'il ne possède pas une aussi forte dominance du privé (29,32) que le Fiji (98,9) ce qui indique qu'il ne dépense pas vraiment dans l'éducation.

Un schéma plus synthétique est présenté ci-dessous. Nous avons calculé les moyennes des coordonnées des différents groupes éducatifs dans le premier plan principal. La taille des bulles est une proportion de la moyenne du taux d'achèvement.

Graphique12 : Représentation des moyennes des systèmes éducatifs



Fait par les auteurs (XLSTAT)

Nous voyons premièrement que la répartition de la qualité de l'éducation va dans le même sens que le taux d'achèvement. Aussi, les îles dépensent peu dans l'éducation, contrairement aux pays de l'ex bloc soviétique qui sont d'ailleurs les meilleurs. Nous voyons également que les pays de l'Afrique francophone dépensent assez dans l'éducation mais n'ont pas de bons résultats ni de bons taux d'achèvement.

IV- Construction d'un modèle :régression par la méthode des moindres carrés ordinaire

4-1) Présentation du modèle

Nous avons ajusté un modèle utilisant la méthode des moindres carrés ordinaires. La variable dépendante est 'complet' ; le taux d'achèvement.

4-2) Résultats du modèle

Les résultats obtenus de cette régression sont récapitulés dans le tableau ci-dessous. Pour la lecture de ce modèle, nous précisons que l'effet de chacune des variables sur le score final standardisé est obtenu sur la colonne 'Valeur estimée des paramètres'. Si l'effet est négatif, on remarquera le signe – (moins) associé au coefficient et si l'effet est positif, il n'y a pas de signe devant le coefficient. Comme dans la plus part des modèles en éducation, la significativité est associée à un seuil d'erreur qui varie de 1 à 10%. Ainsi un effet jugé très significatif est associé à un seuil d'erreur inférieur ou égal à 1%, un effet significatif correspond à un seuil d'erreur compris entre 1 et 5% et un effet moyennement significatif de 5 à 10%.

Tableau20 : Présentation des modèles (MCO)

Modèles avec 'complet' comme variable dépendante				
	Modèle1		Modèle2 (sans literacy)	
Variable	Valeur estimée des paramètres	Tolérance	Valeur estimée des paramètres	Tolérance
Intercept	62.421**	.	86.335***	.
literacy	0.327***	0.25896		
internet	0.185	0.51761	0.158	0.51868
loggdp	2.197	0.31067	2.329	0.31078
female	0.078	0.25207	0.245***	0.37207
ptr	-0.218*	0.34332	-0.263**	0.34872
repet	-0.709***	0.50546	-0.684***	0.50634
popgrowt	-1.681	0.39632	-2.876*	0.41958
Corrupt	-0.739	0.73813	-0.524	0.74005
response	-6.982	0.73005	-11.079	0.74417
private	0.079	0.78226	0.068	0.78439
dur_gpe	-0.214	0.60123	-0.499	0.63927
language	0.086	0.79708	0.001	0.79907
logpop	-0.373	0.76595	-0.821	0.80343
expend	-0.263	0.69526	-0.298	0.69795
R^2	0.751		0.706	
R^2_{aid}	0,739		0,694	
F	18.14		51,058	
Pr > F	<.0001		<.0001	

Fait par les auteurs (SAS)

Note : Le signe *** signifie que l'effet est très significatif, ** indique que l'effet est significatif et * indique que l'effet est moyennement significatif.

- Interprétation des coefficients :

Parmi les coefficients du modèle1, il y a l'effet de la taille de classe qui est moyennement significative (p_{val} = 0,08). Ensuite, on a la constante qui est significative (p_{val} = 0,0358). Enfin, nous avons les effets du taux d'alphabétisation et du taux de redoublement qui sont très significatifs (p_{val} = 0,049 et 0,0011 respectivement).

Les variables literacy, internet, logdp, female, private ; comme envisagés plus haut à partir de la littérature, sont des facteurs qui améliorent la qualité de l'éducation. D'autre part les variables ptr, repet, popgrowt, corrupt, response, logpop ont un effet négatif sur la qualité de l'éducation. Cependant, le coefficient de la variable language est positif, ce qui voudrait signifier qu'elle aurait un effet positif sur la qualité de l'éducation. Nous constatons aussi que le coefficient de la variable expend est négatif, mais non significatif ce qui rejoindrait les études Hanushek et Kimko (2000) et de Altinok qui n'ont pas trouvé d'effet significatif de la dépense sur l'éducation. La variable de durée de participation au partenariat mondial pour l'éducation a également un coefficient négatif. Notons toutefois que ces effets contraires soulignés ne sont pas significatifs. Ils peuvent être également dus à la présence de multicolinéarité entre les variables explicatives. Il s'avère donc important de contourner ce problème de multicolinéarités.

- Tolérance et VIF :

La tolérance est une statistique utilisée pour déterminer le degré de la liaison linéaire entre une variable explicative X_i et les autres X_j.

Tolérance (X_i) = $1 - R^2$ (X_i; autres X_j)

Il est préférable d'observer une tolérance supérieure à 0,33. Par contre une variable qui a une très faible tolérance (proche de 0) contribue très faiblement dans le modèle et peut causer des problèmes dans les calculs.

La variance Inflation Factor (VIF) quant à elle se définit comme l'inverse de la tolérance : VIF = 1/ tolérance. Par ailleurs, il est préférable d'observer un VIF inférieur à 3.

Pour notre modèle, nous pouvons voir que la variable literacy a une tolérance de 0,25896 qui est inférieure à 0,33 ; ce qui signifie que cette variable est liée à plusieurs variables dans le modèle. Ce résultat est justifié car la variable literacy ou taux d'alphabétisation des adultes est, comme nous l'avons vu précédemment, très liée au taux d'achèvement ; elle mesure aussi en quelque sorte un taux d'achèvement passé. Il en est de même pour les variables gdp et female qui ont respectivement une tolérance de 0,31 et 0,25 ; montrant ainsi qu'elles sont liées à d'autres variables dans le modèle. Comme souligné plus haut, cette liaison peut entrainer des erreurs dans les calculs.

- Le coefficient de détermination :

Le coefficient de détermination ajusté R^2 mesure la qualité de l'ajustement des estimations de l'équation de régression. Il permet d'avoir une idée globale de l'ajustement du modèle. Il s'interprète comme la part de variance de la variable Y expliquée par la régression, varie entre 0 et 1 et s'exprime souvent en pourcentage. Un R^2 proche de 1 suffit pour dire que l'ajustement est bon.

Dans notre modèle, on a $R^2 = 75{,}1\%$, donc 75,1% de la variation du taux d'achèvement est expliqué par ce modèle. Ceci est très satisfaisant comme résultat, surtout dans le domaine de l'éducation où les R^2 mesurés sont généralement plus faibles.

Toutefois, en régression multiple, une valeur élevée du coefficient de détermination n'est pas suffisante pour affirmer que le modèle est bon, il est nécessaire d'effectuer un test sur la significativité de R^2 afin de savoir s'il existe une relation entre Y et les X_i. Ce test revient à effectuer un test de significativité globale sur le modèle à l'aide du test de Fisher. Notons que de faibles valeurs de F sont associées à des valeurs du R^2 proche de 0, et de fortes valeurs de F à des valeurs de R^2 proches de 1.

Nous testons donc l'hypothèse globale : $H_0 : \beta_i = 0$ contre H1 : il existe i tel que $\beta_i \neq 0$.

D'après la table ANOVA, la valeur de la statistique de Fisher observée est égale à :

$F_{obs} = 18{,}14$; la p value de ce test est donc très faible, ce qui implique que l'on rejette l'hypothèse H_0. Rejeter l'hypothèse H_0 revient à dire que le modèle est globalement significatif, c'est-à-dire que l'ensemble des variables expliquent bien le taux d'achèvement.

Lorsqu'on enlève la variable literacy du modèle (Modèle2), la variable female devient très significative. Le R^2 est de 74,64% pour ce modèle. Cela confirme le fait que les deux variables literacy et female soient corrélées et jouent à peu près le même rôle dans le modèle.

- Sélection des variables :

Le mode sélection STEPWISE nous donne le modèle :

Tableau21 : Sélection stepwise

Nb. de variables	Variables	R^2 ajusté
4	literacy / gdp / ptr / repet	0,739

Fait par les auteurs

Ce résultat est très intéressant car la taille de classe (ptr) et le taux de redoublement (repet) sont les variables clés du cadre Indicatif Fast Track utilisé dans les pays participant au Partenariat Mondial pour l'Education pour augmenter le taux d'achèvement. Nous confirmons donc l'intérêt du cadre indicatif Fast Track pour le pilotage des systèmes éducatifs sur une base empirique.

Nous remarquons que ce modèle réduit à quatre variables a les mêmes indices de qualités (R ajusté) que le modèle globale. En appliquant la sélection stepwise au modèle sans literacy nous obtenons :

Tableau22 : Sélection Stepwise sans la variable Literacy

Nb. de variables	Variables	R^2 ajusté
4	gdp / ptr / repet / female	0,692

Fait par les auteurs

Nous n'avons exactement le même R^2 ajusté que le modèle de départ sans la variable literacy. En effet, l'absence de certaines variables dans les sélections voudrait normalement dire qu'elles n'ont pas d'effet significatif, or nous voyons par exemple pour la variable female que c'est un problème de colinéarité. Le problème de colinéarité est donc flagrant dans notre modèle, ce qui nous pousse à avoir recourt à une méthode plus sophistiquée.

4-3) Effet des systèmes éducatifs

Dans le tableau suivant, nous avons ajouté des variables catégorielles représentants les différents systèmes cités plus haut dans le modèle, en gardant comme modalité de référence l'appartenance au block soviétique :

Tableau23 : Présentation des modèles MCO avec les systèmes éducatifs

Variable	Valeur estimée des paramètres	Tolérance	Valeur estimée des paramètres	Tolérance
Intercept	47.167*	.	46.895*	.
Literacy	0.328***	0.25083	0.381***	0.34630
Internet	0.256	0.49643	0.261	0.49677
Gdp	1.952	0.26536	2.279	0.27126
Female	0.094	0.21825		
Ptr	-0.227*	0.29797	-0.226*	0.29798
Repet	-0.446*	0.37040	-0.471**	0.37556
popgrowt	-2.35911	0.27529	-2.333	0.27535
Corrupt	-0.340	0.57141	-0.375	0.57174
response	-4.247	0.64143	-3.019	0.65241
Private	0.056	0.67610	0.045	0.69655

Variable	Valeur estimée des paramètres	Tolérance	Valeur estimée des paramètres	Tolérance
dur_gpe	0.009	0.52590	0.011	0.52593
language	6.708	0.65873	5.997	0.67912
Pop	-0.507	0.52290	-0.455	0.52542
expend	-0.175	0.70993	-0.166	0.71206
Anglo	7.434	0.16790	6.012	0.17796
Arabes	6.695	0.33632	5.32865	0.35736
Latins	11.142*	0.34645	10.706*	0.34860
Iles	11.827*	0.39263	10.833*	0.40275
Asie	11.258**	0.29540	10.050*	0.31387
Franc	0.681	0.10449	-1.223	0.11297
R^2	0.8174		0.8151	

Fait par les auteurs (SAS)

Note : Le signe *** signifie que l'effet est très significatif, ** indique que l'effet est significatif et * indique que l'effet est moyennement significatif.

Nous remarquons que le modèle n'est pas perturbé par l'inclusion des variables catégorielles. Aussi, la modalité franc a le vif le plus élevé de 9,57 et n'est pas du tout significative, suivie de female qui a un vif de 4,58. Le modèle en sans la variable female fournit les mêmes résultats de qualité, sauf que la modalité franc devient significative.

Nous pouvons donc conclure que l'effet des femmes est aussi dû à groupe de pays, car par exemple dans les pays du bloc soviétique il y a beaucoup de femmes enseignantes, 87,8% en moyenne, contrairement aux pays africains où la place des femmes n'est pas la même (35,3% en moyenne).

V- Régression des Moindres carrés partiels :

5-1) Présentation de la régression des moindres carrés partiels

5-1-1) Généralités

L'approche PLS (partial least squares) permet d'estimer un modèle d'équations structurelles, c'est-à-dire les équations à variables latentes et variables manifestes. Elle a été introduite pour la première fois par Wold 1979.

Dans le cadre des modèles d'équations structurelles, deux méthodes s'opposent : d'une part, la méthode par analyse de la structure de covariance (bien souvent appelée LISREL) développée par Jöreskog (1970) et, d'autre part, l'approche PLS. Herman Wold a toujours

opposé la première qui utilisait, selon ses termes, une « modélisation dure » (« hard modeling », hypothèses de distribution fortes, nécessité d'avoir plusieurs centaines d'observations) à la seconde basée sur une « modélisation douce » (« soft modeling », peu d'hypothèses de distribution et un très petit nombre d'observations suffit à son application). Les deux approches ont été comparées dans Jöreskog et Wold (1982).

L'approche PLS est une méthode très générale qui contient comme cas particulier l'analyse en composantes principales[11], l'analyse canonique, l'analyse des redondances, la régression PLS, l'analyse canonique généralisée au sens de Horst ou de Carroll, au niveau de la première composante (Tenehaus, 1999). Elle est issue d'une théorie ancienne, celle de l'estimation des moindres carrés et elle se base sur des régressions simples et multiples. En conséquent, elle nécessite peu d'hypothèses et c'est pour cette raison qu'elle est appelée modélisation douce (soft modeling, Wold (1982)).

Initialement mise au point pour des applications dans le domaine de la chimie, la régression PLS est aujourd'hui couramment appliquée dans les sciences humaines et sociales, tout particulièrement dans le domaine de l'économétrie.

PLS est une technique quantitative de décomposition spectrale étroitement liée à la régression sur composantes principale (PCR). Cependant, dans PLS, la décomposition est faite d'une manière légèrement différente.

Au lieu de décomposer d'abord la matrice spectrale *X* en un ensemble de vecteurs propres et de scores, et de les régresser contre les *Y* dans une étape séparée, PLS utilise l'information de *Y* en même temps que le processus de décomposition. Ceci implique que les variables expliquant le mieux Y seront plus fortement pondérées. De plus, les vecteurs propres et les scores calculés en utilisant PLS seront différents de ceux de PCR. L'idée principale de PLS est de donner le plus d'information possible sur *Y* dans les premiers vecteurs construits.

5-1-2) Notions de base et conditions d'utilisation de la méthode :

a- Quelques définitions

Variable manifeste : Une variable manifeste est une variable pour laquelle une mesure peut être directement recueillie (observée, mesurée, etc.).

Variable latente : Une variable latente correspond à une caractéristique qui n'est pas directement observable et qui ne peut donc pas être mesurée directement.

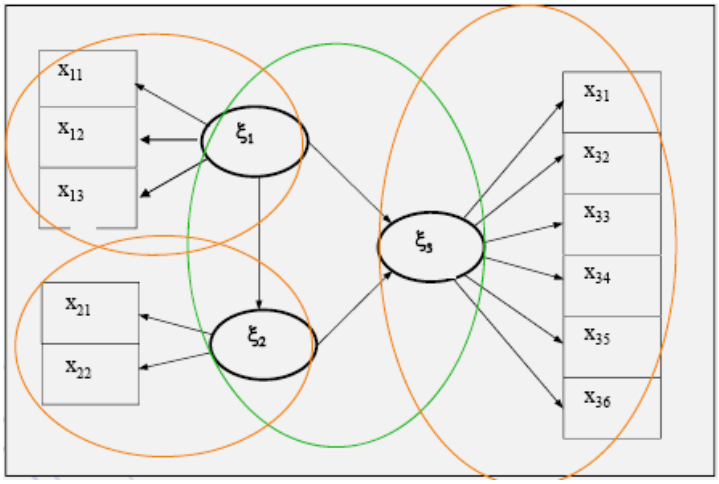
Cette séparation entre variables manifeste et variable latente est particulièrement adaptée à la psychométrie. On parle de trait latent pour parler de l'aptitude d'un enfant à un test et le score est une variable manifeste censée mesurée cette aptitude. Toute une théorie a été développée autour de ces concepts par le mathématicien Rasch dans les années 1960.

Le modèle de mesure ou externe est une sous-partie du modèle complet incluant les relations entre variables manifestes et latentes.

[11] Voir cour ACP

Le modèle de structure ou interne est une sous partie du modèle complet incluant les relations entre les variables latentes.

Graphique13 : Présentation des types de modèles



[12]

Modèle Interne Modèle Externe

Les X_{ij} désignent les variables manifestes et les ξ_i les variables latentes.

Il existe plusieurs schémas de modélisation du modèle externe qui modifieront la manière dont les variables latentes seront construites. Il existe trois façons de relier les variables manifestes aux variables latentes :

Le schéma réflectif : C'est un schéma de modélisation du modèle externe, c'est-à-dire de liaison des variables latentes avec leurs variables manifestes. Ce schéma est souvent adopté dans l'utilisation des modèles d'équation structurelles à variables latentes. Chaque variable manifeste est reliée à sa variable latente par une régression simple. Pour cela, il faut donc que les blocs soient unidimensionels, c'est-à-dire que chaque variable doit être étroitement liée à la variable latente.

Le schéma formatif : moins fréquemment utilisé, il suppose que chaque variable latente est une combinaison linéaire de ses variables manifestes correspondantes.

Le schéma MIMIC : est un mélange des deux premiers schémas.

Le schéma suivant nous montre clairement ces différents modes de liaison entre variable latente et variable manifeste.

[12] Source : Les modèles d'équations structurelles à variables latente, Emmanuel Jakobowicz.

Graphique14 : Schéma de sélection des variables latentes

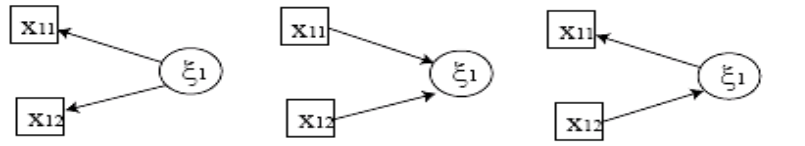


Schéma réflectif Schéma formatif Schéma MIMIC [13]

b- Conditions d'utilisation de la méthode:

La méthode des moindres carrés partiels est valide seulement sous certaines conditions:

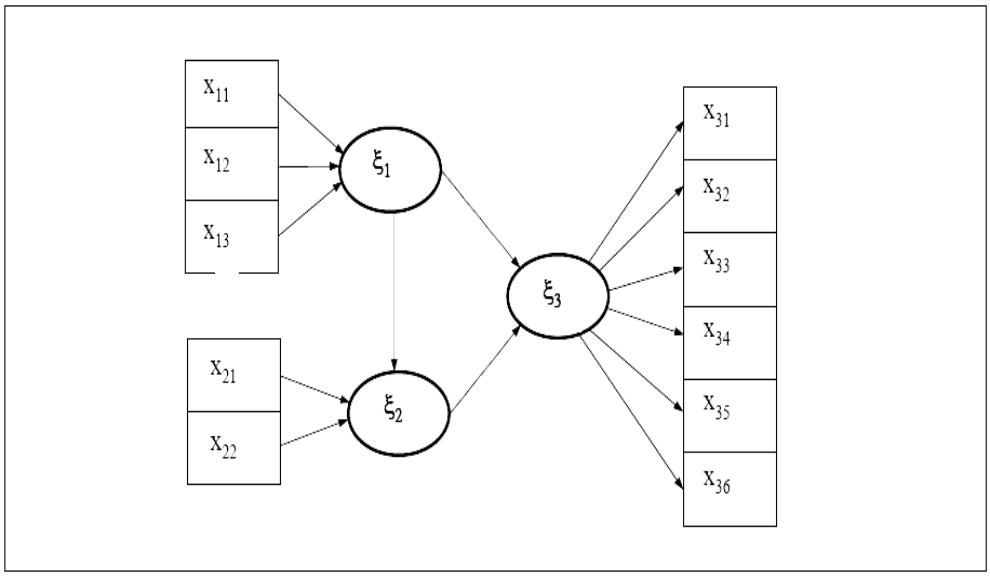
- Indépendance des observations (multiniveaux possible)
- Unidimensionnalité des blocs (dans le cas réflectif). Cela veut dire que pour le schéma réflectif, les variables doivent être assez corrélées entre elles.

5-1-3) Principe de la méthode

Comme souligné plus haut, cette méthode a été créée pour palier au problème de multicolinéarités dont souffrent plusieurs modèles de régression. Lorsque nous voulons estimer des coefficients d'un modèle de régression et qu'il y a un grand nombre de variables explicatives qui sont fortement corrélées, on parle de multicolinéarités. En présence de colinéarité parfaite, il suffit d'éliminer une variable pour rendre la matrice XX' inversible. Si la colinéarité n'est pas parfaite entre les variables explicatives, il est donc possible d'obtenir des estimations des moindres carrées ordinaire mais avec une forte variabilité car le déterminant de la matrice XX' sera proche de zéro. La technique de régression des moindres carrés partiels permet d'éviter que l'estimation des coefficients de régression soit perturbée par cette situation de multicolinéarité. L'idée de cette méthode est que le fait que les variables soient colinéaires implique qu'il existe d'autres variables appelées variables latentes ou composantes qui agiraient en même temps sur plusieurs variables du modèle comme le présente la figure ci-dessous tiré de '***Les modèles d'équations structurelles à variables latentes'*** de Emmanuel Jakobowicz (2006):

[13] Source : Emmanuel Jacobowicz 'Contributions aux modèles d'équations structurelles à variables latentes'.

Graphique15 : Présentation du modèles PLS



[14]

Les variables x_{11}, x_{12}, x_{13}, x_{21}, x_{22}, x_{31}, x_{32}, x_{33}, x_{34}, x_{35} et x_{36} sont les variables explicatives du modèle et ξ_1, ξ_2 et ξ_3 sont des variables latentes ou composantes qui agissent sur plusieurs variables explicatives et qui peuvent aussi agir mutuellement entre elles.

Les variables explicatives peuvent donc être réparties entre Q blocs de variables ou chaque bloc est fortement influencé par une variable latente unique.

$X = [X_1, \ldots, X_q, \ldots, X_Q]$

Dans cas précédent, nous avons par exemple trois blocs de variables (Q=3).

Chaque équation structurelle est composée de deux sous modèles à savoir le modèle externe ou de mesure et le modèle interne ou structurel.

Le modèle de structure comme définit plus haut peut donc s'écrire :

$$\xi_j = \beta_{0j} + \sum_q \beta qj\xi q + \epsilon_j$$

Où ξ_j est la variable générique endogène, β_{qj} le coefficient mesurant l'effet de la variable exogène ξ_q sur la variable endogène ξ_j et ϵ_j l'erreur.

[14] Source : Emmanuel Jacobowicz 'Contributions aux modèles d'équations structurelles à variables latentes'.

Le modèle de mesure quant à lui est une régression de l'estimation de la variable latente du block q sur les variables observées. Il existe plusieurs schémas de représentation du modèle externe comme cité plus haut.

Indépendamment du mode de mesure, à la convergence, chaque variable latente est exprimée comme combinaison linéaire des variables de son block :

$$\xi_q = \sum_{p=1}^{Pq} Wpq Xpq$$

Avec P_q le nombre de variables dans le block q. Les variables x_{pq} sont centrées et les w_{pq} sont leur poids externes. Ceux-ci sont obtenus à la convergence de l'algorithme et ensuite transformés pour donner des scores normalisés. Les poids externes peuvent être négatifs ; l'appellation « poids » n'est qu'un abus de langage désignant l'impact des variables manifestes sur leur variable latente.

5-1-4) Algorithme

L'approche PLS est basée sur un algorithme itératif qui alterne une construction des variables latentes en se basant sur le modèle externe avec une autre construction se basant sur le modèle interne. Après convergence, les coefficients du modèle peuvent être estimés par régressions ordinaires simples ou multiples.

L'algorithme d'estimation des scores des variables latentes des moindres carrés partiels se déroule comme expliqué dans le livre « Handbook of Partial Least Squares » écrit par Esposito (2010). Elle utilise trois méthodes, à savoir : le schéma barycentrique, les matrices des variables explicatives et la standardisation des scores de variables latentes et la régression des moindres carrés ordinaires.

Algorithme[15] :

Input : $X= [X_1, \ldots, X_q, \ldots, X_Q]$

Output: W_q, ξ_q, $\boldsymbol{\beta}_j$.

1) Pour tout q = 1,… , Q faire:
2) Initialiser W_q : deux options sont possibles. Soit on considère les valeurs du premier vecteur propre obtenu après une analyse en composantes principales, soit on affecte à toutes les variables la valeur 1 sauf à la dernière de chaque block qui prends la valeur -1. Les variables x_{pq} sont centrées et réduites.
3) **Estimation du modèle externe** : Soit v_q la variable latente centrée et réduite q. On calcule les scores du modèle externe, c'est-à-dire que l'on estime v_q comme étant une combinaison linéaire des x_{pq} pondérée par les w_{pq} :

$$v_q \propto \pm \sum_{p=1}^{Pq} W_{pq} X_{pq} = \pm X_q W_q$$

[15] Tiré de « Handbook of Partial Least Squares » écrit par V.Esposito, W.W.Chin, J.Henseler et H.Wang(2010), P. 56.

Le symbole ∝ signifie que le terme de gauche est égal au terme de droite standardisé et ± montre l'ambiguïté du signe. On choisit le signe de façon à ce que la variable latente ξ_q soit positivement corrélée au plus de colonnes de X_q possibles. Les éléments de W_q sont appelés les poids externes.

4) **<u>Estimation interne</u>** : On calcule les scores des variables latentes en se basant sur le modèle interne (chaque score associé à une variable latente est calculé en fonction des autres variables latentes qui lui sont liées)

$$V_q \alpha \sum_{q'=1}^{Q'} e_{qq'} v_{q'}$$

En posant :
$e_{qq'}$ = sign[cor(v_q, $v_{q'}$)] pour le schéma centroïde
$e_{qq'}$ = [cor (v_q, $v_{q'}$)], Schéma factoriel.
$e_{qq'}$ = Coefficient de régression dans la régression de v_q sur $v_{q'}$ si ces variables sont liées ou cor (v_q, $v_{q'}$), pour ce qui est du schéma structurel.
Ce choix s'effectue sans aucune condition, le plus utilisé est le schéma centroïde.

5) A partir de cette étape, il existe deux façons de mettre à jour les poids externes :

w_{pq} = cov(x_{pq} , V_q) pour le cas réflectif (Mode A)

$Wq = (X_q'X_q/N)^{-1} (X_q' V_q /N)$ pour le cas formatif (Mode B)

6) fin de la boucle
7) les étapes 1 à 7 sont répétées jusqu'à ce que l'algorithme converge, c'est-à-dire

$$\text{Max}\{w_{pq,\text{itération actuelle}} - w_{pq,\text{itération précédente}}\} < \Delta$$

Où Δ est un seuil de convergence, souvent il est mis à 0,0001 ou moins.

Il faudra noter qu'il n'y a pas de preuve rigoureuse que l'algorithme converge lorsqu'il y a plus de deux blocks de variables. Mais elle converge toujours en pratique.

8) Après la convergence, il faut calculer les coefficients structurels par régressions ordinaires.

$$\beta_j = (\Xi'\Xi)^{-1} \Xi'\xi_j$$

Avec $\Xi = [\xi_1 \ \ldots\ldots\ldots \xi_q]$

Pour résumer, l'algorithme d'estimation des poids se comme suit:

(1) Fixer arbitrairement les poids externes w_{kj}
(2) Calculer υ_q avec l'équation (3)
(3) Calculer Vq avec l'équation (4)
(4) Recalculer les poids externes en appliquant le mode A ou le mode B.
(5) Si converge, aller en (8)
(6) Aller en (3)
(7) Calculer les coefficients structurels par régressions ordinaires.

L'algorithme converge avec une probabilité 1 pour Q≤ 2 (Lyttkens et al., 1975). Au-delà de deux blocs, cette convergence n'a été que constatée dans la pratique.

5-1-5) Les indices de qualité :

- L'indice de communauté, évalue la qualité du modèle externe pour le q^{ème} bloc.

$$\text{Com}_q = \frac{1}{Pq}\sum_{p=1}^{Pq} cor^2(Xpq, \xi q) \quad \forall \text{ q : Pq >1}$$

C'est la moyenne quadratique des corrélations de la variable latente avec les variables du q^{ème} bloc. Cet indice mesure à quel degré la variabilité de chaque bloc de variables observées est expliquée par la variable latente. En d'autres termes, la proportion des variables manifestes expliquée par leur variable latente associée. Une mesure générale est :

$$\overline{\text{Com}} = \frac{1}{\sum_{q:Pq>1} Pq}\sum_{q:Pq>1} PqComq$$

- L'indice de redondance : Pour montrer la qualité de prédiction du modèle, cet indice permet de mesurer à quel niveau la variable latente endogène ξ_j du block j explique la variabilité des variables exogènes qui agissent sur elles. Il évalue la qualité du modèle structurel pour chaque bloc endogène q en prenant en compte le modèle de mesure.

$$\text{Red}_j = \text{Com}_j * R^2(\xi_j , \xi_{q\,:\,\xi q \to \xi j})$$

Une mesure plus globale du modèle est :

$$\overline{\text{Red}} = \frac{1}{J}\sum_{j=1}^{J} Redj$$

Cette mesure part de la même idée que la communauté mais la variable latente est remplacée par son estimation à partir des variables latentes voisines.

Il n'y a pas d'indice d'ajustement global dans la régression des moindres carrés partiels, mais Tenehaus et al.(2004) a proposé un indice de mesure de la qualité :

- L'indice GoF (*Goodness of Fit*) : Elle est la moyenne géométrique de la moyenne des communautés sur l'ensemble des variables latentes (Com) et de la moyenne des R² associés aux variables latentes endogènes (R²).

$\overline{\text{GoF}} = \sqrt{\overline{Com} * \overline{R^2}}$ ($\overline{\text{Com}}$ et $\overline{\text{R}}^2$ sont des moyennes sur tous les blocks)

$\overline{R}^2 = \frac{1}{J}\sum R^2(\xi_j , \xi_{q\,:\,\xi q \to \xi j})$.

5-1-6) Limites de la méthode

L'approche PLS étant basée sur un algorithme itératif, peu de propriétés théoriques ont été démontrées. Certaines sont communément supposées car elles se vérifient dans la pratique. La

multiplicité des modes et des schémas d'estimation rend l'obtention de propriétés générales d'autant plus difficile.

De plus dans les données sur l'éducation, il faut pouvoir prendre en compte les poids de sondage (très fréquents) et également le caractère hiérarchique des données (niveau élève, niveau classe, niveau école, ce qui a tendance à compliquer les analyses.

De plus, contrairement aux MCO, la méthode PLS ne donne pas d'intervalles de confiance pour les prévisions.

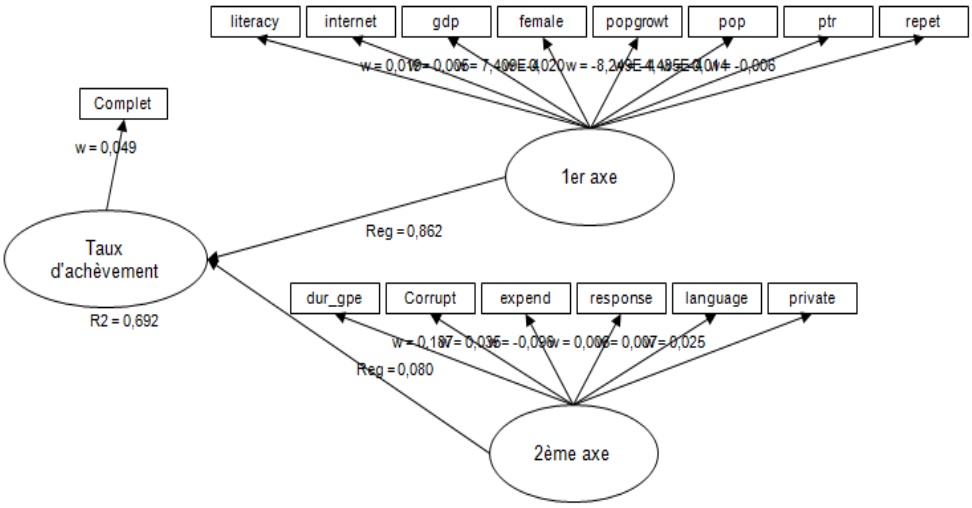
5-2) Application

Nous cherchons à effectuer une régression des moindres carrés partiels de la réponse univariée qui est le taux d'achèvement de l'école primaire sur les facteurs : le taux d'alphabétisation des adultes, un indice de l'accès à internet, le PIB par habitant (qui est un indice du revenu des parents), le taux de femmes enseignantes, le nombre d'élèves par enseignants, le taux de répétition, la croissance démographique, un indice de corruption, le taux de réponse, la part du privé, la durée de participation à l'initiative fastrack, la diversité des langues, la population du pays en 2009 et la dépense dans l'éducation.

Modèle0 :

Nous pouvons donc dans un premier temps partir des résultats de l'ACP pour faire un modèle (Modèle0). Nous prendrons donc deux variables latentes explicatives dont la première regrouperait les variables qui forment le premier axe et la deuxième, celles qui forment le deuxième axe. Nous obtenons le modèle suivant :

Graphique16 : Modèle0



Fait par les auteurs (XLSTAT)

Nous pouvons voir ici les indices de communauté qui sont les carrés de corrélation des variables manifestes avec leur variable latente :

Tableau24 : Indices de communalités (Modèle0)

Variable latente	Variables manifestes	Communalités
Taux d'achèvement	Complet	
1er axe	Literacy	0,842
	Internet	0,359
	Gdp	0,504
	Female	0,889
	Popgrowt	0,507
	Pop	0,041
	Ptr	0,605
	Repet	0,360
2ème axe	dur_gpe	0,286
	Corrupt	0,039
	Expend	0,457
	Response	0,018
	Language	0,024
	Private	0,165

Fait par les auteurs (XLSTAT)

Les variables literacy, female, gdp,ptr et popgrowt ont les indices de communauté les plus élevés et tous supérieurs à 0,5. Ces mêmes variables sont celles, en plus de repet, qui avaient servies à expliquer le premier axe de l'ACP. Les indices de communauté de la deuxième variable latente que nous avons noté 2ème axe ne sont pas relativement élevés, comme les contributions dans l'ACP. Toutefois, comme on pouvait le prévoir à partir de l'ACP, les variables les plus corrélées sont les dépenses, la durée de gpe et ensuite le taux d'écoles privé dans le pays.

L'indice de redondance est ici le R^2 de la régression car nous sommes en présence d'une seule variable dépendante. Il est de 0,692 comme indiqué sur le diagramme.

Le tableau suivant nous montre l'influence de chaque variable latente dans le modèle :

Tableau25 : Effet des variables latentes (Modèle0)

Variable latente	Valeur	Ecart-type	t	Pr > |t|
1er axe	0,862	0,066	13,136	0,000
2ème axe	0,080	0,066	1,216	0,227

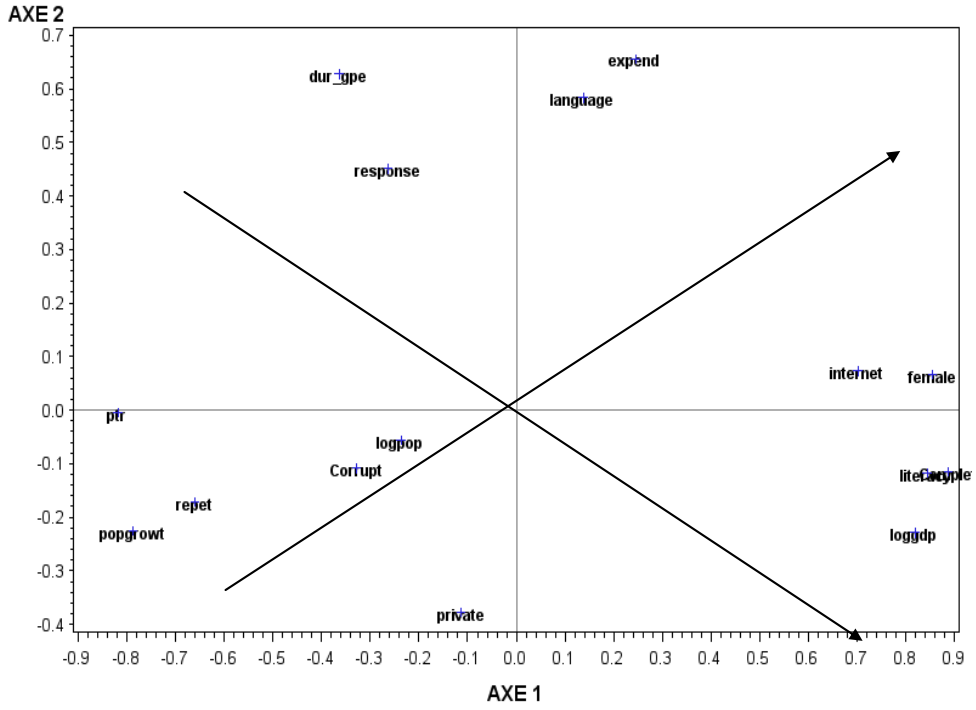
Fait par les auteurs (XLSTAT)

Nous voyons que la première variable a un effet important (0,862) et très significatif dans le modèle contrairement à la deuxième variable latente dont l'effet n'est pas significatif. Aussi, le Gof du modèle, qui mesure à la fois la qualité du modèle interne et du modèle externe, est

de 0,502, ce qui nous montre que le modèle est assez bon et que le premier axe est le principal facteur qui explique le taux d'achèvement. Ce modèle est donc une confirmation de notre étude ACP.

Partant toujours de l'ACP, nous pouvons faire un modèle en tenant compte de deux axes qui forment le premier axe :

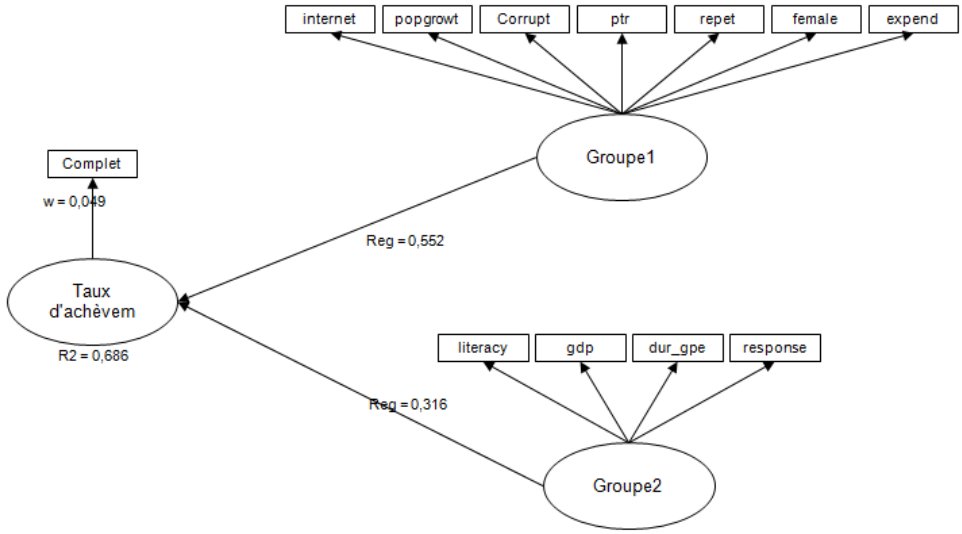
Graphique17 : Représentation des variables dans le premier plan (choix d'un modèle)



Fait par les auteurs

Nous allons donc regrouper les variables qui sont liées (positivement ou négativement) à chaque axe que nous allons regrouper en deux variables latentes.

Graphique18 : Modèle0'



Fait par les auteurs (XLSTAT)

Nous remarquons que les indices de qualité sont plus parlant ceux du premier modèle :

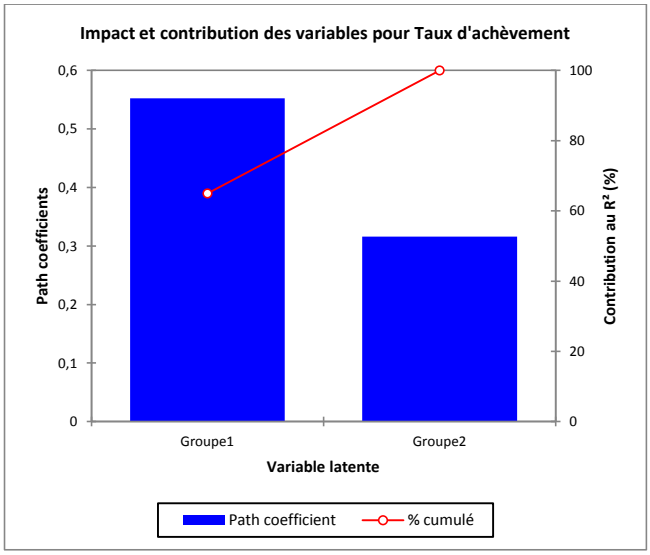
Tableau26 : Indices de communalités (Modèle0')

Variable latente	Variables manifestes	Communalités
Taux d'achèvement	Complet	
Groupe1	internet	0,400
	popgrowt	0,491
	Corrupt	0,042
	ptr	0,669
	repet	0,430
	female	0,906
	expend	0,062
Groupe2	literacy	1,000
	gdp	0,402
	dur_gpe	0,149
	response	0,082

Fait par les auteurs (XLSTAT)

Nous voyons également que les deux groupes ont un effet significatif sur le taux d'achèvement, ce qui n'était pas le cas pour le premier modèle.

Graphique19 : Effet des variables latentes (Modèle0')



Fait par les auteurs (XLSTAT)

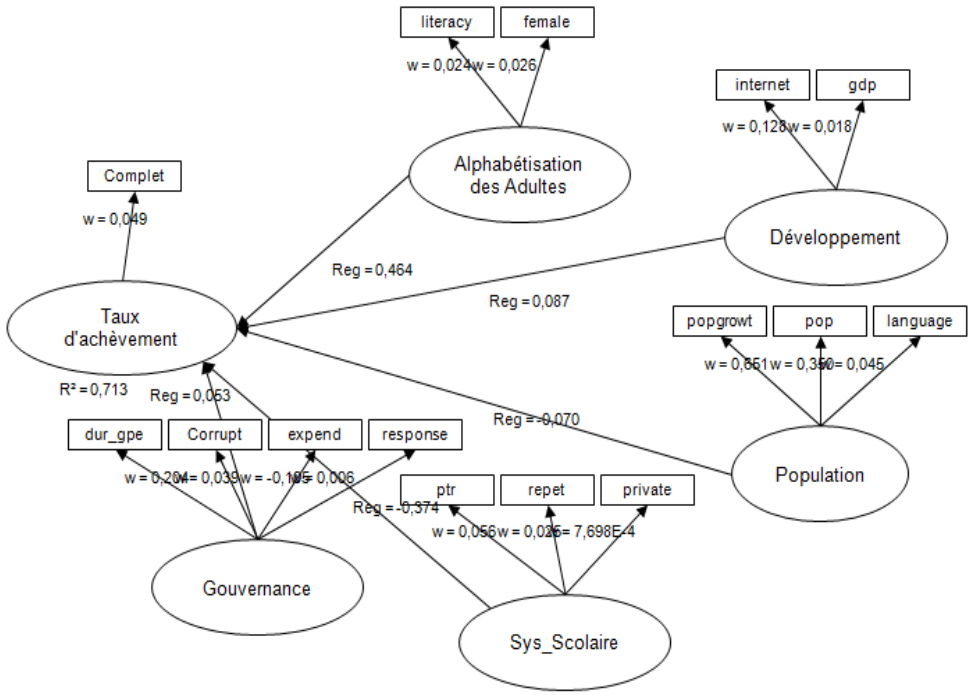
Modèle1 :

Le but de notre modélisation est de faire ressortir l'effet des facteurs latents qui sont mesurés à travers les différentes variables. Nous allons donc aller plus loin en catégorisant les variables d'un point de vue théorique en leur affectant des définitions que nous pensons logiques et qui seront vérifiées par la suite. Nous avons regroupé les variables explicatives en cinq groupes :

6- Alphabétisation des adultes : cette variable latente contient les variables literacy et female. Literacy désigne directement le taux d'alphabétisation des adultes, et female la proportion de femmes enseignantes qui indique aussi le taux d'alphabétisation des femmes.
7- Développement : cette variable est en quelques sortes une mesure du niveau de vie des pays. Elle contient la variable internet qui est le taux d'internautes et gdp, le PIB par habitant.
8- Population : c'est une mesure de la réalité démographique du pays. Elle contient le nombre d'habitants du pays, le taux de croissance démographique et la diversité linguistique au sein de la population.
9- Sys_Scolaire : elle mesure la structure du système scolaire dans le sens négatif. Elle contient les variables ptr et repet qui sont des défaillances de certains systèmes scolaires et la variable private qui est le taux de privé dans le pays. Nous l'appellerons aussi défaillance du système scolaire.
10- Gouvernance : c'est une mesure des stratégies du gouvernement. Il s'agit par exemple des dépenses dans l'éducation, de la corruption dans le pays, du taux de réponse du pays et de la durée de participation au partenariat mondial d'aide à l'éducation.

Nous pouvons donc faire notre premier modèle (Modèle1) en utilisant comme variable réponse le taux d'achèvement :

Graphique20 : Modèle1



Fait par les auteurs (XLSTAT)

Nous pouvons voir à travers le tableau suivant les indices de communauté de chaque variable manifeste avec sa variable latente :

Tableau27 : Indices de communalités (Modèle1)

Variable latente	Variables manifestes	Communalités
Taux d'achèvement	Complet	
Alphabétisation	literacy	0,883
	female	0,922
Développement	internet	1,000
	gdp	0,356
Population	popgrowt	0,599
	pop	0,552
	language	0,135
Sys_Scolaire	ptr	0,975
	repet	0,493
	private	0,000
Gouvernance	dur_gpe	0,452
	Corrupt	0,036

	expend	0,446
	response	0,014

Fait par les auteurs (XLSTAT)

Nous remarquons que la répartition plus fine est plus représentative que la première à travers les valeurs élevées des indices de communautés. La variable alphabétisation est très liée à ses variables latentes, de même que la variable Développement. La variable population l'est également à l'exception de la diversité des langues qui n'est pas très corrélée. Le système scolaire est quant à lui bien corrélé avec ptr et repet, mais repet n'est pas corrélé. Au niveau de la gouvernance aussi les variables corrupt et response ne sont pas très corrélées à la variable latente.

Nous remarquons également au niveau du R^2 que ce modèle est meilleur au précédent (R^2=0,713). L'indice Gof est de 0,591, plus élevé que celui du modèle 0.

Regardons à présent à travers ce tableau les variables qui impactent le taux d'achèvement à l'école primaire :

Tableau28 : Effet des variables latentes (Modèle1)

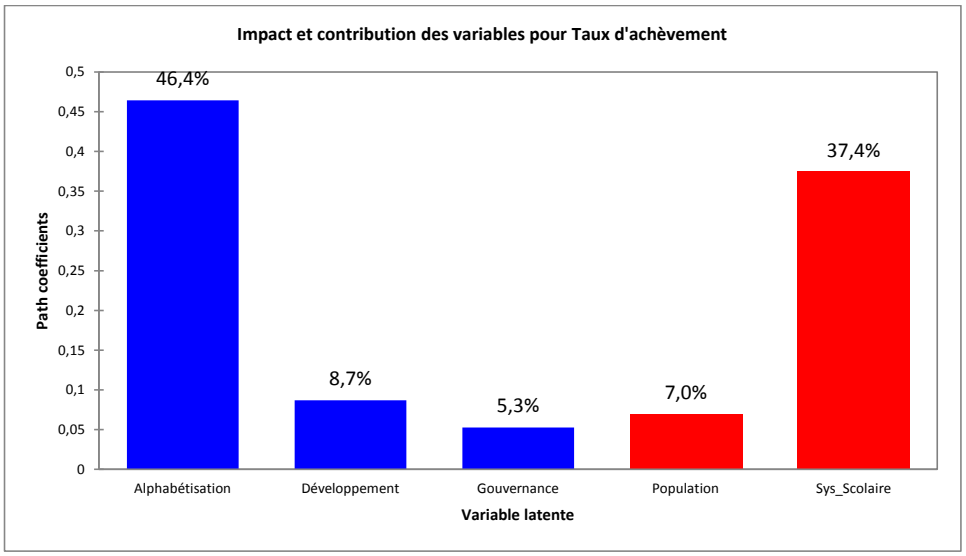
Variable latente	Valeur	Ecart-type	t	Pr > |t|
Alphabétisation des adultes	0,464	0,087	5,336	0,000
Développement	0,087	0,075	1,167	0,247
Population	-0,070	0,074	-0,939	0,351
Sys_Scolaire	-0,374	0,085	-4,390	0,000
Gouvernance	0,053	0,065	0,816	0,417

Fait par les auteurs (XLSTAT)

Les deux variables significatives sont l'alphabétisation des adultes et le système scolaire. Nous pouvons donc conclure comme souligné dans la littérature que l'alphabétisation des adultes est un véritable facteur encourageant la réussite des élèves au primaire. Certains systèmes scolaires favorisent le redoublement et de grandes tailles de classes, nous voyons ici que cela à un effet négatif significatif sur la réussite des élèves. Aussi, les facteurs de population ont un effet négatif mais qui n'est pas significatif. En effet, les facteurs démographiques constituent un défi qui peut être couvert par la prévoyance de l'Etat. Bien que n'étant pas significatif, nous voyons aussi qu'une bonne gouvernance va dans le même sens que le taux d'achèvement des élèves.

Nous pouvons voir cet effet de chaque variable à travers ce graphique :

Graphique21 : Effet des variables latentes (Modèle1)



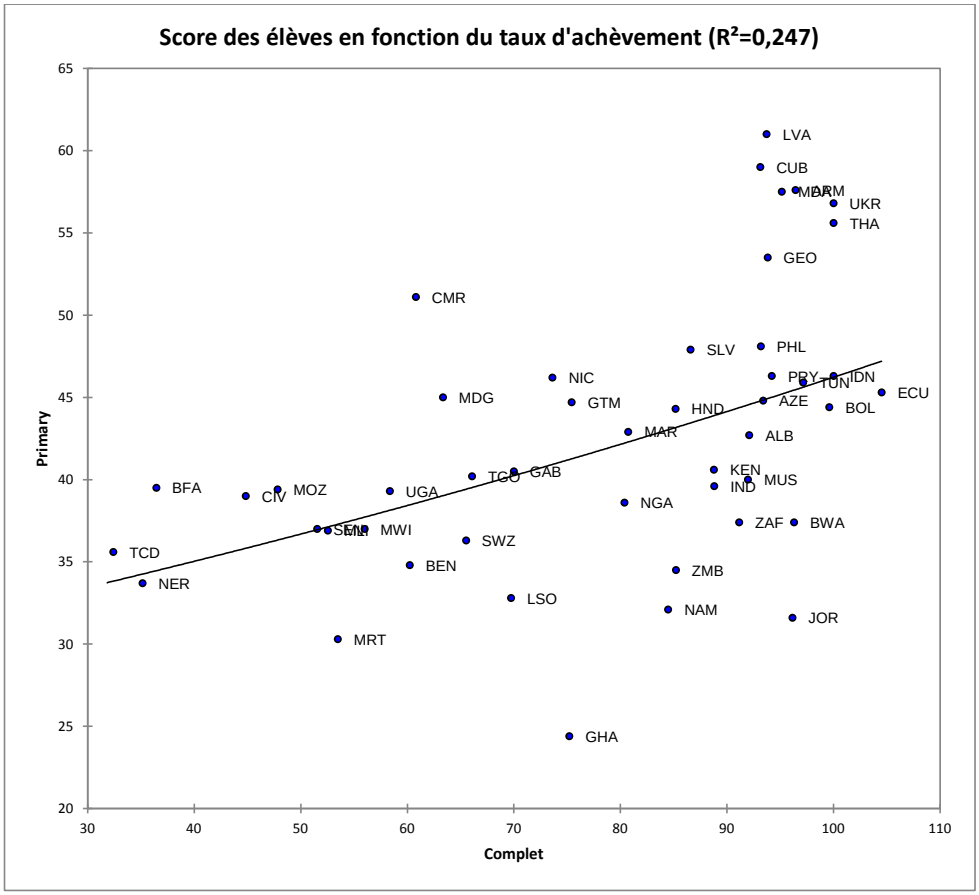
Fait par les auteurs (XLSTAT)

Note : En bleu les effets positifs et en rouge les effets négatifs.

Modèle2 :

En plus du taux d'achèvement, nous avons les résultats aux tests de 49 pays comme mesure de la qualité de l'éducation. Nous pouvons voir à travers le graphe la répartition des scores en fonction du taux d'achèvement :

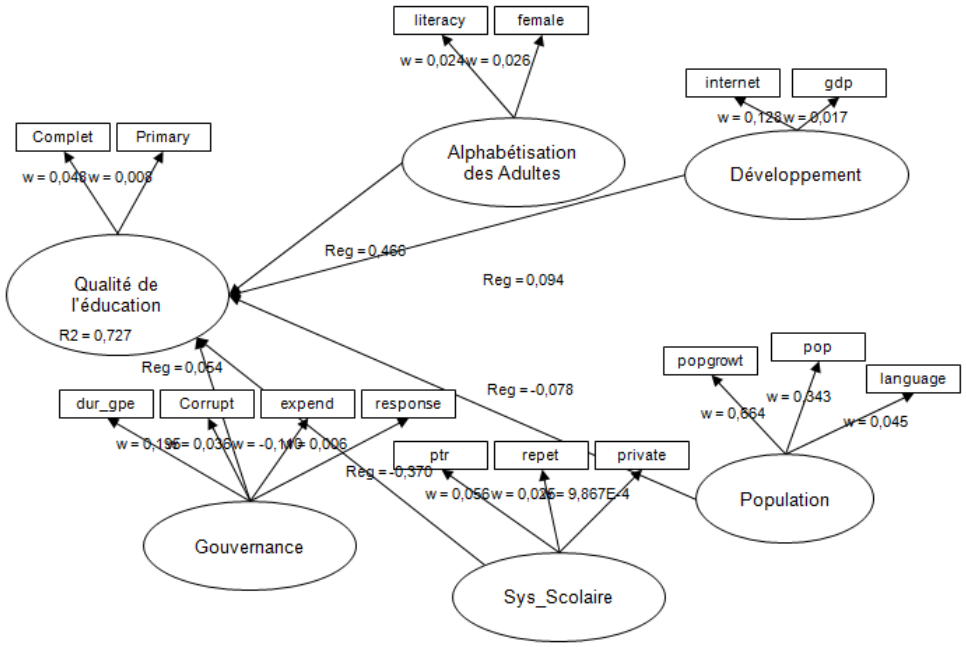
Graphique22 : Score des élèves en fonction du taux d'achèvement



Fait par les auteurs (XLSTAT)

Nous pouvons donc faire le modèle2, où nous auront deux variables dépendantes qui seront affectées à une variable latente :

Graphique23 : Modèle2



Fait par les auteurs (XLSTAT)

Nous pouvons voir les indices de communauté dans le tableau suivant :

Tableau29 : Indices de communalités (Modèle2)

Variable latente	Variables manifestes	Communalités
Qualité de l'éducation	Complet	0,998
	Primary	0,150
Alphabétisation	literacy	0,883
	female	0,922
Développement	internet	1,000
	gdp	0,356
Population	popgrowt	0,616
	pop	0,535
	language	0,135
Sys_Scolaire	ptr	0,975
	repet	0,493
	Private	0,000
Gouvernance	dur_gpe	0,404
	Corrupt	0,035
	Expend	0,494
	Response	0,012

Fait par les auteurs (XLSTAT)

Nous remarquons que le taux d'achèvement est fortement corrélé à la variable qualité de l'éducation tandis que la variable de résultat aux tests lui est faiblement corrélé. Pour ce qui est des autres indices, ils sont répartis comme dans le modèle1. Nous pouvons voir sur le diagramme que le R^2 à légèrement augmenté (0,727 contre 0,713 au modèle1). L'indice de mesure global du modèle a également augmenté légèrement, soit 0,603 contre 0,591 modèle1. L'indice de redondance relatif au taux d'achèvement est de 0,726 tandis que celui relatif aux résultats des scores est de 0,109. Nous pouvons également voir à travers le tableau ci-dessous que les variables qui agissent sont les mêmes que celles du modèle1.

Tableau30 : Effet des variables latentes (Modèle2)

Variable latente	Valeur	Ecart-type	t	Pr > |t|
Alphabétisation	0,466	0,085	5,476	0,000
Développement	0,094	0,073	1,287	0,202
Population	-0,078	0,073	-1,065	0,290
Sys_Scolaire	-0,370	0,083	-4,441	0,000
Gouvernance	0,054	0,063	0,854	0,395

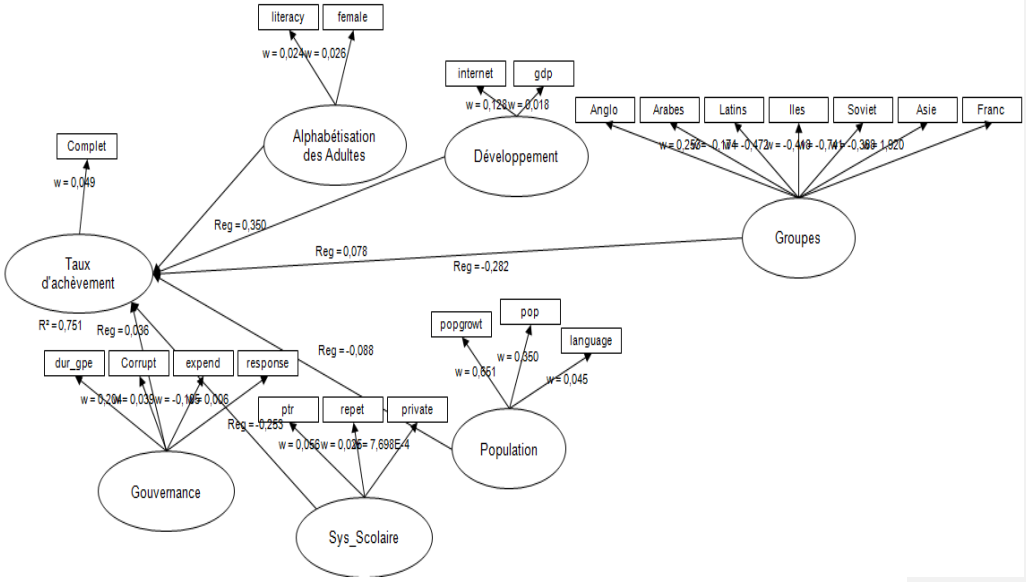
Fait par les auteurs (XLSTAT)

Vu que les modèles 1 et 2 donnent les mêmes résultats et présentent des indices de qualité semblables, il serait donc mieux d'opter pour le modèle1 pour plus de simplicité.

Modèle3 :

Nous allons associer ici la variable de groupes de systèmes éducatifs que nous avions introduit pour voir sa pertinence son effet dans les modèles PLS :

Graphique24: Modèle3



Fait par les auteurs (XLSTAT)

Les indices de communauté restent les mêmes pour les autres variables. Pour ce qui est de la variable groupe désignant les différents groupes éducatif, nous voyons que la variable est plus corrélée avec la variable indicatrice d'appartenance aux pays de l'Afrique francophone. Cela est dû à leur nombre élevé dans la base par rapport aux autres.

Tableau31 : Indices de communalités (Modèle3)

Variable latente	Variables manifestes	Communalités
Taux d'achèvement	Complet	
Alphabétisation	literacy	0,883
	female	0,922
Développement	internet	1,000
	gdp	0,356
Population	popgrowt	0,599
	pop	0,552
	language	0,135
Sys_Scolaire	ptr	0,975
	repet	0,493
	private	0,000
Gouvernance	dur_gpe	0,452
	Corrupt	0,036
	expend	0,446
	response	0,014

Groupes	Anglo	0,001
	Arabes	0,025
	Latins	0,067
	Iles	0,037
	Soviet	0,153
	Asie	0,083
	Franc	0,909

Fait par les auteurs (XLSTAT)

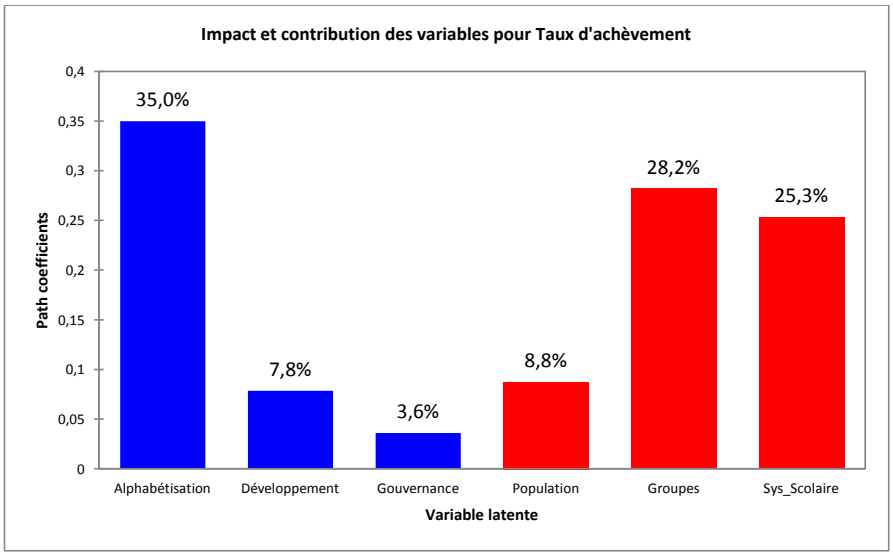
Comme nous le voyons sur le diagramme, le R^2 a augmenté (0,751). L'indice Gof est de 0,534. Le modèle interne s'est donc amélioré, tandis que le modèle externe, c'est-à-dire la mesure des variables manifestes par les variables latentes, est moins bon. Voyons la contribution de chaque variable au modèle :

Tableau32 : Effet des variables latentes (Modèle3)

Variable latente	Valeur	Ecart-type	t	Pr > |t|
Alphabétisation	0,350	0,088	3,991	0,000
Développement	0,078	0,070	1,121	0,266
Population	-0,088	0,070	-1,259	0,212
Sys_Scolaire	-0,253	0,087	-2,917	0,005
Gouvernance	0,036	0,061	0,592	0,555
Groupes	-0,282	0,079	-3,561	0,001

Fait par les auteurs (XLSTAT)

Graphique25 : Effet des variables latentes (Modèle3)



Fait par les auteurs (XLSTAT)

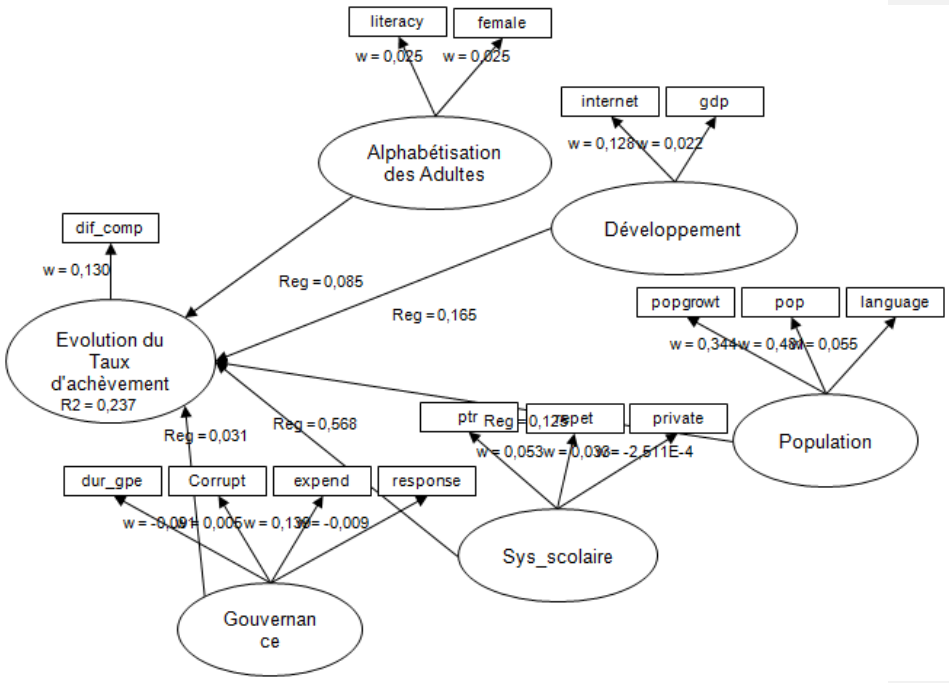
Note : en bleu les effets positif et en rouge les effets négatifs. Effet des groupes par rapport à la catégorie de référence

Nous voyons qu'en plus des variables identifiées préalablement dans le modèle1, la variable groupe a un effet négatif, significatif dans le modèle. La négativité de l'effet est dû au fait que cette variable soit fortement corrélée à l'appartenance aux pays africains francophones, qui ont les moins bons taux d'achèvement. Il y a donc un véritable effet de groupe qui influence les résultats scolaires.

Modèle4 :

Nous avons souligné plus haut qu'il existe deux types de modèles, à savoir les modèles de flux et les modèles de stock. Les modèles de stock servent à mesurer le résultat actuel, c'est-à-dire une quantité actuelle, comme dans notre cas le taux d'achèvement. Le modèle de flux quant à lui sert à mesure l'évolution d'une quantité dans le temps. Après avoir testé de modèle de stock, nous allons donc tester un modèle de flux en utilisant comme variable dépendante l'évolution moyenne du taux d'achèvement de 2004 à 2009.

Graphique26: Modèle4



Fait par les auteurs (XLSTAT)

Nous pouvons avant de passer à la qualité du modèle voir à travers les indices de communauté (tableau) comment sont représentées les variables manifestes par les différentes variables latentes qui leur sont associées :

Tableau33 : Indices de communalités (Modèle4)

Variable latente	Variables manifestes	Communalités
Evolution du Taux d'achèvement	dif_comp	
Alphabétisation	literacy	0,892
	female	0,914
Développement	internet	1,000
	gdp	0,360
Population	popgrowt	0,244
	pop	0,873
	language	0,111
Sys_Scolaire	ptr	0,959
	repet	0,540
	private	0,000
Gouvernance	dur_gpe	0,054
	Corrupt	0,017
	expend	0,892
	response	0,000

Fait par les auteurs (XLSTAT)

Nous pouvons voir que pour les variables alphabétisation, développement, population et système scolaire, nous avons des corrélations semblables au modèle1. Au niveau de la variable gouvernance, la variable *dur_gpe* n'est plus corrélée tandis que la variable *expend* devient très corrélée avec la gouvernance.

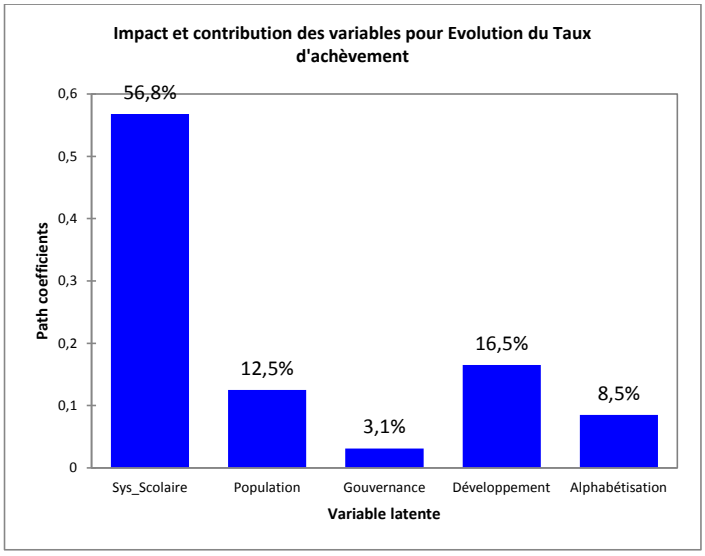
Comme nous pouvons le voir sur le diagramme, le R^2 de ce modèle est de 0,237. Il est relativement faible par rapport aux modèles observés plus haut. Il n'en est pas moins de son indice Gof qui est de 0,341. Nous pouvons voir les variables qui influent sur l'évolution du taux d'achèvement :

Tableau34 : Effet des variables latentes (Modèle4)

Variable latente	Valeur	Ecart-type	t	Pr > |t|
Alphabétisation	0,085	0,135	0,630	0,530
Développement	0,165	0,121	1,365	0,176
Population	0,125	0,106	1,177	0,242
Sys_Scolaire	0,568	0,139	4,081	0,000
Gouvernance	0,031	0,103	0,301	0,764

Fait par les auteurs (XLSTAT)

Graphique27 : Effet des variables latentes (Modèle4)



Fait par les auteurs (XLSTAT)

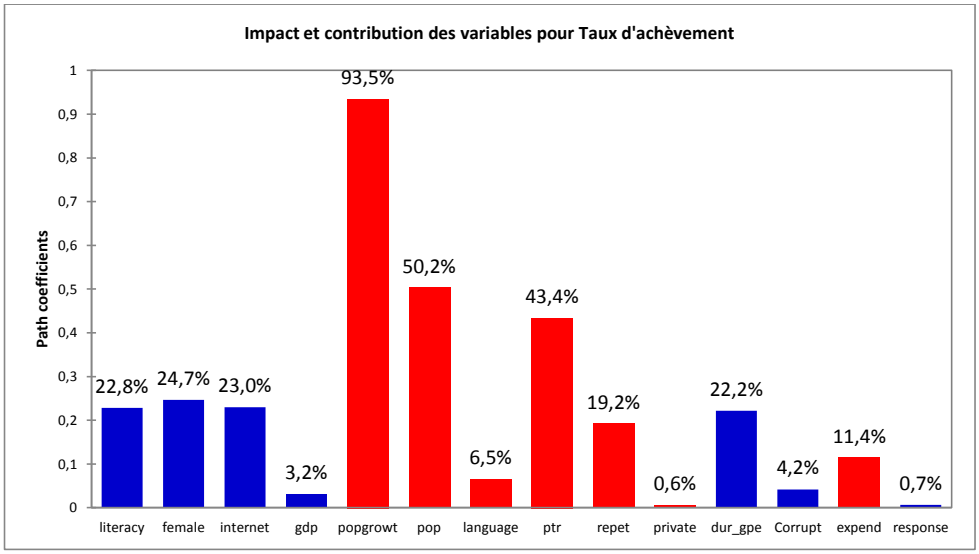
Nous voyons que la seule variable latente qui a un effet positif très significatif sur l'évolution du taux d'achèvement est le système scolaire. En effet, cela peut être un effet de la croissance de la population qui fait accroître à la fois le taux d'achèvement, et également augmente la pression sur les systèmes scolaires. Aussi, dans la base de données, les pays ayant de forts taux de croissance sont les pays de l'Afrique francophone qui ont également de forts taux de croissance. L'effet de cette variable est donc un effet de l'Afrique francophone.

On notera que ces trois indicateurs constituent la base d'un outil de monitoring du Partenariat Mondial pour l'Education, appelé Cadre indicatif Fast Track qui a servi à élaborer des plans de développement de l'éducation et des budgets au début des années 2000.

Nous retenons donc le meilleur modèle plus simple avec le taux d'achèvement comme variable dépendante (modèle1).

Grâce aux poids externes, nous pouvons revenir à notre modèle de départ et voir les variables qui influencent le taux d'achèvement des élèves au primaire :

Graphique28 : Effet des variables (Modèle1)



Fait par les auteurs (XLSTAT)

Conformément à la littérature, nous voyons que le taux d'alphabétisation, et le revenu par habitant sont des facteurs qui favorisent la réussite des élèves au primaire tandis que la taille de classe, le taux de répétition et les dépenses ont un impact négatif sur le taux d'achèvement du primaire. Le taux d'écoles privées a également un effet négatif sur le taux d'achèvement, cela est surement dû au fait que dans certains privés n'étant pas contrôlés par l'Etat ne dispensent pas forcément une formation de qualité. Cela peut indiquer que les pays où l'offre de privé est forte, le gouvernement à répondre à la demande sociale.

En plus des variables observées dans la littérature, nous remarquons qu'un fort taux d'internaute a un grand impact sur le taux d'achèvement, de même qu'un pourcentage élevé de femmes enseignants. Le Partenariat Mondial pour l'Education a également un impact positif sur les résultats des élèves. Le taux de corruption a un effet non significatif sur le taux d'achèvement. D'autre part, une grande population ou un fort taux de croissance est un énorme frein à l'alphabétisation des enfants. En effet, le défi de la plus des pays en développement, notamment les pays africains, est de redoubler chaque année de ressources afin de répondre d'abord à la demande existante avant d'en mettre à la disposition des sphères en marge de l'éducation. La diversité linguistique a également un impact négatif sur le taux d'achèvement. Ce constat devrait plus encourager les gouvernements à promouvoir l'éducation dans les langues nationales mais également à mettre en œuvre des politiques de maîtrise de la démographie. L'atteinte de la scolarisation primaire universelle dans de nombreux pays d'Asie Centrale, ex Républiques de l'Union soviétique, est à mettre en relation avec la faible natalité et une population vieillisante dans ces pays.

La contribution très élevée de la croissance démographique nous montre que nos hypothèses de facteurs contextuels exerçant un fort impact sur les taux de scolarisation sont

vérifiées. Il en est de même, mais dans une moindre mesure pour la part des femmes enseignantes, l'accès à Internet et la participation au Partenariat mondial pour l'éducation, variables qui n'étaient pas bien traitées par la littérature jusqu'à présent.

➔ **A cette étape, l'objectif principal du projet de fin d'études qui était de trouver un modèle explicatif de la qualité de l'éducation a été atteint.**

Nous allons voir maintenant en dehors de l'explication des différences de qualité de l'éducation observées, en quoi nos modèles permettent de prévoir les taux d'achèvement, à conditions économiques et sociopolitiques données.

5-3) Prédictions

Les modèles faits précédemment montrent de part leurs indices de qualité leur capacité à prédire le taux d'achèvement. Ainsi, pour les moindres carrés ordinaires, nous avons donc le tableau de prédiction suivant :

Tableau35 : Valeurs observées et prédites (MCO) du taux d'achèvement

Abr	Taux d'axchèvement observé	Taux d'axchèvement Prévu	Borne inf 95%	Borne sup95%	Différence Prévu--Observé (en valeur absolue)
VUT	82	82	78	86	**0**
ZAF	91	91	86	96	**0**
KGZ	95	95	90	100	**0**
WSM	97	97	92	101	**0**
NPL	59	58	53	63	**1**
LSO	70	71	65	77	**1**
HND	85	84	82	87	**1**
SLV	87	86	81	90	**1**
PHL	93	92	88	97	**1**
PRY	94	93	90	96	**1**
JOR	96	97	93	100	**1**
UZB	96	97	92	102	**1**
BDI	42	44	36	52	**2**
ETH	47	49	43	56	**2**
CMR	61	62	58	67	**2**
TZA	74	72	66	77	**2**
GHA	75	73	70	76	**2**
NAM	84	82	78	87	**2**
MDA	95	97	93	102	**2**
BWA	96	94	89	99	**2**
UKR	100	102	98	106	**2**
MDV	100	98	94	101	**2**

BGD	58	61	58	64	**3**
LAO	72	69	65	73	**3**
GTM	75	78	75	81	**3**
MWI	56	52	44	61	**4**
GIN	58	54	49	59	**4**
MUS	92	96	92	101	**4**
ARM	96	100	96	104	**4**
TJK	98	95	89	100	**4**
TON	100	96	92	100	**4**
CAF	32	37	28	46	**5**
ZAR	53	58	50	65	**5**
LBR	61	65	57	74	**5**
NIC	74	78	76	81	**5**
BTN	78	73	67	78	**5**
WBG	88	93	89	97	**5**
MNG	100	95	91	100	**5**
LKA	100	95	92	98	**5**
BOL	100	94	91	98	**5**
VNM	100	94	90	98	**6**
THA	100	94	90	98	**6**
IDN	100	94	90	97	**6**
TCD	32	39	33	46	**7**
UGA	58	65	62	69	**7**
NGA	80	73	70	77	**7**
ALB	92	99	96	103	**7**
MDG	63	56	51	60	**8**
GAB	70	78	70	87	**8**
EGY	93	85	80	89	**8**
GEO	94	102	97	107	**8**
FJI	97	88	85	92	**8**
MOZ	48	56	51	62	**9**
BEN	60	52	47	56	**9**
GMB	73	64	59	69	**9**
KEN	89	80	75	85	**9**
CPV	90	81	77	84	**9**
SEN	52	62	57	66	**10**
MLI	53	43	37	49	**10**
PAK	61	71	66	75	**10**
AZE	93	104	100	108	**10**

TUN	97	88	83	92	**10**
ECU	105	95	91	99	**10**
BFA	36	48	41	54	**11**
CIV	45	56	51	60	**11**
RWA	47	58	52	64	**11**
SLE	60	49	43	54	**11**
IRQ	70	81	76	85	**11**
COG	75	64	57	71	**11**
MAR	81	70	65	75	**11**
ERI	48	60	55	64	**12**
COM	72	60	53	67	**12**
LVA	94	106	101	111	**12**
SYR	100	88	85	92	**12**
TGO	66	53	47	59	**13**
KHM	82	70	66	74	**13**
IND	89	75	71	80	**13**
CUB	93	106	101	110	**13**
GUY	100	87	84	90	**13**
YEM	61	74	70	79	**14**
SWZ	66	80	75	84	**14**
MRT	53	69	65	72	**15**
BLZ	100	85	81	89	**15**
NER	35	53	46	60	**17**
AGO	47	64	57	70	**17**
ZMB	85	68	61	74	**17**
KIR	100	82	79	84	**18**
TMP	80	55	50	60	**25**
SDN	52	78	75	81	**26**
DJI	35	78	75	80	**42**
	76	**76**	**72**	**81**	**8**

Fait par les auteurs (XLSTAT)

Ce tableau montre que dans plus de deux tiers des pays (63 sur 90), l'écart entre taux d'achèvement prévu et observé n'excède pas 10 points en valeur absolue, montrant le très bon caractère prédictif du modèle. Dans plus d'un tiers des pays (41 sur 90), l'écart n'excède pas cinq points.

Un tel tableau peut également être très utile au Partenariat Mondial pour l'Education et servir d'outil de discussion pour établir un objectif de taux d'achèvement qui tienne compte des paramètres sociodémographiques propre à chaque pays et affiner les outils de simulation existants.

Nous pouvons aussi voir les prédictions faites par les moindres carrés partiels :

Tableau 36: Valeurs du taux d'achèvement prédites par les moindres carrés partiels

PAYS	abr	Taux d'achèvement Observé	Taux d'achèvement Prévu	Différence prévu - observé (en valeur absolue)
Cambodia	KHM	82	83	0
South Africa	ZAF	91	91	0
Ghana	GHA	75	75	0
Burkina Faso	BFA	36	36	1
Timor-Leste	TMP	80	79	1
Gambia, The	GMB	73	71	2
Cote d'Ivoire	CIV	45	43	2
Tunisia	TUN	97	99	2
Syrian Arab Republic	SYR	100	102	2
Namibia	NAM	84	87	2
Rwanda	RWA	47	45	2
Mauritania	MRT	53	51	2
Botswana	BWA	96	94	2
Burundi	BDI	42	45	3
Indonesia	IDN	100	103	3
Malawi	MWI	56	53	3
Egypt, Arab Rep,	EGY	93	96	3
Sierra Leone	SLE	60	56	4
Tanzania	TZA	74	69	5
Congo, Dem, Rep,	ZAR	53	48	5
Mauritius	MUS	92	97	5
Senegal	SEN	52	46	6
Angola	AGO	47	53	6
India	IND	89	83	6
Nigeria	NGA	80	74	6
Thailand	THA	100	107	7
Niger	NER	35	43	7
Ethiopia	ETH	47	55	7
Chad	TCD	32	25	8
Uganda	UGA	58	67	8
Sri Lanka	LKA	100	108	8
Morocco	MAR	81	89	9
Mozambique	MOZ	48	38	10
Bhutan	BTN	78	88	10
Belize	BLZ	100	111	11
Gabon	GAB	70	59	11
Jordan	JOR	96	107	11
Cameroon	CMR	61	50	11
Liberia	LBR	61	72	12
Central African Republic	CAF	32	20	12
Bolivia	BOL	100	112	12

Philippines	PHL	93	105	12
Kenya	KEN	89	76	13
Madagascar	MDG	63	51	13
Mongolia	MNG	100	113	13
Ecuador	ECU	100	113	13
West Bank and Gaza	WBG	88	102	14
Vietnam	VNM	100	115	15
Guinea	GIN	58	43	15
Fiji	FJI	97	112	15
Kiribati	KIR	100	116	16
Lesotho	LSO	70	87	17
Cape Verde	CPV	90	72	17
Bangladesh	BGD	58	76	18
Lao PDR	LAO	72	91	18
Zambia	ZMB	85	67	18
Paraguay	PRY	94	113	19
Nepal	NPL	59	78	19
Mali	MLI	53	33	20
El Salvador	SLV	87	106	20
Djibouti	DJI	35	55	20
Eritrea	ERI	48	68	21
Comoros	COM	72	51	21
Tajikistan	TJK	98	120	22
Tonga	TON	100	122	22
Pakistan	PAK	61	82	22
Guyana	GUY	100	122	22
Yemen, Rep,	YEM	61	83	22
Swaziland	SWZ	66	88	22
Benin	BEN	60	37	23
Samoa	WSM	97	120	24
Togo	TGO	66	42	24
Honduras	HND	85	110	25
Maldives	MDV	100	125	25
Iraq	IRQ	70	98	28
Guatemala	GTM	75	103	28
Vanuatu	VUT	82	110	28
Uzbekistan	UZB	96	125	29
Congo, Rep,	COG	75	46	30
Ukraine	UKR	100	132	32
Armenia	ARM	96	129	33
Cuba	CUB	93	126	33
Albania	ALB	92	126	34
Nicaragua	NIC	74	109	35
Moldova	MDA	95	131	36
Kyrgyz Republic	KGZ	95	131	36

Azerbaijan	AZE	93	133	39
Sudan	SDN	52	91	39
Georgia	GEO	94	134	41
Latvia	LVA	94	142	48
Moyenne		76	85	16

Fait par les auteurs (XLSTAT)

Nous remarquons que les différences sont plus grandes que celle observées avec les moindres carrés ordinaires. Les pays qui ont de grandes différences entre le taux d'achèvement prévu et le taux d'achèvement observé sont les pays extrêmes de l'analyse, c'est-à-dire les pays de l'Afrique francophone et les pays de l'ex bloc soviétique. Les modèles PLS ont tendance à surestimer le taux d'achèvement car le taux prévu dépasse 100% dans de nombreux pays.

Nous remarquons que le taux d'achèvement prédit par les moindres carrés ordinaires du Burundi est de 43,9 ; celui prédit par les moindres carrés partiels est de 44,6 et la valeur observée est de 41,9. Le modèle macroéconomique semble donc bien s'accorder aux valeurs observées. Toutes fois, une l'étude macroéconomique seule ne permet pas d'expliquer entièrement le taux d'achèvement. En effet nous pouvons tenir compte des variétés entre pays pour expliquer le taux d'achèvement, mais chaque a ces réalités, au sein de ces régions et de chaque école. Il est donc indispensable de faire une étude microéconomique afin de déterminer d'autres facteurs de l'augmentation du taux d'achèvement et des compétences des élèves.

2ème Partie : Analyse Microéconomique des données du Burundi

I- Construction de la base de données et des indicateurs

1-1) Revue de la littérature :

Plusieurs études ont été également faites du point de vue microéconomique sur l'analyse des facteurs qui influencent les acquis scolaires. Nous avons retenu ici trois études qui traitent respectivement de la mesure du niveau de vie, de la comparaison des effets de contexte avec les variables de politiques éducatives et de la mesure des comportements en classe des enseignants.

- **Mesure du niveau de vie :**

FILMER & PRITCHETT (2001) ont testés la validité d'une analyse en composantes principales qui estime la relation entre richesse des ménages et la scolarisation des enfants par un indice de biens, en absence de données sur les revenus ou sur les dépenses des ménages. Pour établir la validité de cette approche, les auteurs ont comparé la classification des ménages obtenue à partir de l'indice de biens pondérés et la classification des ménages à partir des dépenses de consommation. L'étude se sert des données des sondages DHS[16]/NFH. Le sondage NFHS fournit des informations sur la population, la santé et la nutrition dans les

[16] National Family Health Survey

29 états de l'Inde. Ce sondage basé sur un échantillon de ménage représentatif au niveau national mais aussi des états.

Ils ont trouvé que l'indice de biens, obtenu grâce à l'analyse en composante principale est une bonne solution pour estimer la richesse des ménages en cas d'absence de données sur les revenus. Cependant, le classement obtenu à partir de l'indice de biens classe les ménages ruraux comme étant plus pauvres que ne le fait la classification à partir des dépenses de consommation.

Les résultats fournis par l'analyse en composantes principales montrent de grandes différences de scolarisation en fonction de la richesse des ménages qui varie considérablement entre les Etats de l'Inde. En moyenne, un enfant riche a 31% de plus qu'un enfant pauvre d'être scolarisé. Mais cet écart est seulement de 4,6 points de pourcentage au Kerala tandis qu'il est de 38% dans l'Uttar Pradesh et de 42,6 dans le Bihar.

Ces méthodes ont été utilisées depuis de manière systématique par l'UNICEF mais aussi les programmes de mesure des acquis comme le PASEC et SACMEQ pour construire une échelle de niveau de vie des ménages ou des élèves.

- Comparaison des effets de contexte et les variables de politiques éducatives :

Dans leur article '*Differences in pupil achievement in Kenya: Implication for policy and practice, Internal Journal of Educational Development*' Njora Hungi et Florence W. Thuku ont quant à eux comparé les effets de contexte avec les variables de politiques éducatives. Cette étude a été menée dans le cadre du projet SACMEQ[17] relatif à la qualité de l'enseignement dispensé dans les écoles primaires au Kenya et dans 13 autres pays africains. Le but est d'identifier les facteurs de progrès des acquis scolaires à trois niveaux, élèves, classes, et école, qui permettent d'expliquer les différences de niveau scolaire des élèves en 6ème année du primaire au Kenya de manière à orienter les politiques éducatives publiques.

Les données ont été recueillies auprès de 3299 élèves dans 320 classes dans 185 écoles dans huit provinces du Kenya en 2002. La méthode d'échantillonnage, la construction des questionnaires et la construction des tests en lecture et mathématiques sont semblables à celles utilisées dans le PISA[18], en plus du fait que les enseignants répondent aussi à des tests, qui ont des items communs avec les tests des élèves de manière à rendre la comparaison possible. Les tests élèves ont été étudiés pour garantir leur adéquation aux différents programmes enseignés dans les régions du Kenya.

Deux modèles multiniveaux (élèves, classe, école) ont été testés, l'un avec comme variable réponse les résultats au test de mathématique et l'autre le score élève en lecture. Les variables explicatives ont été choisies sur la base d'études pionnières menées dans différents pays en développement. Les auteurs ont donc construit des modèles en commençant par régresser uniquement les variables relatives aux élèves sur la variable endogène, ensuite seulement les variables relatives à la classe et enfin les variables relatives à l'école.

[17] Southern an Eastern Africa Consortium for Monitoring Educational Quality.
[18] Programme for International Student Assessment.

L'influence de ces variables est plus ou moins forte selon la région étudiée.

Le modèle de régression multiniveaux a permis de vérifier l'hypothèse faite sur l'influence positive de la richesse des ménages sur les progrès des acquis scolaires mais celle-ci a moins d'impact que les indicateurs de qualité des écoles et d'autres variables. Ce résultat est cohérent avec les résultats d'études antérieures dans les pays en développement qui concluent que les caractéristiques socioéconomiques ont peu d'influence sur le réussite scolaire des élèves comparativement aux facteurs de niveau classe et école (Heyneman et Loxley,1983 ; Gamoran et Long, 2006).

Le partitionnement de la variance a permis de mettre en exergue que des différences importantes en terme de réussite scolaire des élèves de 6ème année existent entre école, mais que les différences entre classes d'une même école sont négligeables. Mais de manière générale, la variance totale expliquée par les variables de classe et d'école est beaucoup plus importante que la variance expliquée par les variables propres aux élèves. Les facteurs <<scolaires>>, gestion de l'école et pratiques pédagogiques des enseignants seraient des facteurs clés de la réussite scolaire. La variance restant en grande partie inexpliquée (65,8% et 59% pour les mathématiques et la lecture, respectivement), les conclusions des modèles développés dans cette étude sont à relativiser car il semblerait qu'il y ait des facteurs influençant la réussite scolaire des élèves, non inclus dans ces modèles. Cette étude a également révélé une inégalité entre les sexes plus importante au Nord en termes de réussite scolaire et des disparités entre établissements à l'est et dans la vallée du Rift.

Sur la base des variables explicatives de la réussite scolaire au Kenya mise en évidence dans cette étude, les auteurs invitent le ministère de l'éducation à mettre en place des politiques éducatives dont ils font quelques propositions et à renforcer des politiques déjà existantes mais non suivies comme l'obligation de scolarisation à partir de l'âge de 12 ans.

- Mesure des comportements en classe des enseignants

Enfin, dans leur article '*Interpersonal Teacher Behaviour and Student Outcomes, School Effectiveness and School Improvement*', Brok & Brekelmans & Wubbels, (2004) ont traité des comportements des enseignants en classe. Leur objectif était de proposer un modèle et des techniques d'analyses de l'influence du comportement des enseignants sur la réussite scolaire (résultats cognitifs et affectifs), en combinant les concepts, les méthodes et les instruments déjà élaborés dans les travaux de recherche précédant sur le sujet. Ils vont donc procéder, dans un premier temps à l'évaluation du comportement des enseignants par les élèves, pour mettre en exergue l'importance de la perception du profil de l'enseignant par les élèves. Ces données sont recueillies auprès des élèves de 3ème année de collège. Ensuite ils vont déterminer le profil de l'enseignement au travers des points de vue sur la relation enseignant/élèves. Cette appréciation peut être du point de vue organisationnel (place de l'enseignant dans l'administration de l'école), du point de vue moral (valeurs transmises), du point de vue du type d'activités ou du point de vue contenu.

Les auteurs proposent un modèle appelé modèle circonflexe, qui a des propriétés statistiques particulières pour analyser le comportement des enseignants dans une perspective

relationnelle. Deux dimensions sont étudiées (8 profils d'enseignants agrégés), le premier est relatif au caractère plus ou moins autoritaire de l'enseignant, le second est relatif au caractère plus ou moins coopératif de ce dernier.

Après avoir répertorié les limites des études précédentes qui ont toutes pour effet de surestimer l'influence du profil relationnel de l'enseignant sur les résultats des élèves, notamment les limites des techniques d'analyses de la variance couramment utilisées (ANOVA et analyse multiple), les auteurs ont préconisé l'utilisation d'une analyse multiniveaux (élèves, classe) de la variance et l'introduction dans le modèle de profils relationnels enseignants/élèves qui soient indépendants, de manière à résoudre le problème de surestimation.

Les résultats de cette étude sont : premièrement que la perception qu'ont les élèves du caractère autoritaire ou coopératif de leur enseignant n'a aucun effet sur leurs résultats affectifs, c'est-à-dire sur leur motivation. Ensuite, la perception qu'ont les élèves du caractère autoritaire de leur enseignant est un déterminant de leurs résultats cognitifs en physique mais n'a aucun effet sur les résultats en anglais langue étrangère. Enfin, la perception qu'ont les élèves du caractère coopératif de leur enseignant n'a d'effet ni sur les résultats cognitifs en anglais, ni sur ceux en physique. Ceci peut être expliqué par la nature des cours comparés, car il existe plus d'opportunités de pratiquer l'anglais en dehors de la classe que d'opportunités de faire de la physique, ce qui prouve la nécessité de bien mesurer le contexte extra-scolaire dans les recherches en éducation.

Cette étude admet plusieurs limites, tel que la comparaison entre échantillons <<physique>> et <<anglais>> qui est fragile car les deux échantillons sont indépendants, étant composés d'écoles différentes, de classe différentes et donc d'élèves différents. Aussi, le fait qu'il ait considéré uniquement la dimension <<relation enseignant/élève>> favorise une surestimation de l'effet du relationnel sur les résultats cognitifs des élèves. Enfin, le nombre d'enseignants sur lequel porte l'étude, les niveaux d'analyse (élèves et enseignant/classe) et le nombre de variables de contrôle introduites dans le modèle ne sont pas très élevés.

Plusieurs études ont été menées récemment au Burundi et nous permettent d'avoir une idée des facteurs d'amélioration des acquis scolaires propres à ces pays. Il s'agit :

- D'une évaluation PASEC menée en 2009 sur 180 écoles et testant les compétences en kirundi (la langue du pays), français et mathématiques
- Une évaluation nationale menée en 2010 sur un échantillon d'élèves de 3ème année, reprenant les tests PASEC
- Une évaluation dite EGRA des compétences en lecture réalisée sur 120 écoles et 1800 élèves de 2ème année
- Enfin une évaluation EGRA menée en 2012 sur 90 écoles et 1800 élèves de 2ème et 3ème année reprenant les outils 2011 et dont les données seront analysées dans le cadre de ce rapport.

Nous allons présenter tout d'abord les résultats des deux premières études qui reposent sur les mêmes instruments (tests).

1-2) **Note sur les évaluations nationale et PASEC du Burundi**

L'évaluation nationale a été faite suite à la mise en place d'une commission d'évaluation du système éducatif Burundais, en juillet 2010, pour promouvoir un système d'évaluation capable d'assurer une gestion locale de la qualité, au service des corps d'encadrement et des équipes enseignantes. Elle sera également susceptibles d'interroger et de faire évoluer aussi bien les pratiques de dotation en ressources humaines et financières que les pratiques d'animation et de gestion pédagogique au niveau des circonscriptions et des établissements.

1-2-1) caractéristiques et conditions de vie des élèves

Le modèle montre, pour ce qui est des caractéristiques individuelles, que le sexe de l'élève ne semble pas être lié à un effet significatif. Pour sa part l'évaluation PASEC montre qu'en 5ème année, les résultats des filles ont diminués de 5,6% par rapport à ceux des garçons (Rapport de l'évaluation diagnostique 3ème Année 2010/2011 Page 39), ce qui peut aussi être expliqué par le travail domestique demandé aux élèves filles. L'âge avancé de l'élève est associé à un impact négatif et significatif sur les performances des élèves. Il est utile de signaler que 68,7% des élèves testés ont 10 ans et plus et qu'à travers une moyenne d'âge de 10,7 ans pour les filles et de 11 ans pour les garçons en 3ème année (cf document indicateurs 2009/2010 Bureau de la Planification), au lieu d'un âge normal de 9 ans en fonction d'un début de scolarité primaire à 7 ans, on peut considérer les effets directs du redoublement important au Burundi. L'effet de la gratuité en 2006 peut cependant encore expliquer un accès tardif de certains élèves entrés en 2008.

Pour ce qui est des facteurs scolaires, l'influence semble plus visible. Le taux de redoublement des élèves fait partit des plus élevés de l'Afrique. Dans l'enquête réalisée, on peut constater que 62,4% des élèves testés (à travers un choix aléatoire) ont au moins redoublé une fois, alors qu'ils ne sont qu'en début de troisième année ; ceci a un effet négatif très significatif sur les scores agrégés des élèves. Ce fort taux de redoublement est agréé par les enseignants (60% d'après le PASEC 2008) et les parents car ils disent que cela permet à l'élève d'être mieux formé. Même si on pouvait considérer certains effets positifs du redoublement dans les premières années, son mauvais rapport coût/efficacité doit faire reconsidérer les stratégies en faveur de la qualité. L'absentéisme de l'élève a clairement un impact négatif et fortement significatif. En effet, l'élève qui s'absente accuse un retard de 2.6 point comparativement à l'élève assidu (l'absentéisme a été considéré dans l'enquête après un retard de deux semaines). L'analyse révèle que 42.3% des élèves ayant participé à l'évaluation ont eu ce taux d'absentéisme. La fréquentation du préscolaire n'a pas d'effet significatif.

Regardons à présent l'effet de l'environnement socio-économique. Le milieu rural a un effet positif sur les disciplines du Kirundi et des Mathématiques. Cependant, il faut noter que le milieu rural ou urbain peut être difficile à définir au Burundi, et si l'on considère que 90% des écoles se trouvent en milieu rural, l'effet positif significatif de ce milieu peut correspondre à un effet de nombre. Parlant de l'alimentation de l'élève, 50,8% des élèves testés prennent un petit déjeuner le matin et 92,1% prennent un déjeuner. Le modèle, comme au PASEC, indique qu'il y a un effet positif significatif du déjeuner sur les acquis scolaires. Pour sa part, le petit

déjeuner ne semble pas être lié à un impact significatif. En milieu rural, les enfants sont amenés à effectuer des travaux domestiques, avant ou après la classe. Bien que le modèle ne reflète pas un impact significatif, nous pensons que le cumul des travaux avant et après la classe, associé souvent à la non prise d'un petit déjeuner, permettra difficilement à ces enfants de réaliser des performances scolaires acceptables. L'effet de l'encadrement à la maison, apporté par les frères ou les parents, est positif mais non significatif. 26% des foyers des élèves testés disposent d'électricité. Ce facteur a un effet positif significatif sur les résultats des élèves. On peut l'associer au milieu urbain et au niveau de vie des familles.

1-2-2) Caractéristiques des enseignants et de l'école

Cinq variables sont associés à un impact significatif positif ou négatif sur les acquisitions des élèves: L'ancienneté du directeur (négatif), l'ancienneté de l'équipe (positif), le fait que l'enseignant vive sur place ou en famille (positif), le fait qu'il ait réalisé 75% du programme (positif).

Quant au genre de l'enseignant, le modèle indique un effet négatif de la présence des femmes enseignantes par rapport à celle des hommes sur les scores des élèves (non significatif cependant). Au niveau de l'impact par discipline, par contre, il y a un effet positif significatif des enseignants femmes sur le score en Français, que l'on retrouve d'ailleurs dans les analyses du PASEC (plus important de 9.5% par rapport à celui d'un enseignant de sexe masculin). Si l'on veut tenir compte de la tendance négative exprimée, elle pourrait éventuellement s'expliquer par un plus fort taux d'absentéisme féminin, dû aux congés de maternité. Le modèle indique un effet négatif de l'absentéisme de l'enseignant sur les scores des élèves, qui n'est cependant pas statistiquement significatif par rapport à l'ensemble des variables considérées. Mais le taux d'absentéisme des enseignants n'est pas négligeable, 27% se sont absentés entre 2 et 4 semaines l'an dernier et 8% ont eu une absence supérieure à 4 semaines.

L'ancienneté de l'équipe pédagogique a un effet positif significatif. 67% des enseignants ont entre 5 et 15 ans d'ancienneté. Cela indique une stabilité des équipes pédagogiques qui pourrait être utilement mise à profit dans l'intérêt des élèves. De la même manière que pour l'ancienneté pédagogique, le fait que l'enseignant vive sur place a un effet positif très significatif. 53% des enseignants des élèves testés vivent en effet sur place. Ce facteur est en effet très important dans le milieu rural où les enseignants ont généralement de longs déplacements à effectuer pour se rendre en cours.

Pour ce qui est de l'ancienneté du directeur, la tranche d'ancienneté 5-15 ans des enseignants, qui est bénéfique pour les élèves est négative quand il s'agit des directeurs d'école. Par contre, sur la tranche des jeunes directeurs (de 1 à 5 ans), on peut noter un effet positif. Il semble donc que la durée n'améliore pas les compétences de gestion pédagogique des directeurs. L'analyse du PASEC indique qu'il n'y a pas d'effet de l'ancienneté du directeur sur les scores des élèves. Concernant la couverture du programme, 96% des enseignants de la classe précédente (2ème année) ont effectué une couverture d'au moins 75% du programme. Cette couverture a un effet positif significatif que l'on retrouve dans l'analyse PASEC, avec un taux de couverture similaire. La visite des autorités scolaires et les réunions des directeurs avec l'ensemble de l'équipe pédagogique ne semblent pas avoir un effet significatif sur le score des élèves.

1-3) Modèle du rapport EGRA

Le manuel pour l'évaluation des compétences fondamentales en lecture- EGRA (Sprenger-Charolles, 2009) dit que : « *la capacité de lecture et de compréhension d'un simple texte est l'une des compétences les plus fondamentales qu'un enfant puisse acquérir*».

Les outils EGRA (Early Grades Reading Assessment ou d'évaluation des compétences fondamentales en lecture) mis au point par la société RTI seront les instruments de mesure des apprentissages des élèves au Burundi. Les tests EGRA sont des tests critériés permettant de savoir si un enfant possède les compétences ou non. Ils reposent sur le principe de la fluidité ou vitesse de lecture. Un score est donc attribué à chaque élève en fonction du nombre de mots lus par minute.

Mazunya & Varly (2011) ont donc testé l'influence des variables sur les scores obtenus. Ils ont vu que l'âge de l'élève, l'absentéisme de l'élève, le redoublement et l'ancienneté du directeur ont un effet négatif sur les résultats des scores des élèves.

D'autre part, l'ancienneté de l'équipe, le fait que l'enseignant vive sur place ou en famille, le fait qu'il ait réalisé 75% du programme, le fait que l'élève prenne le déjeuné, qu'il ait de l'électricité et l'appartenance au milieu rural ont un effet négatif sur ces mêmes résultats. Les travaux domestiques avant et après la classe sont associés à un effet négatif mais non significatif tandis que l'encadrement à la maison a un effet positif mais non significatif.
Nous pouvons résumer toutes ces analyses par les tableaux ci-dessous. Le premier tableau nous montre l'impact des caractéristiques de l'élève sur leurs résultats selon chaque auteur :

Tableau37 : Effet des variables testées dans la littérature (caractéristiques de l'élève)

Variables	Auteurs	Effet
Genre de l'élève	PASEC	NS
	Evaluation nationale	NS
Age	PASEC	+
	Evaluation nationale	-
	EGRA	+
Redoublement	PASEC	-
	Evaluation nationale	-
	EGRA	+
Absentéisme de l'élève	PASEC	-
	Evaluation nationale	-
	EGRA	+
Fréquentation du préscolaire	Evaluation nationale	NS
Milieu rural	Evaluation nationale	+
	EGRA	-
Bonne alimentation de l'élève (petit déjeûner)	PASEC	+
	Evaluation nationale	+
	EGRA	-
Travaux domestiques	Evaluation nationale	NS
	EGRA	NS

Encadrement à la maison	Evaluation nationale	NS
	EGRA	NS
Eclairage du foyer	Evaluation nationale	+
	EGRA	-

Fait par les auteurs à partir de la revue de la littérature

Le tableau suivant nous montre l'impact des conditions de l'école et des enseignants sur les résultats des élèves :

Tableau38 : Effet des variables testées dans la littérature (Caractéristiques de l'école et de l'enseignant)

Variables	Auteurs	Effet
Ancienneté du Directeur	PASEC	NS
	Evaluation nationale	+ (1-5 ans) - (+5ans)
	EGRA	+
Couverture du programme (75%)	PASEC	+
	Evaluation nationale	+
	EGRA	-
Visite des autorités	Evaluation nationale	NS
Enseignant femme	PASEC	+
	Evaluation nationale	-
Absentéisme de l'enseignant	Evaluation nationale	NS
Ancienneté de l'équipe pédagogique	Evaluation nationale	+
	EGRA	-
Enseignant vivant sur place ou en famille	Evaluation nationale	+
	EGRA	-

Fait par les auteurs à partir de la revue de la littérature

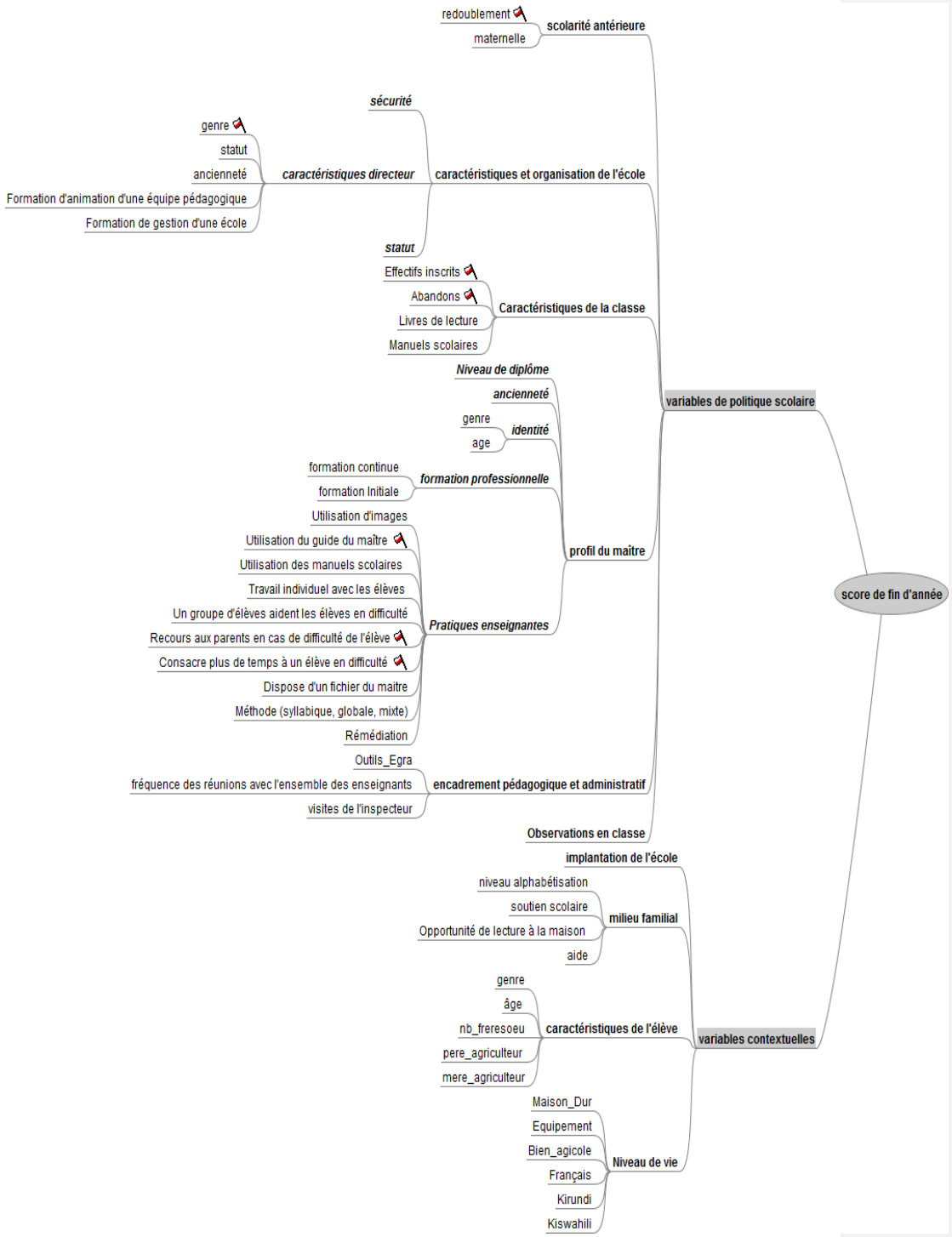
1-4) Variables ajoutées et calcul des indicateurs

Afin d'obtenir les données de l'enquête qui a été menée cette année au Burundi, nous avons aidé à réaliser des masques de saisie pour les opérateurs sur place. Ensuite nous avons eu à vérifié la cohérence des données reçu par des tests logiques ainsi qu'à réaliser des imputations des valeurs manquantes. Le nombre d'écoles est de 90 et le nombre d'élèves de de 1793.

Nous avons tenus compte des résultats aux tests en 2011 et nous sommes partis du modèle fait l'année dernière par une stagiaire dans les locaux de Varlyproject, Richard (2011) auquel nous avons ajouté d'autres variables dont les observations en classe, ce qui est une nouveauté du protocole EGRA 2012.

Les principaux facteurs mesurés par les analyses 2011 sont :

Graphique29 : Variables de l'analyse microéconomique



Fait par les auteurs (Freemind)

Nous avons deux principaux groupes de variables à savoir les variables contextuelles et les variables de politique scolaire.

1-4-1) Les variables contextuelles :

Il s'agit des différentes réalités auxquelles est soumis chaque élève. Ces différentes conditions peuvent avoir un impact sur l'évolution scolaire de l'élève. Nous avons pris en compte le niveau de vie de l'élève, les caractéristiques de l'élève, le milieu familial et l'implantation de l'école.

Le niveau de vie de l'élève regroupe la qualité de sa maison, son équipement, les biens agricoles des parents et les langues parlées à la maison. La qualité de la maison de l'élève (maison en dur) est mesurée ici à travers la qualité de la toiture ; qu'elle soit en dur, semi dur, en tôle ou en tuile. L'équipement quant à lui est une somme de biens d'équipements de la maison, à savoir un robinet, les toilettes, l'électricité, un réfrigérateur, une télévision et un ordinateur. Les biens agricoles quant à eux sont aussi une somme des propriétés agricoles des parents de l'élève à savoir des vaches, chèvres, porcs, autres animaux et des arbres. Français, kirundi et kiswahili sont des variables binaires qui indiquent si ces langues sont parlées par l'élève chez lui.

Les caractéristiques de l'élève regroupent son genre, son âge, le nombre de frères et sœurs qu'il a et la profession des parents. Le milieu familial nous montre la possibilité d'encadrement que peut recevoir l'élève de ces parents. Nous y voyons l'aide à la maison, le niveau d'alphabétisation des parents, le soutien scolaire et les opportunités de lecture à la maison. L'indice d'opportunités de lecture est calculé selon un score explicité dans les deux tableaux suivant :

Tableau39 : Calcul de l'indice d'opportunité de lecture(1)

Var.	LIVR_CLASS	LIVR_MAISON	PAPA_LIT	MAMAN_LIT	LIS_MAISON	LIS_SCOLAIRES	LIVRES_ENFANTS	LIVRES_AUTRES
Pts	2	1	1	1	1	1	2	2

Fait par les auteurs

Puis

Tableau40 : Calcul de l'indice d'opportunité de lecture(2)

Var	LIV_FRANCAIS	LIV_KIRUNDI	LIV_KISWAHILI	LIV_AUTRES	AIDE_PERSO	AIDE_PARENTS	AIDE_FRERES	AIDE_TUTEUR	AIDE_MAITRE	AIDE_REPET	AIDE_AUTRES
Pts	1	2	1	1	0	1	1	1	2	2	1

Fait par les auteurs

Dans le premier tableau, l'élève mentionne s'il a des livres en classe, s'il a des livres dans sa maison. Les variables PAPA_LIT et MAMAN_LIT nous renseignent sur le fait que les parents de l'élève lui font des lectures ou pas. Les variables LIS_SCOLAIRES et LIS_MAISON nous montrent si l'enfant lit à la maison et s'il lit à l'école. Les deux dernières variables nous montrent si l'élève possède des livres pour enfant et s'il en possède d'autres.

Le deuxième tableau, les variables LIV_FRANCAIS, LIV_KIRUNDI,, LIV_KISWAHILI et LIV_AUTRES indiquent si l'élève a des livres dans les différentes langues ou dans d'autres langues à préciser. La variable aide_perso veut dire que l'élève ne reçoit de l'aide de personne. Les variables AIDE_PARENTS, AIDE_FRERES, AIDE_TUTEUR, AIDE_MAITRE, AIDE_REPET et AIDE_AUTRES indiquent si l'élève reçoit respectivement de l'aide de ses parents, de ses frères, de son tuteur, de son maitre, de son répétiteur ou d'autres personnes.

1-4-2) Variables de politiques scolaires

Parmi les variables de politiques scolaires, nous avons la scolarité antérieure de l'élève, les caractéristiques et l'organisation de l'école, les caractéristiques de la classe, le profile du maitre, l'encadrement pédagogique et administratif et les observations en classe.

La scolarité antérieure de l'élève mesure les failles du passé scolaires de l'élève. Nous y avons la variable maternelle qui indique si l'élève a fait la maternelle ou pas et la variable redouble qui indique s'il a déjà redoublé au moins une fois.

Les caractéristiques et l'organisation de l'école : Il s'agit du statut de l'école, si celle-ci est publique ou pas, des caractéristiques du directeur, son genre, son statut (fonctionnaire ou pas), son ancienneté, s'il a suivi une formation d'animation et (ou) de gestion d'une équipe pédagogique ou pas, et de l'insécurité dans l'école. Ce dernier indice (insécurité) est une somme de variables binaires comme l'indique le tableau suivant :

Tableau41 : Calcul de l'indice d'insécurité

VOLS	DEGRADATIONS	AGRESSIONS ECOLE	AGRESSIONS HORS
1	1	1	1

Fait par les auteurs

Ces variables indiquent respectivement s'il y a eu au cours de l'année scolaire des vols, des dégradations, des agressions dans l'école et des agressions hors de l'école.

Les caractéristiques de la classe renferment la taille de la classe, c'est-à-dire l'effectif des inscrits, le nombre d'abandons, les livres de lectures et les manuels scolaires de la classe.

Nous avons ensuite le profil du maitre qui comprends son identité (le genre et l'âge, le niveau de son diplôme et son ancienneté. Il comprend également les formations professionnelles qu'il a reçues : formation initiale ou formation continue. Ainsi que les pratiques enseignantes. Ces dernières sont diverses, nous avons entre autres l'utilisation du guide du maitre et des manuels scolaires, les techniques de suivit des élèves et de remédiation en cas de retard ou de difficulté d'un élève à assimiler les cours et l'utilisation d'un fichier du maitre.

Pour ce qui est de l'encadrement pédagogique et administratif, nous avons : la fréquence de réunion avec l'ensemble des enseignants, la fréquence des visites de l'inspecteur, et l'utilisation des outils Egra. L'indice d'utilisation des outils Egra mesure si le directeur a participé à l'évaluation Egra et s'il applique les méthodes recommandées.

Nous avons ensuite les observations en classe faites par les enquêteurs. Il s'agit de noter si les élèves font des lectures silencieuses ou individuelles à haute voix, s'ils écrivent sur un support, au tableau ; s'ils écoutent et regardent l'enseignant, s'ils répètent ou récitent, posent des question ou répondent aux questions de l'enseignant. En résumé ces observations veulent mesurer la réactivité des élèves et voir si les cours sont interactifs et encouragent la participation des élèves.

Il s'agira donc dans cette étude de voir ce qu'il en est des variables déjà observées auparavant et tester la pertinence des nouvelles variables introduites.

II- Etude préliminaire

Nous allons dans un premier temps voir les variables susceptibles d'avoir un effet significatif sur la performance en lecture des élèves à travers les corrélations.

Tableau42 : Corrélation des variables avec le nombre de mots lus par minute

Corrélations avec le nombre de mots lus par minutes					
motstext_perminut2011_AGR	0,362	nb_freresoeu_ES	0,041	manuel_kiru	0,037
Niveau	0,393	pere_agriculteur	-0,034	fichier_kirundi	-0,019
Fille	0,112	mere_agriculteur	0,007	mat_innovateur	0,017
age_ES	0,105	dir_femme	0,018	m_globale	0,030
maison_dur	0,057	dir_nouvo	0,026	m_mixte	-0,015
Equipement	0,026	dir_fonct	-0,060	remediation	0,036
Bien_agricol	-0,031	dir_formanim	0,048	kirundi_dif	-0,066
Boutique	0,012	dir_formgest	-0,069	RCOR3	-0,115
Mange	-0,004	nb_visinsp_ES	0,023	RCOR9	-0,031
trav_dom_svt	0,023	freq_reunion_mois	0,048	RCOR10	0,074
Maternelle	0,047	outil_egra	-0,012	RCOR12	0,068
REDOUBL	-0,080	insecurite	0,161	RCOR13	-0,081
Francais	0,070	femme	-0,114	RCOR14	0,016
Kirundi	0,038	DUREE_FI	0,029	RCOR19	0,069
livr_class	0,046	formcont	-0,021	RCOR20	0,028
opport_lect	0,268	presents	-0,078		
aide_perso	-0,148	Txabandon_ES	-0,069		

Fait par les auteurs

Tableau43 : Libellé des observations

Observations en classe	Libellés
RCOR3	L'enseignant parle à un seul élève
RCOR9	L'enseignant écoute les élèves
RCOR10	L'enseignant pose une question
RCOR12	Les élèves lisent en chœur
RCOR13	Elèves_lecture individuelle à haute voix

RCOR14	Elèves_lecture silencieuse
RCOR19	Elèves_pose une question
RCOR20	Elèves_répond à une question

Fait par les auteurs

Les corrélations ici, comme dans la plus part des études microéconomiques ne sont pas aussi fortes que celles des études macroéconomiques, aussi en raison du nombre d'observation. Nous voyons notamment que le niveau de la classe et les résultats de l'année passée sont les variables les plus corrélées (positivement) avec le nombre de mots lus par minutes. Ensuite nous avons l'indice d'opportunités de lecture qui est positivement corrélé au nombre de mots lus par minute. Le genre féminin et l'âge sont positivement corrélés avec les performances des élèves en lecture. Les biens ou la richesse des parents d'élèves sont ici très peu corrélés avec leur performance en lecture, mais nous pouvons voir que l'équipement en biens « moderne » est corrélé positivement tandis que les biens agricoles sont corrélés négativement. Ce manque de corrélation est dû au fait qu'il est difficile au Burundi de mesurer la richesse des parents à travers leurs biens car la population est essentiellement rurale.

Le fait que l'enseignant soit une femme est ici corrélée négativement avec la performance en lecture des élèves. Dans l'évaluation précédente, ils ont trouvé que cela avait un effet négatif et l'ont expliqué par le fait que les femmes ont tendance à plus s'absenter à cause des congés de maternité. Nous verrons plus loin ce qu'il en ait ici. Cela contredit quelque peu les résultats obtenus dans les modèles macroéconomiques qui montraient un meilleur taux d'achèvement dans les pays où la proportion de femmes est plus importante.

Nous remarquons aussi en général que les méthodes et formations de l'enseignant et du directeur ne sont pas corrélées avec la performance des élèves. Les observations en classe ne sont pas vraiment corrélées avec le nombre de mots lus par minute et les pratiques pédagogiques semblent très homogènes au Burundi.

Il nous faut donc voir plus clairement les variables qui influencent la performance en lecture des élèves.

III- Application des moindres carrés ordinaires

3-1) Présentation du modèle

Nous avons appliqué les moindres carrés ordinaires tenant compte de toutes les variables.

Tableau44 : Modèle1 (MCO Micro)

Source	Valeur	Ecart-type	t	Pr > |t|	Borne inf (95%)	Borne sup (95%)	VIF
Constante	-24,595	4,588	-5,360	< 0,0001	-33,596	-15,594	
motstext_perminut2011_AGR	0,669	0,047	14,220	< 0,0001	0,577	0,761	1,752
niveau	10,191	0,686	14,850	< 0,0001	8,845	11,537	1,497
fille	2,118	0,572	3,701	0,000	0,995	3,240	1,039
age_ES	-0,231	0,195	-1,187	0,236	-0,614	0,151	1,426

maison_dur	1,302	0,802	1,623	0,105	-0,271	2,876	1,178
Equipement	0,073	0,359	0,204	0,839	-0,632	0,778	1,536
Bien_agricol	-0,282	0,265	-1,065	0,287	-0,801	0,237	1,275
boutique	-0,959	1,205	-0,796	0,426	-3,323	1,404	1,086
mange	0,120	0,600	0,200	0,842	-1,058	1,298	1,145
trav_dom_svt	1,307	0,707	1,850	0,065	-0,079	2,693	1,085
maternelle	0,002	0,736	0,003	0,997	-1,440	1,445	1,139
REDOUBL	-1,866	0,734	-2,542	0,011	-3,306	-0,426	1,168
francais	7,425	3,125	2,376	0,018	1,294	13,556	1,078
kirundi	1,523	1,400	1,087	0,277	-1,224	4,269	1,241
livr_class	0,119	1,550	0,077	0,939	-2,921	3,159	1,053
opport_lect	0,847	0,139	6,117	< 0,0001	0,576	1,119	2,177
aide_perso	1,563	0,819	1,908	0,057	-0,044	3,170	2,030
nb_freresoeu_ES	-0,106	0,140	-0,759	0,448	-0,380	0,168	1,092
pere_agriculteur	-0,477	0,642	-0,743	0,458	-1,737	0,783	1,308
mere_agriculteur	0,807	0,782	1,032	0,302	-0,727	2,342	1,402
dir_femme	-0,757	0,776	-0,975	0,330	-2,280	0,766	1,404
dir_nouvo	3,787	0,917	4,128	< 0,0001	1,987	5,587	1,669
dir_fonct	-2,497	1,059	-2,358	0,019	-4,574	-0,419	1,438
dir_formanim	2,323	0,753	3,086	0,002	0,846	3,799	1,683
dir_formgest	-2,882	0,720	-4,001	< 0,0001	-4,295	-1,469	1,612
nb_visinsp_ES	0,005	0,297	0,016	0,987	-0,578	0,587	1,352
freq_reunion_mois	1,194	0,811	1,473	0,141	-0,396	2,784	1,527
outil_egra	-0,404	0,252	-1,599	0,110	-0,898	0,091	1,624
insecurite	0,697	0,257	2,716	0,007	0,193	1,200	1,513
femme	1,803	0,740	2,436	0,015	0,351	3,256	1,626
DUREE_FI	0,028	0,406	0,069	0,945	-0,769	0,825	1,367
formcont	1,971	0,735	2,681	0,007	0,529	3,413	1,200
presents	0,013	0,018	0,709	0,478	-0,023	0,049	1,460
Txabandon_ES	-3,637	2,690	-1,352	0,177	-8,915	1,640	1,370
manuel_kiru	2,707	1,424	1,901	0,058	-0,087	5,501	1,424
fichier_kirundi	-1,790	0,768	-2,331	0,020	-3,297	-0,284	1,565
mat_innovateur	2,124	0,760	2,796	0,005	0,633	3,615	1,427
m_globale	-0,225	0,733	-0,308	0,758	-1,663	1,213	1,692
m_mixte	-0,814	0,804	-1,012	0,311	-2,392	0,763	1,977
remediation	-0,354	0,403	-0,880	0,379	-1,145	0,436	1,433
kirundi_dif	-0,432	1,707	-0,253	0,800	-3,781	2,917	1,576
RCOR3	1,198	1,322	0,906	0,365	-1,396	3,793	1,602
RCOR9	-0,572	1,691	-0,338	0,735	-3,890	2,746	2,084
RCOR10	10,569	3,576	2,956	0,003	3,554	17,584	2,410
RCOR12	-5,185	2,633	-1,969	0,049	-10,350	-0,020	2,298
RCOR13	-0,522	2,096	-0,249	0,803	-4,633	3,589	2,967
RCOR14	6,023	4,489	1,342	0,180	-2,783	14,829	1,512
RCOR19	-15,973	15,062	-1,060	0,289	-45,522	13,575	1,378
RCOR20	-13,442	4,322	-3,110	0,002	-21,922	-4,963	1,901

Fait par les auteurs(XLSTAT)

3-2) Résultats du modèle

- Interprétation des coefficients :

Les variables significatives du modèle sont :

Tableau45 : Variables significatives Modèle1 (MCO Micro)

Source	Valeur	Ecart-type	t	Pr > |t|	Borne inf (95%)	Borne sup (95%)
Constante	-24,595	4,588	-5,360	< 0,0001	-33,596	-15,594
motstext_perminut2011_AGR	0,669	0,047	14,220	< 0,0001	0,577	0,761
niveau	10,191	0,686	14,850	< 0,0001	8,845	11,537
fille	2,118	0,572	3,701	0,000	0,995	3,240
maison_dur	1,302	0,802	1,623	0,105	-0,271	2,876
trav_dom_svt	1,307	0,707	1,850	0,065	-0,079	2,693
REDOUBL	-1,866	0,734	-2,542	0,011	-3,306	-0,426
francais	7,425	3,125	2,376	0,018	1,294	13,556
opport_lect	0,847	0,139	6,117	< 0,0001	0,576	1,119
aide_perso	1,563	0,819	1,908	0,057	-0,044	3,170
dir_nouvo	3,787	0,917	4,128	< 0,0001	1,987	5,587
dir_fonct	-2,497	1,059	-2,358	0,019	-4,574	-0,419
dir_formanim	2,323	0,753	3,086	0,002	0,846	3,799
dir_formgest	-2,882	0,720	-4,001	< 0,0001	-4,295	-1,469
outil_egra	-0,404	0,252	-1,599	0,110	-0,898	0,091
insecurite	0,697	0,257	2,716	0,007	0,193	1,200
femme	1,803	0,740	2,436	0,015	0,351	3,256
formcont	1,971	0,735	2,681	0,007	0,529	3,413
manuel_kiru	2,707	1,424	1,901	0,058	-0,087	5,501
fichier_kirundi	-1,790	0,768	-2,331	0,020	-3,297	-0,284
mat_innovateur	2,124	0,760	2,796	0,005	0,633	3,615
RCOR10	10,569	3,576	2,956	0,003	3,554	17,584
RCOR12	-5,185	2,633	-1,969	0,049	-10,350	-0,020
RCOR20	-13,442	4,322	-3,110	0,002	-21,922	-4,963

Fait par les auteurs(XLSTAT)

Nous voyons donc, comme au niveau des corrélations que les variables les plus pertinentes dans le modèle sont le niveau de la classe et les tests de l'année passée qui exercent un effet positif sur la performance des élèves. Le fait que l'élève soit une fille a un effet positif significatif. Bien que l'équipement et les biens agricoles ne soient pas significatifs, nous voyons que la qualité de la maison a un effet positif.

Les travaux domestiques ont un effet positif significatif, cela peut être dû au fait que certains travaux domestiques, comme la tenue d'une caisse, ne sont pas vraiment contraignants et rendent habiles les élèves.

Nous voyons que le parler du français à la maison a un grand effet positif sur la performance en lecture de l'élève, de même que l'indice d'opportunité de lecture. Le redoublement, comme attendu, a une mauvaise influence sur la performance des élèves.

Pour ce qui est des caractéristiques de l'école, le fait que le directeur soit nouveau ici a une influence positive. Cela peut être dû au fait qu'à partir d'un certain âge l'ancienneté devient un frein à la performance des élèves comme l'évaluation passée l'à soulignée. La formation pédagogique est également bénéfique pour les résultats en lecture.

Les enseignants femmes ont ici un effet positif sur les performances des élèves. Nous remarquons que les méthodes d'apprentissage ne sont pas significatives, mais que l'utilisation des manuels scolaires et de matériels innovateurs tels que les journaux ou les fichiers audio améliorent les rendements des élèves.

Pour les observations, le fait que l'enseignant pose une question a un effet positif, cela montrerait en effet les qualités pédagogiques de celui-ci. Le fait que les élèves lisent en chœur (RCOR12) a un effet négatif sur les résultats des élèves, ainsi que le fait que l'élève réponde à une question.

- Tolérance et VIF :

Comme mentionné plus haut, elle permet de pour déterminer le degré de la liaison linéaire entre une variable explicative X_i et les autres X_j.

Dans le cadre des études microéconomiques avec plusieurs observations, il est préférable d'avoir des VIF inférieurs à 2.

Nous voyons dans quelques colinéarités (VIF>2) et ce surtout dans les observations.

- Le coefficient de détermination :

Il est de 0,388 pour notre modèle, ce qui est relativement aux études microéconomiques un bon résultat.

Nous pouvons également observer la table ANOVA pour voir la significativité de notre modèle.

Tableau46 : Table ANOVA Modèle1 (MCO, Micro)

Source	DDL	Somme des carrés	Moyenne des carrés	F	Pr > F
Modèle	49	90352,081	1843,920	17,109	< 0,0001
Erreur	1320	142262,401	107,775		
Total corrigé	1369	232614,482			

Fait par les auteurs(XLSTAT)

Le modèle est donc globalement très significatif ; l'ensemble des variables expliquent bien le nombre de mots lus par minute.

Le fait que certaines variables soient colinéaires constitue un frein pour la mesure réelle de leur impact sur la performance en lecture des élèves. Nous allons, pour palier à ce problème et aussi afin de regrouper nos variables par catégories, appliquer la modélisation des moindres carrés partiels.

IV- Application des moindres carrés partiels

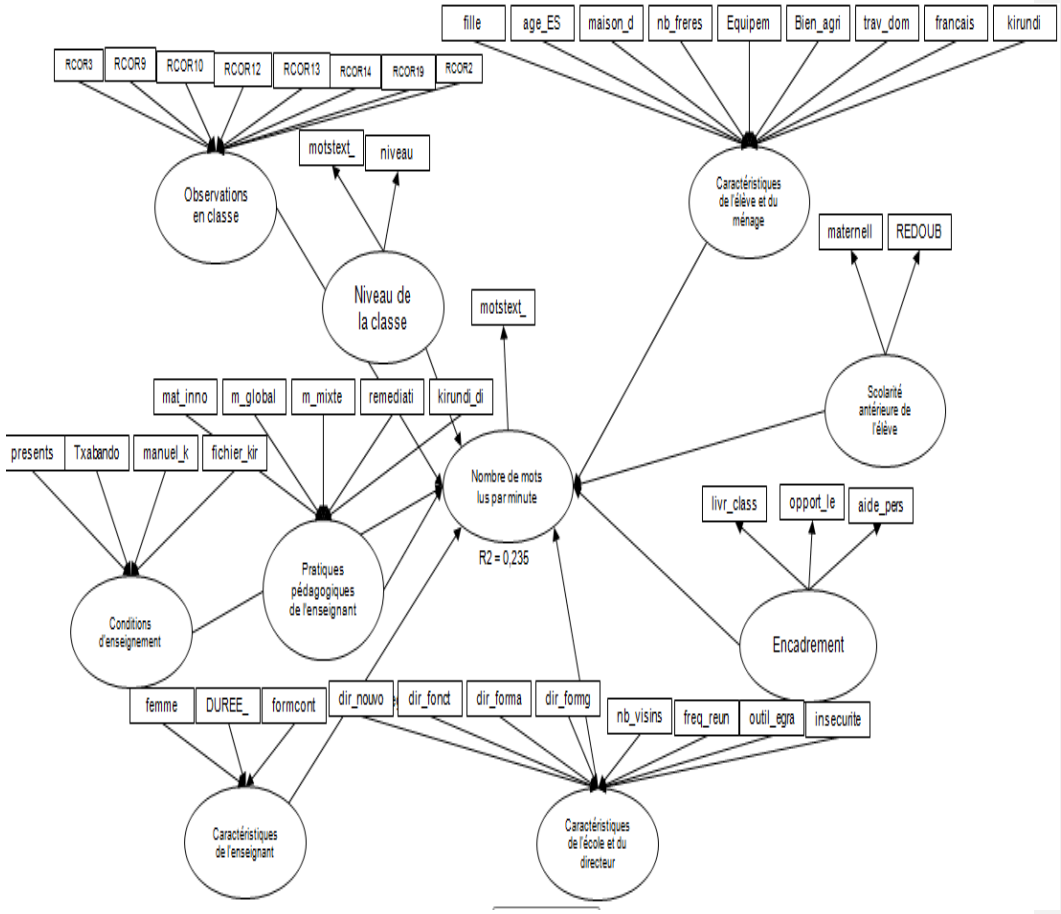
4-1) Présentation du modèle

La modélisation des moindres carrés partiels (*pls-path modeling*) sert à regrouper plusieurs variables, appelé bloc de variables, en un seul trait latent. Elle permet ainsi de réduire le nombre de variables explicatives et permet une meilleure présentation du modèle. Contrairement à l'analyse macroéconomique nous avons ici plusieurs variables qui mesurent un même trait latent et qui ne sont pas forcément corrélées. Nous avons dans ce cas appliqué un modèle formatif, qui ne nécessite pas forcément l'unidimensionnalité des blocs de variables. Il s'agit notamment des caractéristiques de l'élève et du ménage. Elles contiennent le sexe de l'élève (fille) qui n'est pas forcément lié à son âge, au nombre de frères et sœurs qu'il possède, les biens que possèdent ses parents, ni le fait de parler français à la maison. Mais nous pouvons regrouper ces variables pour en tirer une variable qui mesure les caractéristiques de l'élève et du ménage. Il s'agit de former une variable à partir de ces différentes variable, d'où le mode « formatif ». Il en est de même pour les observations en classe ; le fait que les élèves font une lecture silencieuse n'est pas forcément lié au fait qu'un élève pose une question ou qu'un élève réponde à une question. Toutes les variables qui forment les conditions d'enseignement également ne sont pas non plus colinéaires. Il en est de même pour les caractéristiques de l'enseignant et les caractéristiques de l'école et du directeur.

4-2) Modèle1

Nous avons dans un premier temps fait notre modélisation en tenant compte de toutes les variables explicatives :

Graphique30 : Modèle1 (PLS Micro)



Fait par les auteurs(XLSTAT)

Les indices de communalités nous permettent de mesurer la liaison des variables latentes avec leurs variables manifestes :

Tableau47 : Indices de communalités (Modèle1, Micro)

Variable latente	Variables manifestes	Communalités
Nombre de mots lus par minute	motstext_perminut	
Niveau de la classe	motstext_perminut2011_AGR	1,000
	niveau	0,000
Pratiques pédagogiques de l'enseignant	mat_innovateur	0,040
	m_globale	0,125
	m_mixte	0,032

	remediation	0,183
	kirundi_dif	0,615
Caractéristiques de l'enseignant	femme	0,908
	DUREE_FI	0,058
	formcont	0,030
Caractéristiques de l'école et du directeur	dir_nouvo	0,016
	dir_fonct	0,085
	dir_formanim	0,054
	dir_formgest	0,113
	nb_visinsp_ES	0,012
	freq_reunion_mois	0,054
	outil_egra	0,003
	insecurite	0,615
Encadrement	livr_class	0,003
	opport_lect	1,000
	aide_perso	0,495
Scolarité antérieure de l'élève	maternelle	0,298
	REDOUBL	0,772
Caractéristiques de l'élève et du ménage	fille	0,308
	age_ES	0,269
	maison_dur	0,078
	nb_freresoeu_ES	0,041
	Equipement	0,016
	Bien_agricol	0,024
	trav_dom_svt	0,013
	francais	0,120
	kirundi	0,036
Observations en classe	RCOR3	0,530
	RCOR9	0,039
	RCOR10	0,222
	RCOR12	0,186
	RCOR13	0,265
	RCOR14	0,010
	RCOR19	0,190
	RCOR20	0,031
Conditions d'enseignement	presents	0,508
	Txabandon_ES	0,399
	manuel_kiru	0,117
	fichier_kirundi	0,029

Fait par les auteurs(XLSTAT)

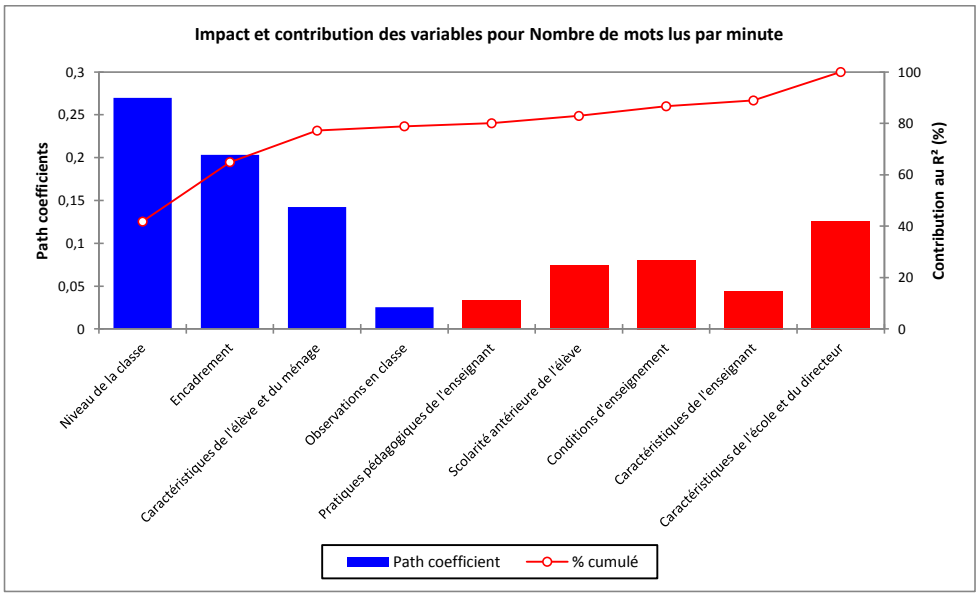
Ces indices sont pour la plus part faibles à cause du nombre d'observation et de l'impact sur la performance des élèves. Comme nous le voyons sur le graphe, le R^2 est de 2,35. Nous pouvons voir ci-dessous l'impact des différentes variables du modèle :

Tableau48 : Effet des variables latentes (Modèle1, Micro)

Variable latente	Valeur	Ecart-type	t	Pr > |t|
Niveau de la classe	0,270	0,027	10,158	0,000
Pratiques pédagogiques de l'enseignant	-0,033	0,024	-1,355	0,176
Caractéristiques de l'enseignant	-0,044	0,025	-1,795	0,073
Caractéristiques de l'école et du directeur	-0,126	0,025	-5,147	0,000
Encadrement	0,203	0,024	8,388	0,000
Scolarité antérieure de l'élève	-0,075	0,024	-3,084	0,002
Caractéristiques de l'élève et du ménage	0,142	0,024	5,873	0,000
Observations en classe	0,025	0,026	0,970	0,332
Conditions d'enseignement	-0,080	0,024	-3,338	0,001

Fait par les auteurs(XLSTAT)

Graphique32 : Effet des variables latentes (Modèle1, Micro)



Fait par les auteurs(XLSTAT)

Nous voyons, comme pour les moindres carrés ordinaires que le niveau de la classe a un grand effet positif sur les performances des élèves. En effet le niveau de la classe renferme la classe en question (2ème ou 3ème année) et les résultats de l'année passée. Il est donc normal que les élèves de troisième année soient plus performants que ceux de la deuxième année et que les résultats passés influent sur ceux de cette année.

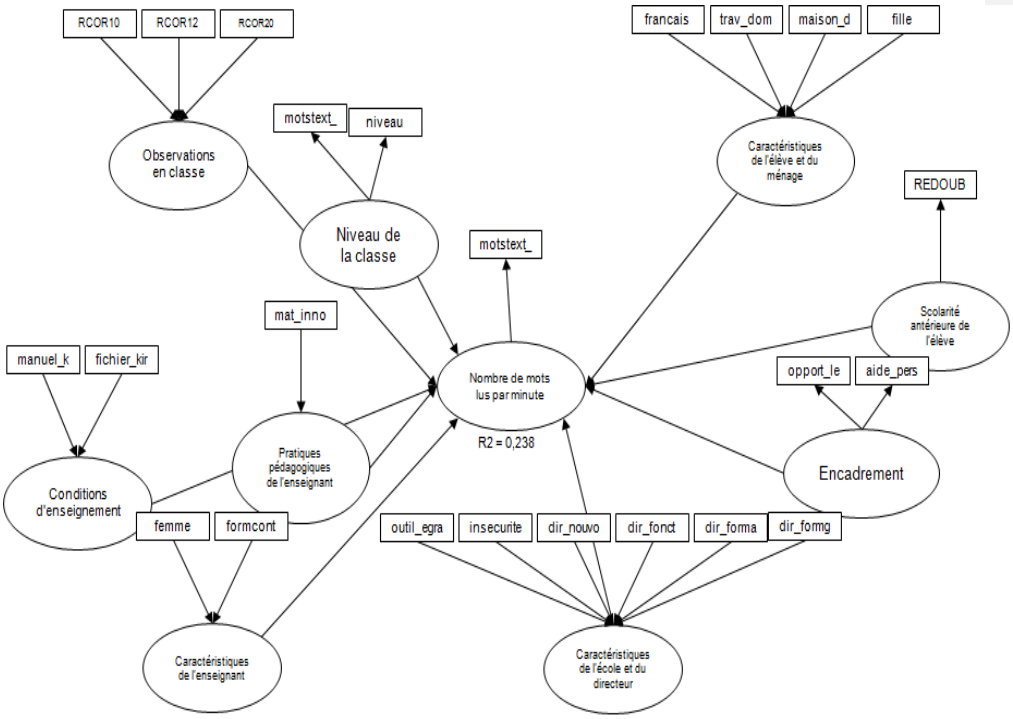
Nous remarquons aussi que l'encadrement de l'élève à la maison a un grand effet positif. De meilleures conditions de vie de l'élève et du ménage sont associées avec de meilleures performances en lecture des élèves. Les observations en classe et les pratiques pédagogiques de l'enseignant mesurées par questionnaires n'ont pas d'effets significatifs. La scolarité antérieure des élèves, qui prends en compte le taux de redoublement constitue un

frein à la réussite en lecture des élèves. Il en est de même pour les conditions de l'enseignement, très liées aux tailles de classes. Les caractéristiques de l'enseignant, notamment sont sexe et les formations pédagogiques reçu ont un effet négatif sur la réussite des élèves. Les caractéristiques de l'école et du directeur, notamment l'insécurité ont ici un impact négatif sur les performances en lecture des élèves.

Les variables non significatives dans le modèle avec les moindres carrés ordinaires n'admettent pas non plus de bons indices de communalités, nous allons donc réaliser un deuxième modèle en ne tenant compte que des variables significatives.

4-3) Modèle2

Graphique31 : Modèle2 Micro



Fait par les auteurs(XLSTAT)

Nous remarquons, malgré le fait que nous ayons diminué des variables que le R^2 a augmenté (R^2 = 0,238). Le tableau suivant nous montre les indices de communalités de chaque variable manifeste avec sa variable latente :

Tableau49 : Indices de communalités (Modèle2, Micro)

Variable latente	Variables manifestes	Communalités
Nombre de mots lus par minute	motstext_perminut	

Niveau de la classe	motstext_perminut2011_AGR	1,000
	Niveau	0,000
Pratiques pédagogiques de l'enseignant	mat_innovateur	
Caractéristiques de l'enseignant	Femme	0,939
	Formcont	0,032
Caractéristiques de l'école et du directeur	dir_nouvo	0,017
	dir_fonct	0,088
	dir_formanim	0,056
	dir_formgest	0,117
	outil_egra	0,004
	Insecurite	0,633
Encadrement	opport_lect	1,000
	aide_perso	0,495
Scolarité antérieure de l'élève	REDOUBL	
Caractéristiques de l'élève et du ménage	Fille	0,630
	maison_dur	0,160
	trav_dom_svt	0,026
	Francais	0,246
Observations en classe	RCOR10	0,511
	RCOR12	0,430
	RCOR20	0,072
Conditions d'enseignement	manuel_kiru	0,677
	fichier_kirundi	0,165

Fait par les auteurs(XLSTAT)

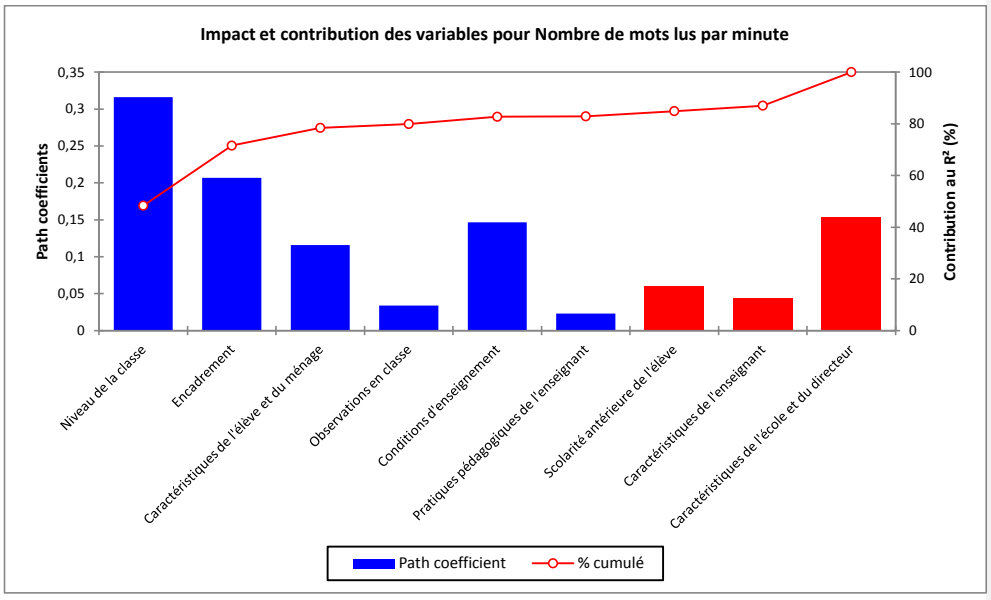
Nous notons une meilleure représentativité des variables latentes par rapport à leurs variables manifestes. Regardons à présent les effets de chaque variable sur le nombre de mots lus par minute :

Tableau50 : Effet des variables latentes (Modèle, Micro)

Variable latente	Valeur	Ecart-type	t	Pr > |t|
Niveau de la classe	0,316	0,025	12,457	0,000
Pratiques pédagogiques de l'enseignant	0,023	0,024	0,949	0,343
Caractéristiques de l'enseignant	-0,043	0,025	-1,741	0,082
Caractéristiques de l'école et du directeur	-0,152	0,025	-6,139	0,000
Encadrement	0,207	0,024	8,549	0,000
Scolarité antérieure de l'élève	-0,060	0,024	-2,517	0,012
Caractéristiques de l'élève et du ménage	0,116	0,024	4,868	0,000
Observations en classe	0,034	0,025	1,358	0,175
Conditions d'enseignement	0,147	0,024	5,985	0,000

Fait par les auteurs(XLSTAT)

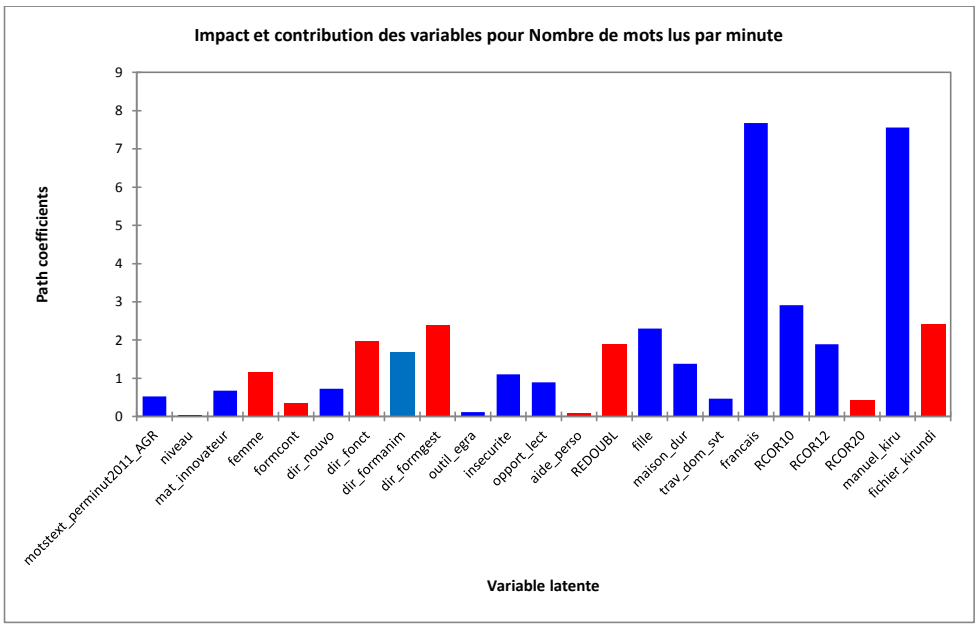
Graphique33 : Effet des variables latentes (Modèle2, Micro)



Fait par les auteurs(XLSTAT)

Les deux variables non significatives sont toujours les observations en classe et les pratiques pédagogiques de l'enseignant.

Graphique34 : Effet des variables (Modèle2, Micro)



Fait par les auteurs(XLSTAT)

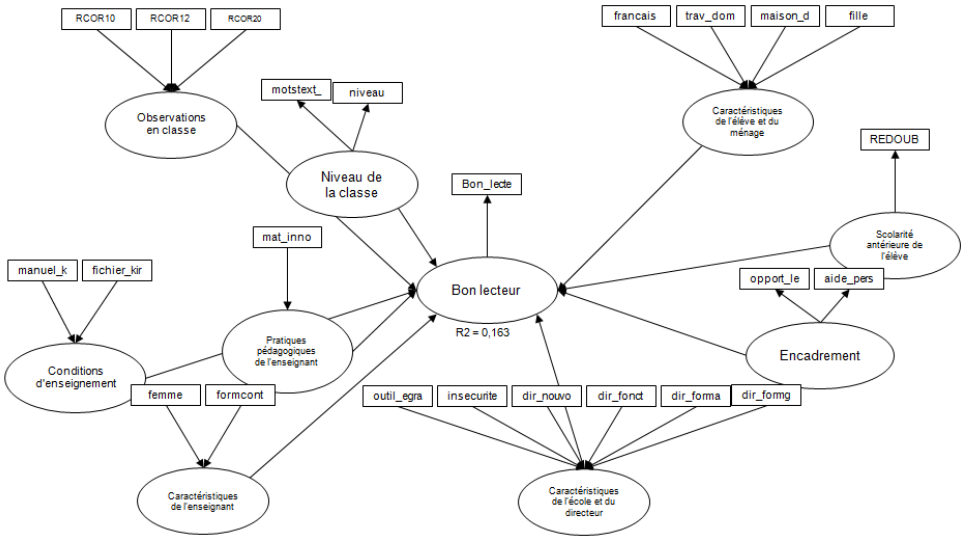
Les variables les plus pertinentes sont le fait de parler le français à la maison et l'utilisation de manuels en kirundi. C'est-à-dire que le plus important pour améliorer les performances de l'élève est de mettre à sa disposition des moyens de formation, d'encadrement. Nous remarquons que les filles réussissent mieux que les garçons. Certaines formations n'améliorent pas la performance des élèves. Elles peuvent même être un frein si elles entrainent l'absence de ce dernier. Le taux de redoublement, comme dans l'étude macroéconomique a un effet négatif sur la performance des élèves.

4-4) Modèle 3

Nous pouvons étudier en effet les variables qui font le nombre de mots lus par minute, mais nous ne savons pas si ces variables font des élèves de bons lecteurs. Il est donc important de voir ce qui différencie le bon lecteur d'un mauvais lecteur. Nous avons créé une variable binaire qui prend la valeur 1 lorsque le nombre de mots lus par minute est supérieur à 21 qui est la moyenne du nombre de mots lus par minute et 0 sinon.

Nous ajustons donc un modèle avec les variables significatives, prises en compte dans le modèle2.

Graphique35 : Modèle3 Micro



Fait par les auteurs(XLSTAT)

Le R^2 de ce modèle, comme indiqué sur le diagramme est de 0,163.

Tableau51 : Indices de communalités (Modèle3, Micro)

Variable latente	Variables manifestes	Communalités
Nombre de mots lus par minute	Bon_lecteur	
Niveau de la classe	motstext_perminut2011_AGR	1,000
	niveau	0,000
Pratiques pédagogiques de l'enseignant	mat_innovateur	
Caractéristiques de l'enseignant	femme	0,997
	formcont	0,000
Caractéristiques de l'école et du directeur	dir_nouvo	0,036
	dir_fonct	0,018
	dir_formanim	0,128
	dir_formgest	0,086
	outil_egra	0,086
	insecurite	0,631
Encadrement	opport_lect	1,000
	aide_perso	0,495

Scolarité antérieure de l'élève	REDOUBL	
Caractéristiques de l'élève et du ménage	fille	0,570
	maison_dur	0,212
	trav_dom_svt	0,002
	francais	0,270
Observations en classe	RCOR10	0,600
	RCOR12	0,335
	RCOR20	0,213
Conditions d'enseignement	manuel_kiru	0,968
	fichier_kirundi	0,000

Fait par les auteurs(XLSTAT)

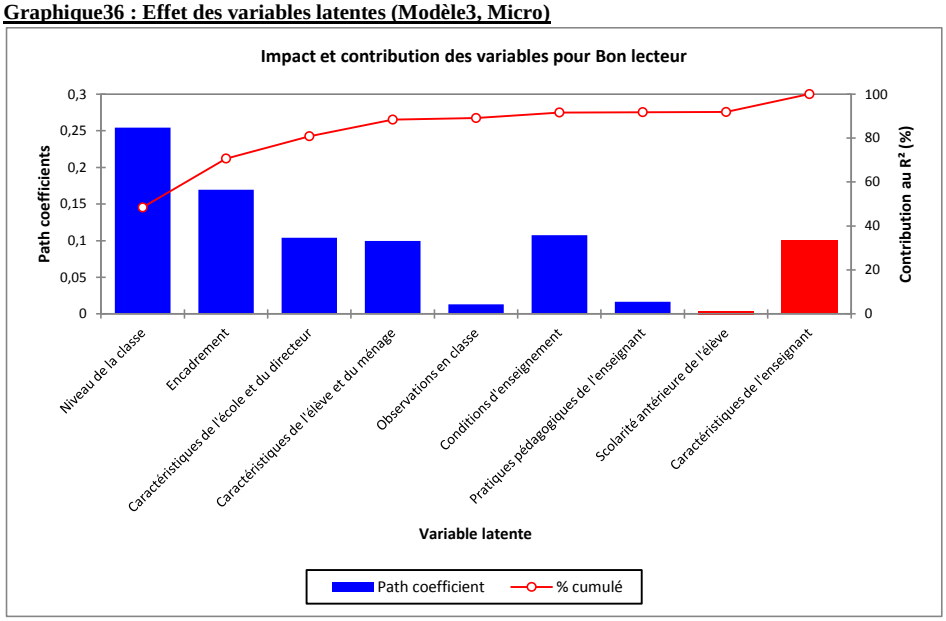
Les indices de communalités sont acceptables, comme dans le modèle précédent. Nous pouvons voir les variables qui favorisent un bon niveau de lecture :

Tableau52 : Effet des variables latentes (Modèle3, Micro)

Variable latente	Valeur	Ecart-type	t	Pr > |t|
Niveau de la classe	0,254	0,026	9,797	0,000
Pratiques pédagogiques de l'enseignant	0,017	0,025	0,662	0,508
Caractéristiques de l'enseignant	-0,101	0,026	-3,927	0,000
Caractéristiques de l'école et du directeur	0,104	0,025	4,123	0,000
Encadrement	0,170	0,025	6,798	0,000
Scolarité antérieure de l'élève	-0,004	0,024	-0,178	0,859
Caractéristiques de l'élève et du ménage	0,100	0,024	4,072	0,000
Observations en classe	0,013	0,026	0,507	0,612
Conditions d'enseignement	0,108	0,025	4,256	0,000

Fait par les auteurs(XLSTAT)

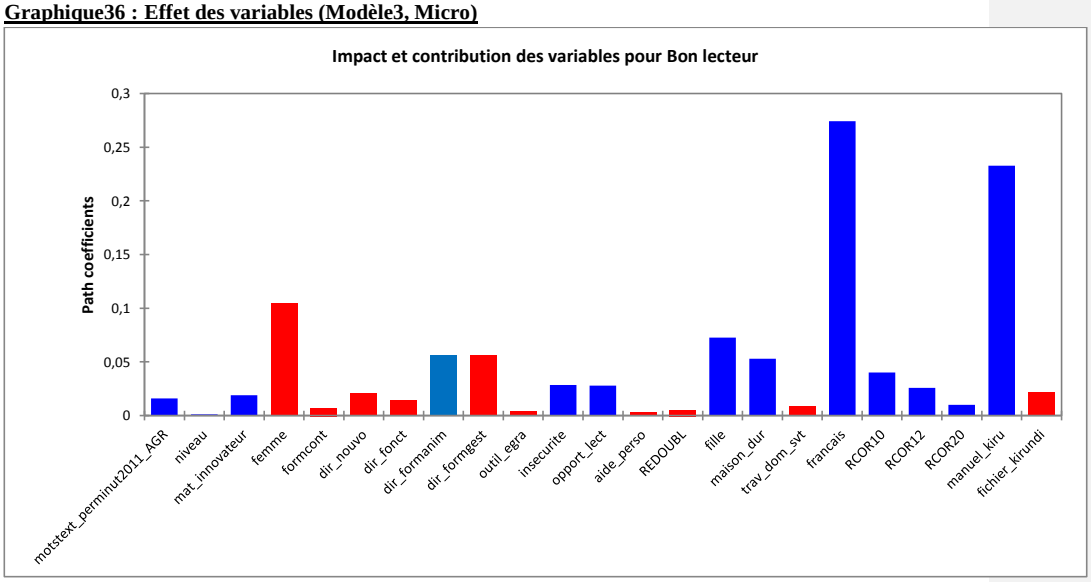
Graphique36 : Effet des variables latentes (Modèle3, Micro)



Fait par les auteurs(XLSTAT)

Nous voyons que la scolarité antérieure de l'élève devient non significative, ce qui est normal car les bons élèves ne redoublent pas. Les observations en classe et les pratiques pédagogiques de l'enseignant demeurent non significatives.

Graphique36 : Effet des variables (Modèle3, Micro)



Fait par les auteurs(XLSTAT)

Comme dans le modèle précédent, les variables les plus pertinentes sont le fait de parler de français à la maison et l'utilisation d'un manuel en kirundi. Le fait que l'enseignant soit une femme a toujours un effet négatif, alors les filles réussissent plus que les garçons.

Conclusion Générale

Au terme de notre étude, nous voyons retenons les deux approches macroéconomiques et microéconomiques sont complémentaires et nécessaires pour une meilleure appréhension de la qualité de l'éducation. La modélisation des moindres carrés partiels (*Structural Equation- Path modelling*) nous a permis d'avoir des résultats fiables et de pouvoir regrouper les variables afin de mieux présenter nos modèles.

Nous avons trouvé, comme dans la littérature que l'alphabétisation des adultes et le revenu des parents ont un impact positif sur le taux d'achèvement des élèves au primaire et aussi que les défaillances de systèmes scolaires tels que la taille de classe et les taux de répétitions constituent un frein à la qualité de l'éducation. Les dépenses, comme dans la littérature n'ont pas d'effet significatif sur le taux d'achèvement du primaire.

L'approche macroéconomique nous a permis de voir l'importance de certaines variables contextuelles autrefois négligées et qui ont une forte influence sur le taux d'achèvement du primaire, comme le taux d'internautes dans le pays qui mesure l'accès à l'information, la diversité linguistique, la croissance démographique, l'indice de corruption et l'appartenance au partenariat d'aide mondiale à l'éducation.

Le taux d'internaute dans le pays est un facteur favorisant la qualité de l'éducation. Nous avons également vu que les facteurs de population, à savoir la population, la croissance démographique et la diversité linguistique sont des entraves au taux d'achèvement, ce qui suggère que des politiques démographiques doivent être mises en œuvre parallèlement aux politiques éducatives. L'indice de corruption quant à lui n'a pas un effet significatif. Nous avons vu enfin que le Partenariat Mondial pour l'Education est bénéfique pour les pays qui y participent. Les différents modèles obtenus ont également de bonnes qualités prédictives qui pourraient permettre aux décideurs de mieux cadrer leurs politiques scolaires.

Les modèles réalisés peuvent grâce à leurs qualités prédictives servir d'outil de discussion pour définir un objectif de taux d'achèvement qui tienne compte des contextes économiques et sociodémographiques propres aux pays. Notre projet prévoit donc d'affiner les outils existants comme le cadre indicatif Fast Track.

D'autre part, dans l'analyse microéconomique, nous avons vu que les biens des parents n'ont pas d'effet significatif sur les résultats en lecture, surement parce qu'ils ne mesurent pas la richesse des parents, étant donné que le Burundi a une forte population rurale. Les filles au Burundi réussissent mieux que les garçons. Le niveau des élèves également a un grand effet sur leur performance en lecture, cela montre la qualité des programmes, l'apport d'une année supplémentaire. Le taux de redoublement comme dans l'analyse macroéconomique a un effet négatif sur les performances en lecture des élèves. Il en est de même pour les conditions de l'école telles que l'insécurité.

Nous remarquons aussi que les formations pédagogiques dispensées aux enseignants et aux directeurs ne se sont pas révélées avoir un effet significatif. Par ailleurs, le fait de parler le français à la maison et d'utiliser des manuels en kirundi a un effet positif sur la performance en lecture des élèves. Pour améliorer le rendement en lecture des élèves il faudrait donc plus les équiper eux même que de se pencher sur la formation des enseignants.

Enfin, nous pouvons noter la complémentarité des études macroéconomique et microéconomique. En effet, Les analyses microéconomiques réalisées avec les méthodes PLS montrent que ce sont avant tout des variables contextuelles qui ont un impact sur les résultats, davantage que les variables de politique éducative telle que la formation des enseignants. Cela rejoint le constat tiré des analyses macroéconomiques qui montre l'importance de ces facteurs contextuels dans l'explication des résultats.

Aussi, le fait que les dépenses ne soient pas significatives dans les modèles macroéconomiques peut être aussi lié à l'usage de ces fonds. On retrouve cet effet dans l'analyse des données du Burundi. Nous avons vu qu'il vaudrait plus investir dans l'encadrement des élèves, c'est-à-dire sensibiliser les parents à assurer un meilleur suivi en leur apprenant à parler français à la maison par exemple ou en offrant aux élèves des livres de lectures qui pourront les aider chez eux. Nous avons aussi vu l'importance de supports tels que les manuels en kirundi qui manquent parfois dans les salles de classe. Les types de formation dispensés aux enseignants et aux directeurs ont soit un effet non significatif dans nos modèles, soit un effet négatif qui peut s'expliquer par l'effet négatif occasionné par l'absence de l'enseignant au cours de ces formations.

Il serait donc plus optimal pour l'Etat d'investir dans les fournitures et l'encadrement scolaire plutôt que d'organiser des formations des enseignants qui n'ont pas vraiment d'effet sur la réussite des élèves, à moins de changer le type de formation dispensée. Un résultat surprenant au Burundi est que, comme remarqué dans l'évaluation passée, la présence des femmes enseignantes exerce un effet négatif sur la performance des élèves. Dans l'analyse macroéconomique, nous avions souligné l'impact positif qu'avait la présence des femmes enseignantes. Ce constat au Burundi pourrait être dû, comme expliqué plus haut aux absences des institutrices en raison notamment des congés de maternité.

Le Maroc également, surement à cause de son faible taux d'alphabétisation des adultes (54,51%) devrait avoir un taux d'achèvement prévu par nos modèles de 69,9% alors que le taux réel taux d'achèvement est de 80,73%, ce qui, a conditions socioéconomiques données, range le Maroc dans la catégorie des pays performants. Une analyse plus détaillée des données du Maroc pourraient donc être entreprise.

L'approche PLS s'est révélée très performante et relativement facile à mettre en œuvre sur le logiciel Xlstat. Elle pourrait être davantage utilisée dans le domaine de l'analyse des données sur l'éducation et le projet de fin de stage a permis d'équiper la société Varlyproject à cette fin.

Bibliographie

Alain Lacroux, 'L'analyse des modèles de relations structurelles par la méthode pls : une approche émergente dans la recherche quantitative en **GRH',** Rapport, Laboratoire ERMMES

Altinok, N., Murseli, H. (2007), 'International database on Human Capital Quality', Economics Letters, 96(2), pp. 237-244

ALTINOK Nadir (2006) , 'Les sources de la qualité de l'éducation', Manuscrit auteur, publié dans « N/P », IREDU (Institut de Recherche sur l'Education)

ALTINOK Nadir (2007), 'A Macroeconomic Estimation of the Education Production Function', Working Paper IREDU

Altinok Nadir (2010), 'Do School Resources Increase School Quality?', Working Paper IREDU

Ardilly . (2006), Techniques de sondage, Technip

Bernard J.M. & Vianou K. & Simon O. (2005), Le redoublement : mirage de l'école africaine : PASEC/CONFEMEN

Commission d'évaluation du système éducatif burundais Dispositif Qualité/Gestion (2010/2011), Rapport de l'évaluation diagnostique des élèves de début de 3ème année de l'enseignement primaire

Cristian Preda et Al, (2005) 'Gestion des données manquantes dans les grandes bases de données en santé', Rapport, Faculté de Médecine, CERIM, Lille, France, Departement of Computer Science, University College of Dublin

Deon FILMER et Lant.H.PRITCHETT, (2001), '*Estimating wealth effects without expenditure data- or tears: an application to educational enrollments in states of India', scientific article, Demography*

Dominique DESBOIS , 'Introduction a la regression des moindres carres partiels avec la procedure pls de sas', Cour, Institut national de la recherche agronomique-Economie et Sociologie Rurales

Doudjidingao Antoine (2011), 'Education et croissance en Afrique', Etudes Africaines, L'Harmattan

Emmanuel Jacobowicz (2007) 'Contribution aux modèles d'équations structurelles à variables latentes', Thèse, Conservatoire National des Arts et Métiers Paris

Emmanuel Jacobowicz (2008), 'Variables latentes et analyse de la satisfaction', Présentation, Journée d'Etude en Statistique, Xlstat

Emmanuel Jacobowicz (2012), 'Les modèles d'équations structurelles à variables latentes', Cours de Statistique Multivariée Approfondie, Addinsoft/Xlstat

Eric A. Hanusek, and D.D. Kimko (2000), Schooling, Labor-Force Quality, and the Growth of Nations, American Economic Review, 90(5), 1184-1208.

Eric A. Hanushek & Ludger Woessmann (2010) 'The economics of international differences in educational achievement', Working Paper 15949, National Bureau of Economis Research, Cambridge

Eric A. Hanushek & Ludger Woessmann 'Do better schools lead to more growth? cognitive skills, economic outcomes, and causation', Working Paper 14633, National Bureau of Economis Research, Cambridge

Esposito (2010), Al, 'Handbook of Partial Least Squares', Springer

Gamoran et Long (2006), 'Equality of Educational Opportunity : A 40-Year Retrospective', WCER Working Paper No. 2006-9

Gregoria Mateos-Aparicio Morales 'Partial Least Square (PLS) methods : Origins, Evolution and Application to Social Sciences', Report, Complutense University of Madrid, Spain

GUPTA S., VERHOEVEN M. et TIONGSON E. (1999), "Does Higher Government Spending Buy Better Results in Education and Health Care ? ", International Monetary Fund (IMF), Working Paper, n°99/21, February

Heyneman et Loxley,(1983), 'The Effect of Primary-School Quality on Academic Achievement Acrross Twenty-nine High-and Low-Income Countries', Publication, American Journal of Sociology Vol.88, No.6 (May 1983), 1162-1194

Institut de Statistiques de l'UNESCO (2011), Le financement de l'éducation en Afrique Subsaharienne, relever les défis de l'expansion, de l'équité et de la qualité, UNESCO BREDA, UNESCO IIPE, UNESCO ISU, Montréal.

Institut de Statistiques de l'UNESCO (2011), Global education digest 2008, UNESCO- ISU, Montréal.

Jöreskog,K.G.(1970) 'A general method for analysis of covariance structures', Biometrika 57(2)

Lee, J.W. and R.J. Barro, 2001, Schooling Quality in a cross Section of Countries, Economica, 38(272), 465-488

Mazunya M. & Varly (2011), Rapport sur l'évaluation des competences fondamentales en lecture avec les outils EGRA, PARSEB/SOFRECO

Njora Hungi & Florence W. Thuku '*Differences in pupil achievement in Kenya: Implication for policy and practice, Internal Journal of Educational Development*'

Perry Den Brok, Mieke Brekelmans et Theo Wubbels, (2004), '*Interpersonal Teacher Behaviour and Student Outcomes, School Effectiveness and School Improvement*'

PROMAN (2011), Monitoring and Evaluation Strategy, Meeting of the Board of Directors, Copenhagen, Denmark, 9-10 November 2011

Randall D. Tobias, 'An Introduction to Partial Least Squares Regression', SAS Institute

Randall E.S. & Richard G.L (2010), 'A Beginner's Guide to Structural Equation Modeling', Third Edition, ROUTEDGE

Rex B. Kline (2011), 'Principles and Practice of Structural Equation Modeling', Third Edition, GUILFORD

Richard J. (2011), Modélisation des facteurs de réussite scolaire : Cas de l'évaluation du système éducatif burundais 2010/2011, Rapport de stage, Université de Lyon/Varlyproject

Ross (2009), Présentation des résultats SACMEQ lors de la réunion des partenaires de l'initiative Fast Track, avril 2009, Copenhague

S.Garcia-Acosta & Al (1999), 'Gestion des données manquantes, abérrantes et incohérentes dans l'étude de cohorte E3N', Report, HAL author manuscript

SAS OnlineDoc, 'The MI procedure', Report, version8

Sévérine Vancolen (2004), 'La regression PLS', Diplôme Postgrade en Statistique, Université de Neuchâtel, Suisse

Tenehaus (1999), 'l'approche PLS', Revue de Statistique Appliquée 47(2)

Tenehaus et al.(2004) 'A global goodness-of-fit index for PLS structural equation modelling', In : '*Atti de la reunion Scientifica della SIS'*, Barri,pp.739-742

Varly P. (2012), Report for the Global Partnership for Education Secretariat, draft

Wold (1980), 'Model construction and evaluation when theoretical knowledge is scarce', In : Kmenta, J., Ramsey, J.B (Eds), 'Evaluation of econometric models', Academic Press, New York, pp.383-407

Wold (1982), 'Soft modeling : the basic design and some extentions', In : Jöreskog, K.G. & Wold H. (Eds), 'Systems underIndirect Observation. Vol2. North-Holland, Amsterdam, pp.1-54. 14, 20, 25, 26, 27, 99

Wold, Jöreskog (1982), 'The ML and PLS techniques for modeling with latent variables : Historical and comparative aspects', In : Jöreskog K.G, Wold H. (Eds.), 'Systems under indirect observation. Vol1. North-Holland, Amsterdam, pp.263-270

Yang C.Yuan, Al,'Multiple Imputation for Missing Data : Concepts and New Development' (Version9.0), SAS Institutes

Méthode ACP:

« Analyse en composantes principales » Ali Kouani, S. El Jamali et M.Talbi

« Analyse en Composantes Principales » C. Duby, S. Robin

Cours de L'analyse des données (Mr NSIRI)

Méthode MCO

Bernard Delyon (2012), 'Régression'

Yadolah Dodge (1999), 'Analyse de régression appliquée'

Cours de régression linéaire (Mr Chaoubi)

Source des données:

www.uis.unesco.org

www.worldbank.org

www.educationfasttrack.org

www.statsilk.com/data-sources

http://esa.un.org/wpp/Excel-Data/DB02_Stock_Indicators

http://esa.un.org/wpp/Excel-Data/DB02_Stock_Indicators/WPP2010_DB2_F01_TOTAL_POPULATION_BOTH_SEXES.XLS

www.transparancy.org

ANNEXE1 : Analyse en Composantes Principales (ACP)

Commentaire [A1]: Le plan a été changé pourquoi cette partie arrive ici ?

1) Objectifs de l'ACP

L'Analyse en Composantes Principales - que nous notons par la suite ACP - est une méthode d'analyse de données. Elle cherche à synthétiser l'information contenue dans un tableau croisant des individus et des variables quantitatives. Produire un résumé d'information au sens de l'ACP, c'est établir d'une part une similarité entre les individus, chercher des groupes d'individus homogènes, mettre en évidence une typologie d'individus. Il s'agit donc de répondre aux questions :

Quels sont les individus qui se ressemblent ? Et quels sont ceux qui s'opposent ?

Existe-t-il des groupes homogènes d'individus ?

Peut-on mettre en évidence une typologie d'individus ?

Quant aux variables c'est mettre en évidence des bilans de liaisons entre elles, moyennant des variables synthétiques et mettre en évidence une typologie de variables. La liaison des variables est mesurée en ACP en termes de corrélation. Il s'agit donc de répondre aux questions suivantes :

Quels sont les variables corrélées positivement ou négativement ?

Existe-t-il des groupes de variables corrélées entre elles ? Et peut-on faire une typologie de variables?

Peut-on résumer l'ensemble des variables par un nombre réduit de variables synthétiques appelées composantes principales, chacune représentant un groupe de variables ?

2) Notations

Les données sont les mesures effectuées sur n unités $\{u_1, u_2, ..., u_i, ...u_n\}$. Les p variables quantitatives qui représentent ces mesures sont $\{v_1, v_2, ..., v_j, ...v_p\}$.

Le tableau des données brutes à partir duquel on va faire l'analyse est noté X et a la forme suivante :

$$\mathbf{X} = \begin{array}{c} \\ u_1 \\ u_2 \\ . \\ u_i \\ . \\ u_n \end{array} \begin{array}{c} \begin{array}{cccccc} v_1 & v_2 & \dots & v_j & \dots & v_p \end{array} \\ \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1j} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2j} & \dots & x_{2p} \\ . & . & \dots & . & \dots & . \\ x_{i1} & x_{i2} & \dots & x_{ij} & \dots & x_{ip} \\ . & . & \dots & . & \dots & . \\ x_{n1} & x_{n2} & \dots & x_{nj} & \dots & x_{np} \end{bmatrix} \end{array}$$

On peut représenter chaque unité par le vecteur de ses mesures sur les p variables :

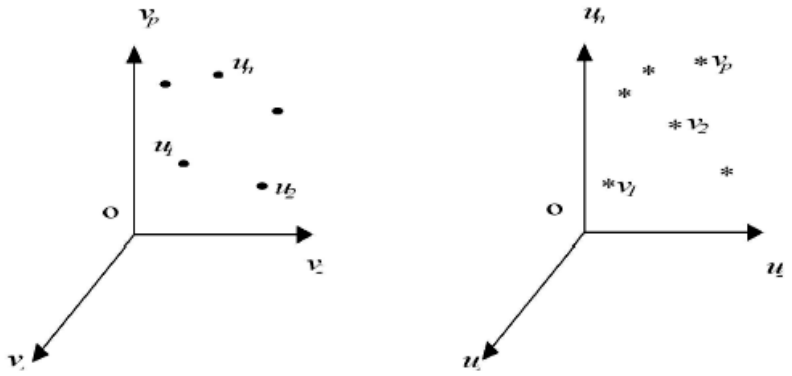
${}^tU_i = [\, x_{i1}\ x_{i2}\ \ldots\ x_{ij}\ \ldots\ x_{ip}\,]$

où ${}^tU_i$ est la transposée de U_i (vecteur de R^p).

De façon analogue, on peut représenter chaque variable par un vecteur de R^n dont les composantes sont les valeurs de la variable pour les n unités :

$$V_j = \begin{bmatrix} x_{1j} \\ x_{2j} \\ . \\ x_{ij} \\ . \\ x_{nj} \end{bmatrix}.$$

Pour avoir une image de l'ensemble des unités, on se place dans un espace affine en choisissant comme origine un vecteur particulier de R^p, par exemple le vecteur dont toutes les coordonnées sont nulles. Alors, chaque unité sera représentée par un point dans cet espace. L'ensemble des points qui représentent les unités est appelé traditionnellement "nuage des individus". En faisant de même dans R^n , chaque variable pourra être représentée par un point de l'espace affine correspondant. L'ensemble des points qui représentent les variables est appelé "nuage des variables". On constate, que ces espaces étant de dimension supérieure en général `a 2 et même 3, on ne peut visualiser ces représentations. L'idée générale des méthodes factorielles est de trouver un système d'axes et de plans tels que les projections de ces nuages de points sur ces axes et ces plans permettent de reconstituer les positions des points les uns par rapport aux autres, c'est-à-dire avoir des images les moins déformées possible.



3) Choix d'une distance

Pour faire une représentation géométrique, il faut choisir une distance entre deux points de l'espace. La distance utilisée par l'ACP dans l'espace où sont représentées les unités, est la distance euclidienne classique. La distance entre deux unités u_i et u_i' est égale à :

$$d^2(u_i, u_i') = \sum_{j=1}^{p}(x_{ij} - x_{i'j})^2$$

Avec cette distance, toutes les variables jouent le même rôle et les axes définis par les variables constituent une base orthogonale. A cette distance on associe un produit scalaire entre deux vecteurs :

$$<\overrightarrow{OU_i}, \overrightarrow{OU_{i'}}> = \sum_{j=1}^{p} x_{ij}x_{i'j} = {}^tU_iU_{i'}$$

Ainsi que la norme d'un vecteur :

$$\|\overrightarrow{ou_i}\|^2 = \sum_{j=1}^{p} x_{ij}^2 = {}^tU_i\ U_i$$

On peut alors définir l'angle α entre deux vecteurs par son cosinus :

$$\cos(\alpha) = \frac{\langle \overrightarrow{ou_i}, \overrightarrow{ou_{i'}} \rangle}{\|\overrightarrow{ou_i}\|\ \|\overrightarrow{ou_{i'}}\|} = \frac{\sum_{j=1}^{p} x_{ij}x_{i'j}}{\sqrt{\sum_{j=1}^{p} x_{ij}^2 \sum_{j=1}^{p} x_{i'j}^2}} = \frac{{}^tU_i\ U_{i'}}{\sqrt{({}^tU_i\ U_i)\ ({}^tU_{i'}\ U_{i'})}}$$

4) Choix de l'origine

Le point o correspondant au vecteur de coordonnées toutes nulles n'est pas forcément une origine satisfaisante, car si les coordonnées des points du nuage des individus sont grandes, le nuage est éloigné de cette origine. Il apparaît plus judicieux de choisir une origine liée au nuage lui-même : le centre de gravité du nuage. Pour définir ce centre de gravité, il faut choisir un système de pondération des unités :

∀ i = 1, ..., n p_i = poids de l'unité u_i tel que $\sum_{i=1}^{n} p_i$. Par définition, le centre de gravité est défini comme le point tel que :

$$\sum_{i=1}^{n} p_i \overrightarrow{GU_i} = \overrightarrow{0}$$

Pour l'ACP on choisit de donner le même poids $\frac{1}{n}$ a tous les individus.

Le centre de gravité G du nuage des individus est alors le point dont les coordonnées sont les valeurs moyennes des variables :

$$G = \begin{pmatrix} \frac{1}{n}\sum_{i=1}^{n} x_{i1} \\ \vdots \\ \frac{1}{n}\sum_{i=1}^{n} x_{ij} \\ \vdots \\ \frac{1}{n}\sum_{i=1}^{n} x_{ip} \end{pmatrix} = \begin{pmatrix} x_{\bullet 1} \\ \vdots \\ x_{\bullet j} \\ \vdots \\ x_{\bullet p} \end{pmatrix}$$

Prendre G comme origine, conformément `a la figure suivante, revient alors à travailler sur le tableau des données centrées :

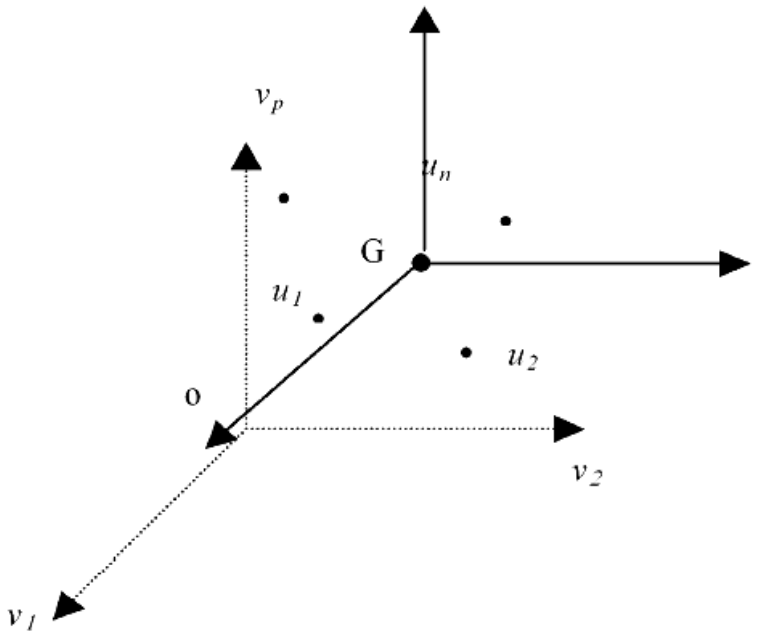
$$\mathbf{X}_c = \begin{bmatrix} x_{11} - x_{\bullet 1} & \cdots & x_{1j} - x_{\bullet j} & \cdots & x_{1p} - x_{\bullet p} \\ \vdots & & . & & \vdots \\ x_{i1} - x_{\bullet 1} & \cdots & x_{ij} - x_{\bullet j} & \cdots & x_{ip} - x_{\bullet p} \\ \vdots & & \vdots & & \vdots \\ x_{n1} - x_{\bullet 1} & \cdots & x_{nj} - x_{\bullet j} & \cdots & x_{np} - x_{\bullet p} \end{bmatrix}$$

Et le vecteur des coordonnées centrées de l'unité u_i est :

$$U_{ci} = \begin{bmatrix} x_{i1} - x_{\bullet 1} \\ x_{i2} - x_{\bullet 2} \\ \vdots \\ x_{ij} - x_{\bullet j} \\ \vdots \\ x_{ip} - x_{\bullet p} \end{bmatrix}$$

Celui des coordonnées centrées de la variable v_j est :

$$V_{cj} = \begin{bmatrix} x_{1j} - x_{\bullet j} \\ \vdots \\ x_{ij} - x_{\bullet j} \\ \vdots \\ x_{nj} - x_{\bullet j} \end{bmatrix}.$$



5) Moments d'inertie

5.1) Inertie totale du nuage des individus

On note I_G le moment d'inertie du nuage des individus par rapport au centre de gravité G :

$$I_G = \frac{1}{n}\sum_{i=1}^{n} d^2(G, u_i) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{p}(x_{ij} - x_{.j})^2 = = \frac{1}{n}\sum_{i=1}^{n} U'_{ci}U_{ci}.$$

Ce moment d'inertie totale est intéressant car c'est une mesure de la dispersion du nuage des individus par rapport `a son centre de gravit´e. Si ce moment d'inertie est grand, cela signifie que le nuage est très dispersé, tandis que s'il est petit, alors le nuage est très concentré sur son centre de gravité.

Remarque : On peut voir, en inversant l'ordre des signes sommes, que I_G peut aussi s'écrire sous la forme suivante :

$$I_G = \sum_{j=1}^{p}\left[\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - x_{\bullet j})^2\right] = \sum_{j=1}^{p}\mathrm{Var}(v_j)$$

où Var(vj) est la variance empirique de la variable v_j . Sous cette forme, on constate que l'inertie totale est égale à la trace de la matrice de covariance ∑.

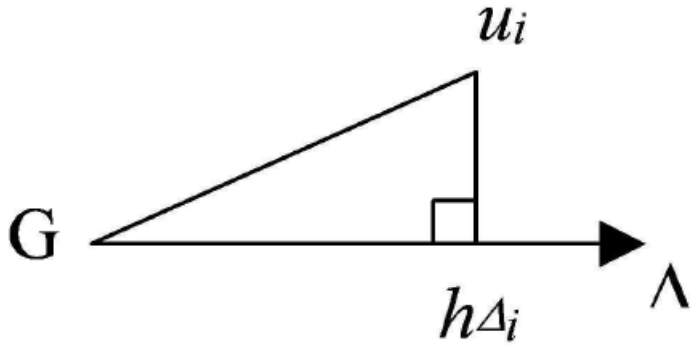
5.2 Inertie du nuage des individus par rapport à un axe passant par G

L'inertie du nuage des individus par rapport à un axe Δ passant par G est égale, par définition, à :

$$I_\Delta = \frac{1}{n}\sum_{i=1}^{n} d^2(h_{\Delta i}, u_i)$$

où $h_{\Delta i}$ est la projection orthogonale de u_i sur l'axe Δ.

Cette inertie mesure la proximité à l'axe Δ du nuage des individus.



5.3 Inertie du nuage des individus par rapport `a un sous-espace vectoriel V passant par G

Cette inertie est, par définition, égale à :

$$I_V = \frac{1}{n}\sum_{i=1}^{n} d^2(h_{Vi}, u_i)$$

où h_{Vi} est la projection orthogonale de u_i sur le sous-espace V.

5.4 Décomposition de l'inertie totale

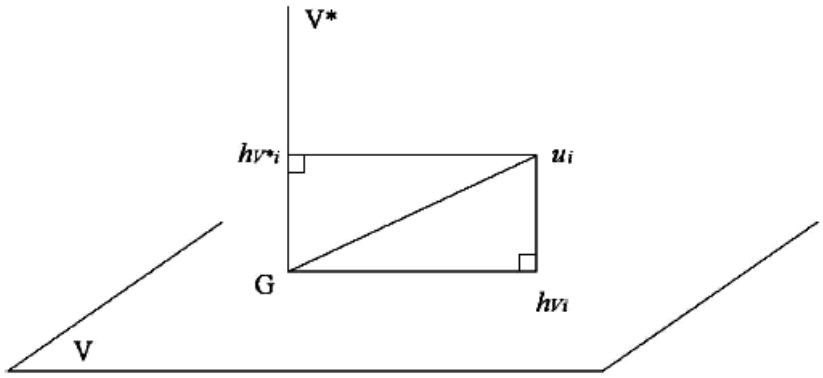
Si on note V* le complémentaire orthogonal de V dans R^p et h_{V^*i} la projection orthogonale de u_i sur V*, en appliquant le théorème de Pythagore, on peut écrire :

$$d^2(h_{Vi}, u_i) + d^2(h_{V^*i}, u_i) = d^2(G, u_i) = d^2(G, h_{Vi}) + d^2(G, h_{V^*i})$$

On en déduit, c'est le théorème de Huygens, que :

$$I_v + I_{v^*} = I_G$$

Dans le cas particulier où le sous-espace est de dimension 1, c'est-à-dire est un axe, I_{V^*}est une mesure de l'allongement du nuage selon cet axe. On emploie pour I_{V^*} les expressions "d'inertie portée par l'axe" ou bien "d'inertie expliquée par l'axe"



En projetant le nuage des individus sur un sous-espace V, on perd l'inertie mesurée par I_V, on ne conserve que celle mesurée par I_{V^*}.

De plus, si on décompose l'espace R^p comme la somme de sous-espaces de dimension

1 et orthogonaux entre eux :

$$\Delta_1 \oplus \Delta_2 \oplus ... \oplus \Delta_p$$

On écrire :

$$I_G = I_{\Delta_1^*} + I_{\Delta_2^*} + ... + I_{\Delta_p^*}$$

6) Recherche de l'axe Δ_1 passant par G d'inertie minimum

On cherche un axe Δ_1 passant par G d'inertie $I_{\Delta 1}$ minimum car c'est l'axe le plus proche de l'ensemble des points du nuage des individus, et donc, si l'on doit projeter ce nuage sur cet axe, c'est lui qui donnera l'image la moins déformée du nuage. Si on utilise la relation entre les inerties données au paragraphe précédent, rechercher Δ_1 tel que $I_{\Delta 1}$ est minimum, est équivalent à chercher Δ_1 tel que $I_{\Delta^* 1}$ est maximum.

$I_{\Delta 1}$ est minimum $\Leftrightarrow$ $I_{\Delta^* 1}$ est maximum

On définit l'axe Δ_1 par son vecteur directeur unitaire $\vec{Ga}_1$

Il faut donc trouver $\vec{Ga}_1$ tel que $I_{\Delta^* 1}$ est maximum sous la contrainte que $||\vec{Ga}_1||$.

6.1 Expressions algébriques de $I_{\Delta 1^*}$ et de $||\vec{Ga}_1||$

$$d^2(G, h_{\Delta_1^* i}) = \left\langle \overrightarrow{Gu_i}, \overrightarrow{Ga_1} \right\rangle^2 = {}^t a_1 U_{ci}\ {}^t U_{ci}\ a_1$$

En utilisant la symétrie du produit scalaire. On en déduit :

$$I_{\Delta_1^*} = \frac{1}{n}\sum_{i=1}^{n} {}^t a_1 U_{ci}\ {}^t U_{ci}\ a_1 = {}^t a_1 \left[\frac{1}{n}\sum_{i=1}^{n} U_{ci}\ {}^t U_{ci}\right]\ a_1$$

Entre crochets on reconnaît la matrice de covariance empirique ∑ des p variables.

$$I_{\Delta_1^*} = {}^t a_1 \Sigma\ a_1$$

Et

$$\left\|\overrightarrow{Ga_1}\right\|^2 = {}^t a_1\ a_1$$

6.2 Recherche du maximum

Le problème à résoudre : trouver a_1 tel que ${}^t a_1 a_1$ soit maximum avec la contrainte ${}^t a_1 a_1 = 1$ est le problème de la recherche d'un optimum d'une fonction de plusieurs variables liées par une contrainte (les inconnues sont les composantes de a_1. La méthode des multiplicateurs de Lagrange peut alors être utilisée.

Dans le cas de la recherche de a_1, il faut calculer les dérivées partielles de :

$$g(a_1) = g(a_{11,}a_{12,}..., a_{1p}) = {}^t a_1 \Sigma\ a_1 - \lambda_1({}^t a_1\ a_1 - 1)$$

En utilisant la dérivée matricielle, on obtient :

$$\frac{\partial g(a_1)}{\partial a_1} = 2\Sigma\ a_1 - 2\lambda_1\ a_1 = 0$$

Le système à résoudre est :

$$\begin{cases} \Sigma\ a_1 - \lambda_1\ a_1 = 0 & (1) \\ {}^t a_1\ a_1 - 1 = 0 & (2) \end{cases}$$

De l'équation matricielle de ce système on déduit que a_1 est vecteur propre de la matrice ∑ associé à la valeur propre λ_1.

En multipliant à gauche par ${}^t a_1$ les deux membres de l'équation (1) on obtient :

$${}^t a_1 \Sigma\ a_1 - \lambda_1\ {}^t a_1\ a_1 = 0$$

Et en utilisant l'équation (2) on trouve que :

$${}^t a_1 \Sigma\ a_1 = \lambda_1$$

On reconnaît que le premier membre de l'équation précédente est égal à l'inertie $I_{\Delta*1}$ qui doit être maximum. Cela signifie que la valeur propre λ_1 est la plus grande valeur propre de la matrice de covariance et que cette valeur propre est égale à l'inertie portée par l'axe Δ_1.

L'axe Δ_1 pour lequel le nuage des individus a l'inertie minimum a comme vecteur directeur unitaire le premier vecteur propre associé à la plus grande valeur propre de la matrice de covariance $\sum$.

7 Recherche des axes suivants

On recherche ensuite un deuxième axe Δ_2 orthogonal au premier et d'inertie minimum. On peut, comme dans le paragraphe précédent, définir l'axe Δ_2 passant par G par son vecteur directeur unitaire a_2. L'inertie du nuage des individus par rapport à son complémentaire orthogonal est égale à :

$$I_{\Delta_2^*} = {}^t a_2 \boldsymbol{\Sigma}\, a_2$$

Elle doit être maximum avec les deux contraintes suivantes :

$${}^t a_2\ a_2 = 1 \quad \text{et} \quad {}^t a_2\ a_1 = 0.$$

La deuxième contrainte exprime que le deuxième axe doit être orthogonal au premier et donc que le produit scalaire des deux vecteurs directeurs est nul. En appliquant la méthode des multiplicateurs de Lagrange, cette fois avec deux contraintes, on trouve que a_2 est le vecteur propre de $\sum$ correspondant à la deuxième plus grande valeur propre. On peut montrer que le plan défini par les axes Δ_1 et Δ_2 est le sous-espace de dimension 2 qui porte l'inertie maximum.

On peut rechercher de nouveaux axes en suivant la même procédure. Les nouveaux axes sont tous vecteurs propres de $\sum$ correspondant aux valeurs propres ordonnées. La matrice de covariance $\sum$ étant une matrice symétrique réelle, elle possède p vecteurs propres réels, formant une base orthogonale de R^p :

$$\begin{cases} \Delta_1 & \perp & \Delta_2 & \perp & \dots & \perp & \Delta_p \\ a_1 & \perp & a_2 & \perp & \dots & \perp & a_p \\ \lambda_1 & \geq & \lambda_2 & \geq & \dots & \geq & \lambda_p \\ I_{\Delta_1^*} & \geq & I_{\Delta_2^*} & \geq & \dots & \geq & I_{\Delta_p^*} \end{cases}$$

On passera de la base orthogonale initiale des variables centrées à la nouvelle base orthogonale des vecteurs propres de $\sum$. On appelle les nouveaux axes, axes principaux.

8-) Contributions des axes à l'inertie totale

En utilisant le théorème de Huygens, on peut décomposer l'inertie totale du nuage des individus

$$I_G = I_{\Delta_1^*} + I_{\Delta_2^*} + ... + I_{\Delta_p^*} = \lambda_1 + \lambda_2 + ... + \lambda_p$$

La contribution absolue de l'axe Δ_k à l'inertie totale du nuage des individus est égale à :

$$\mathrm{ca}(\Delta_k / I_G) = \lambda_k$$

Valeur propre qui lui est associée.

Sa contribution relative est égale à :

$$\mathrm{cr}(\Delta_k / I_G) = \frac{\lambda_k}{\lambda_1 + \lambda_2 + ... + \lambda_p}$$

On emploie souvent l'expression "pourcentage d'inertie expliquée par Δ_k".

On peut étendre ces définitions à tous les sous-espaces engendrés par les nouveaux axes. Ainsi, le pourcentage d'inertie expliqué par le plan engendré par les deux premiers axes Δ_1 et Δ_2 est égal à :

$$\mathrm{cr}(\Delta_1 \oplus \Delta_2 / I_G) = \frac{\lambda_1 + \lambda_2}{\lambda_1 + \lambda_2 + ... + \lambda_p}$$

Ces pourcentages d'inertie sont des indicateurs qui rendent compte de la part de variabilité du nuage des individus expliquée par ces sous-espaces. Si les dernières valeurs propres ont des valeurs faibles, on pourra négliger la variabilité qu'expliquent les axes correspondants.

On se contente souvent de faire des représentations du nuage des individus dans un sous-espace engendré par les d premiers axes si ce sous-espace explique un pourcentage d'inertie proche de 1. On peut ainsi réduire l'analyse `a un sous-espace de dimension d < p.

9-) Représentation des individus dans les nouveaux axes

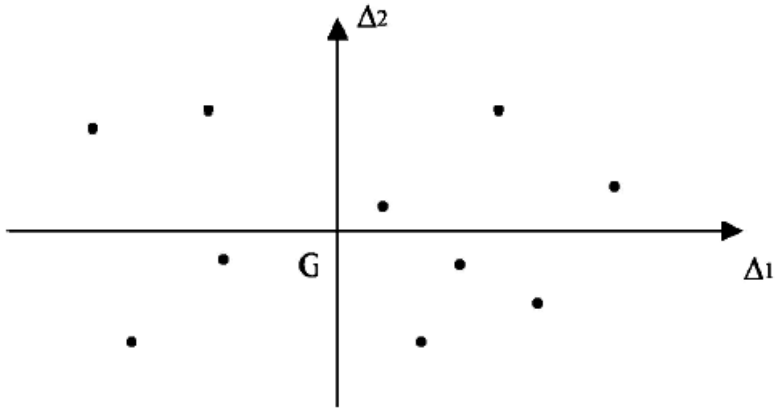
Pour faire la représentation des individus dans les plans définis par les nouveaux axes, il suffit de calculer les coordonnées des individus dans les nouveaux axes. Pour obtenir y_{ik}, coordonnée de l'unité u_i sur l'axe Δ_k, on projette orthogonalement le vecteur $\overrightarrow{Gu_i}$ sur cet axe et on obtient :

$$y_{ik} = \left\langle \overrightarrow{Gu_i}, \overrightarrow{a_k} \right\rangle = {}^t a_k \, U_{ci}$$

Et

$$Y_i = {}^t\mathbf{A} \, U_{ci}$$

où Y_i est le vecteur des coordonnées de l'unité u_i et A est la matrice du changement de base A matrice des vecteurs propres orthogonaux et de norme 1 est une matrice orthogonale, son inverse est égale à sa transposée



Remarque : L'orientation des axes est complètement arbitraire et peut différer d'un logiciel à l'autre. Le signe des coordonnées des individus sur un axe n'a donc pas de signification. En revanche, la comparaison des signes peut s'interpréter. Si deux individus u_i et $u_{i'}$ sont sur un axe Δ, le premier une coordonnée positive et le second une coordonnée négative, cela signifie qu'ils s'opposent sur cet axe.

9.1 Qualité de la représentation des individus

Lorsque des points projections des individus sont éloignés sur un axe (ou sur un plan), on peut assurer que les points représentants ces individus sont éloignés dans l'espace. En revanche, deux individus dont les projections sont proches sur un axe (ou sur un plan) peuvent ne pas être proches dans l'espace.

Pour interpréter correctement la proximité des projections de deux individus sur un plan, il faut donc s'assurer que ces individus sont bien représentés dans le plan. Pour que l'individu u_i soit bien représenté sur un axe (ou sur un plan, ou un sous-espace), il faut que l'angle entre le vecteur $\overrightarrow{Gu_i}$ et l'axe (ou le plan, ou le sous-espace) soit petit. On calcule donc le cosinus de cet angle, ou plutôt le carré de ce cosinus. En effet, en utilisant le théorème de Pythagore, on peut montrer que le carré du cosinus de l'angle d'un vecteur avec un plan engendré par deux vecteurs orthogonaux, est égal à la somme des carrés des cosinus des angles du vecteur avec chacun des deux vecteurs qui engendrent le plan. Cette propriété se généralise `a l'angle d'un vecteur avec un sous-espace de dimension k quelconque. Si le carré du cosinus de l'angle entre $\overrightarrow{Gu_i}$ et l'axe (ou le plan, ou le sous-espace) est proche de 1, alors on pourra dire que l'individu u_i est bien représenté par sa projection sur l'axe (ou le plan, ou le sous-espace). Et si deux individus sont bien représentés en projection sur un axe (ou un plan, ou un sous-espace) et ont des projections proches, alors on pourra dire que ces deux individus sont proches dans l'espace. Le carré du cosinus de l'angle α_{ik} entre $\overrightarrow{Gui}$ et un axe Δ_k de vecteur directeur unitaire a_k est égal à :

$$\cos^2(\alpha_{ik}) = \frac{\left\langle \overrightarrow{Gu_i}, \overrightarrow{Ga_k} \right\rangle^2}{\left\| \overrightarrow{Gu_i} \right\|^2} = \frac{{}^t a_k U_{ci}\ {}^t U_{ci}\ a_k}{{}^t U_{ci}\ U_{ci}} = \frac{\left[\sum_{j=1}^{p}(x_{ij} - x_{\bullet j})\ a_{kj}\right]^2}{\sum_{j=1}^{p}(x_{ij} - x_{\bullet j})^2}$$

En utilisant le théorème de Pythagore on peut calculer le carré du cosinus de l'angle $\alpha_{ikk'}$ entre $\overrightarrow{Gu_i}$ et le plan engendré par deux axes $\Delta_k \oplus \Delta_{k'}$:

$$\cos^2(\alpha_{ikk'}) = \cos^2(\alpha_{ik}) + \cos^2(\alpha_{ik'})$$

Si, après l'étude des pourcentages d'inertie expliqués par les sous-espaces successifs engendrés par les nouveaux axes, on a décidé de ne retenir qu'un sous-espace de dimension d < p, on pourra calculer la qualité de la représentation d'un individu u_i en calculant le carré du cosinus de l'angle de $\overrightarrow{Gu_i}$ avec ce sous-espace.

Remarque : Si un individu est très proche du centre de gravit´e dans l'espace, c'est-à-dire si $\left|\overrightarrow{Gu_i}\right|^2$ est très petit, le point représentant cet individu sur un axe (ou un plan, ou un sous-espace) sera bien représenté.

9.2 Interprétation des nouveaux axes en fonction des individus

Lorsqu'on calcule l'inertie I_{Δ^*k} portée par l'axe Δ_k, on peut voir quelle est la part de cette inertie due à un individu u_i particulier.

9.2.1 Contribution absolue d'un individu à un axe

I_{Δ^*k} étant égale à $\frac{1}{n}\sum_{i=1}^{n} d^2(h_{\Delta_{ki}}, G)$, la contribution absolue de u_i à cette inertie est égale à :

$$\text{ca}(u_i/\Delta_k) = \frac{1}{n}\, d^2(h_{\Delta_{ki}}, G)$$

Puisque tous les individus ont le même poids. Un individu contribuera d'autant plus à la confection d'un axe, que sa projection sur cet axe sera éloignée du centre de gravité du nuage. Inversement, un individu dont la projection sur un axe sera proche du centre de gravité contribuera faiblement à l'inertie portée par cet axe. On se sert de ces contributions pour interpréter les nouveaux axes de l'ACP en fonction des individus.

9.2.2 Contribution relative d'un individu à un axe

On peut aussi, pour un individu particulier u_i, donner sa contribution relative `a l'inertie portée par cet axe :

$$\text{cr}(u_i/\Delta_k) = \frac{\frac{1}{n}d^2(h_{\Delta_{ki}}, G)}{I_{\Delta_k^*}} = \frac{\frac{1}{n}\left\langle \overrightarrow{Gu_i}, \overrightarrow{Ga_k} \right\rangle^2}{\lambda_k} = \frac{\frac{1}{n}\ {}^t a_k U_{ci}\ {}^t U_{ci}\ a_k}{\lambda_k}$$

L'examen de ces contributions permet d'interpréter les axes principaux avec les individus.

On peut remarquer que $\sum_{i=1}^{n} cr(\frac{u_i}{\Delta_k})$ =1.

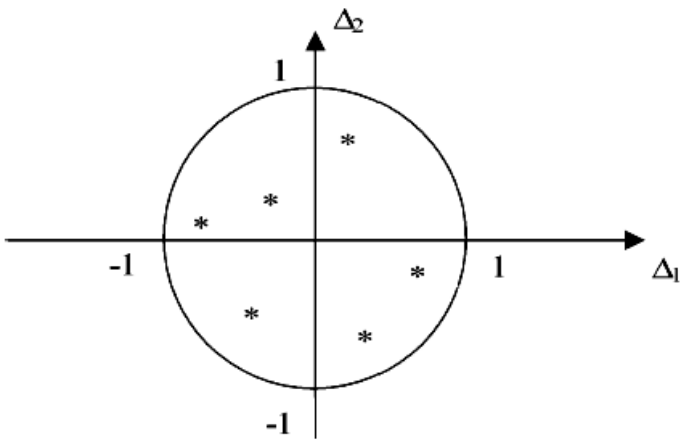
10.) Représentation des variables

On peut envisager le problème de la représentation des variables de façon complètement symétrique de celui des individus. Les raisonnements se font dans R^n au lieu de R^p. Mais dans l'ACP, au-delà de la symétrie formelle entre les individus et les variables, on peut utiliser la dissymétrie liée à la sémantique : les variables n'ont pas la même signification que les individus. On peut alors faire le raisonnement suivant : on a représenté les individus dans l'espace des anciennes variables, et on a fait un changement de base dans cet espace. Les nouveaux axes sont des combinaisons linéaires des anciens axes et peuvent donc être considérés comme de nouvelles variables combinaisons linéaires des anciennes. On appelle communément ces nouvelles variables "composantes principales".

On note Z_1, Z_2, ..., Z_k, ..., Z_p les composantes principales, Z_k étant la nouvelle variable correspondant à l'axe Δ_k :

$Z_k = \sum_{j=1}^{p} a_{kj} V_{cj} = X_c a_k$.

Il est alors intéressant de voir comment les anciennes variables sont liées aux nouvelles et pour cela on calcule les corrélations des anciennes variables avec les nouvelles. La représentation des anciennes variables se fera en prenant comme coordonnées des anciennes variables leurs coefficients de corrélation avec les nouvelles variables. On obtient alors ce que l'on appelle communément le "cercle des corrélations", dénomination qui vient du fait qu'un coefficient de corrélation variant entre -1 et +1, les représentations des variables de départ sont des points qui se trouvent à l'intérieur d'un cercle de rayon 1 si on fait la représentation sur un plan.



On peut montrer que les variances, covariances et coefficients de corrélation empiriques des composantes principales entre elles ou avec les variables de départ sont :

$$\mathrm{Var}(Z_k) = \frac{1}{n}\ {}^t a_k\ {}^t\mathbf{X}_c\ \mathbf{X}_c\ a_k = {}^t a_k\ \mathbf{\Sigma}\ a_k = \lambda_k$$

$$\mathrm{Cov}(Z_k, V_{cj}) = \frac{1}{n}\ {}^t a_k\ {}^t\mathbf{X}_c\ V_{cj} = \frac{1}{n}\ {}^t a_k\ {}^t\mathbf{X}_c\ \mathbf{X}_c \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix}$$

$$= {}^t a_k\ \mathbf{\Sigma} \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix} = \lambda_k\ {}^t a_k \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix} = \lambda_k\ a_{kj}$$

Enfin :

$$\mathrm{Cor}(Z_k, V_{cj}) = \sqrt{\lambda_k}\frac{a_{kj}}{\sqrt{\mathrm{Var}(V_j)}}$$

où a_{kj} est la $j^{ème}$ coordonnée du vecteur directeur unitaire a_k de Δ_k.

De façon générale, la matrice de covariance des composantes principales est égale à :

$$\mathbf{\Sigma_Z} = \frac{1}{n}\ {}^t\mathbf{A}\ {}^t\mathbf{X}_c\ \mathbf{X}_c\ \mathbf{A} = {}^t\mathbf{A}\ \mathbf{\Sigma}\ \mathbf{A} = \mathbf{\Lambda}$$

Où Λ est la matrice diagonale des valeurs propres de $\sum$:

$$\mathbf{\Lambda} = \begin{pmatrix} \lambda_1 & & (\mathbf{0}) \\ & \ddots & \\ (\mathbf{0}) & & \lambda_p \end{pmatrix}$$

Et la matrice des covariances entre les composantes principales et les anciennes variables vaut :

$$\mathrm{Cov}(\mathbf{Z}, \mathbf{V}) = \frac{1}{n}\ {}^t\mathbf{X}_c\ \mathbf{X}_c\ \mathbf{A} = \mathbf{\Sigma}\ \mathbf{A} = \mathbf{A}\ \mathbf{\Lambda}$$

Si on remarque que la variance empirique d'une variable est égale au carré de la norme du vecteur qui la représente dans la géométrie euclidienne choisie et que le coefficient de

corrélation empirique de deux variables est égal au produit scalaire des deux vecteurs qui les représentent, on pourra interpréter les angles des vecteurs comme des corrélations.

10.1 Interprétation des axes en fonction des anciennes variables

On peut interpréter les axes principaux en fonction des anciennes variables. Une ancienne variable V_j expliquera d'autant mieux un axe principal qu'elle sera fortement corrélée avec la composante principale correspondant à cet axe.

10.2 Qualité de la représentation des variables

Pour les mêmes raisons qui ont poussé à se préoccuper de la qualité de la représentation des individus, il faut se préoccuper de la qualité de la représentation des variables sur un axe, un plan ou un sous-espace. Une variable sera d'autant mieux représentée sur un axe que sa corrélation avec la composante principale correspondante est en valeur absolue proche de 1. En effet, le coefficient de corrélation empirique entre une ancienne variable V_{cj} et une nouvelle variable Z_k n'est autre que le cosinus de l'angle du vecteur joignant l'origine au point v_j représentant la variable sur l'axe avec cet axe.

Une variable sera bien représentée sur un plan si elle est proche du bord du cercle des corrélations, car cela signifie que le cosinus de l'angle du vecteur joignant l'origine au point représentant la variable avec le plan est, en valeur absolue, proche de 1, etc.

10.3 Etude des liaisons entre les variables

Sur le graphique du cercle des corrélations, on peut aussi interpréter les positions des anciennes variables les unes par rapport aux autres en termes de corrélations. Deux points très proches du cercle des corrélations, donc bien représentées dans le plan, seront très corrélés positivement entre elles. Si elles sont proches du cercle, mais dans des positions symétriques par rapport à l'origine, elles seront très corrélées négativement.

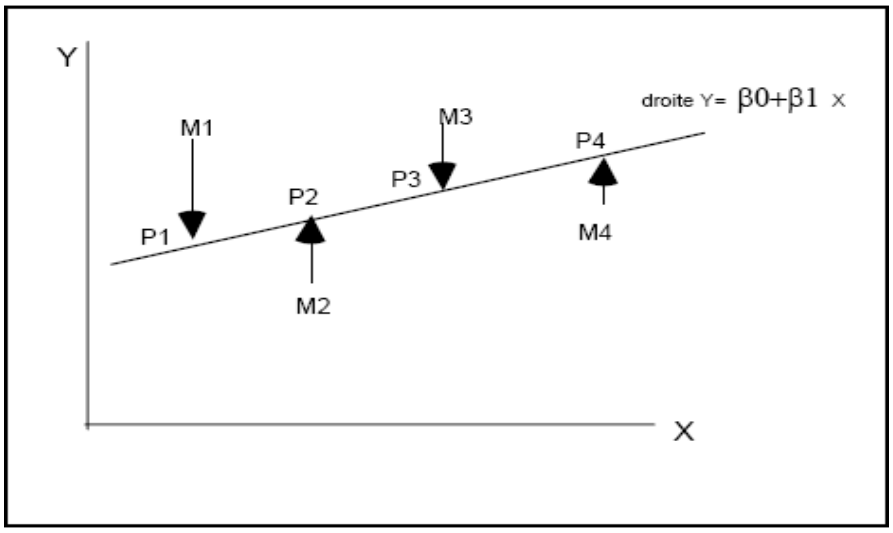
Deux variables proches du cercle des corrélations et dont les vecteurs qui les joignent à l'origine forment un angle droit, ne seront pas corrélées entre elles.

Il faut, pour interpréter correctement ces graphiques des cercles de corrélation, se souvenir qu'un coefficient de corrélation est une mesure de liaison linéaire entre deux variables, et qu'il peut arriver que deux variables très fortement liées aient un coefficient de corrélation nul ou très faible, si leur liaison n'est pas linéaire.

ANNEXE2 : Régression des moindres carrés ordinaires

La **régression linéaire simple** se classe parmi les méthodes d'analyses multivariées qui traitent des données quantitatives.
C'est une méthode d'investigation sur données d'observations, ou d'expérimentations, où l'objectif principal est de rechercher une liaison linéaire entre une variable Y quantitative et une ou plusieurs variables X également quantitatives.
C'est la méthode la plus utilisée pour deux raisons majeures :
• c'est une **méthode ancienne,**
• c'est l'**outil de base** de la plupart des modélisations plus sophistiquées comme la régression logistique, le modèle linéaire généralisé, les méthodes de traitement des séries temporelles, et surtout des modèles économétriques, etc.

Le but de cette méthode est de trouver la droite qui ajuste le mieux le nuage de points, c'est-à-dire trouver la droite qui passe « au plus près » de tous les points du nuage.



La **régression linéaire multiple** est une généralisation à p variables explicatives de la régression linéaire simple. Elle cherche a expliquer, avec plus de précision possible, les valeurs prises par Y_i dite variable endogène à partir d'une série de variables explicatives X_{i1}, …, X_{ip}. Le modèle théorique s'écrit :

$$Y_i = a_0 + a_1X_{i1} + a_2X_{i2} + \ldots + a_pX_{ip} + \epsilon_i, \quad i = 1, \ldots, n$$

Où ϵ_i est l'erreur du modèle. Elle exprime ou résume l'information manquante dans l'explication linéaire des valeurs de Y_i à partir des X_{i1}, …, X_{ip} (problème de spécifications, variables non prises en compte, etc.). Les coefficients a_0, a_1,…, a_p sont les paramètres à estimer.

1- **Estimation des paramètres**

Nous avons n observations (y_i, x_{i1},…, x_{ip}) avec i = 1, …, n qui sont des réalisations des variables aléatoires Y_i, X_{i1}, …, X_{ip} ; l'équation de régression s'écrit :

$Y_i = a_o + a_1x_{i1} + \ldots + a_px_{ip} + \epsilon_i$ $\quad$ i = 1, …, n

La problématique est donc :

- Estimer les paramètres ai en exploitant les observations ;
- Evaluer la précision de ces estimateurs ;
- Mesurer le pouvoir explicatif du modèle ;
- Evaluer l'influence des variables dans le modèle :
  - Globalement (les p variables en bloc)
  - Individuellement (chaque variable)
- Evaluer la qualité du modèle lors de la prédiction ;
- Détecter les observations qui peuvent influencer exagérément les résultats (points atypiques).

a- <u>Notation matricielle</u>

Comme souligné plus haut, chaque observation yi de Yi s'écrit sous la forme de combinaison linéaire des observations xi des variables explicatives Xi. Nous avons donc les équations :

$$\begin{cases} y_1 = a_0 + a_1x_{1,1} + \ldots + a_px_{1,p} + \epsilon_1 \\ y_2 = a_0 + a_1x_{2,1} + \ldots + a_px_{2,p} + \epsilon_2 \\ \ldots \\ y_1 = a_0 + a_1x_{n,1} + \ldots + a_px_{n,p} + \epsilon_n \end{cases}$$

Notation matricielle :

$$\begin{pmatrix} y1 \\ \vdots \\ Yn \end{pmatrix} = \begin{pmatrix} 1 & x_{1,1} \ldots & x_{1,p} \\ \vdots & \vdots & \vdots \\ 1 & x_{n,1} \ldots & x_{n,p} \end{pmatrix} \begin{pmatrix} a_0 \\ \vdots \\ a_p \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

Ou encore : y = X a + ϵ

Avec

- y de dimension (n,1)
- X de dimension (n, p+1)
- ϵ de dimension (n, 1)
- La première colonne sert à indiquer que nous procédons à une régression avec constante.

Comme dans chaque théorie ou méthode, des hypothèses ont été émises afin de déterminer les propriétés des estimateurs (biais, convergence) et leurs lois de distribution (pour les estimations par intervalle et les tests d'hypothèses).

H1 : $E(\epsilon) = 0 \Rightarrow E(\epsilon_i) = 0 \ \forall$ i : l'erreur est d'espérance nulle.

H2 : X est une matrice composée de variables certaines (non aléatoires).

$\Rightarrow$ $Cov(x_{it}, \epsilon_t) = 0$: pas de corrélation entre la variable explicative x_{it} et l'erreur ϵ_t.

H3 : Rg(X) = p et n >p : le nombre d'observations n doit être supérieur au nombre de variables explicatives k et il ne doit pas exister de colinéarité stricte des p variables explicatives.

H4 : $V_\epsilon = E[\epsilon - E(\epsilon))(\epsilon - E(\epsilon))'] = E(\epsilon\ \epsilon') = \sigma_\epsilon^2 I$ où V_ϵ est la matrice des variances-covariances des erreurs ϵ.

H5 : $\frac{X'X}{n}$ tend vers une matrice finie non singulière (inversible ou régulière).

b- <u>Estimateur des moindres carrés ordinaires (EMCO)</u>

Le principe des moindres carrés consiste à chercher des valeurs qui minimisent la somme des carrés des résidus.

Min $\sum_{i=1}^{n} e_i^2$ = min e'e = min (Y – Xâ)'(Y – Xâ) = min S

Où e' est la transposée du vecteur e.

La fonction S est minimale lorsque :

$$\frac{\partial S}{\partial \hat{a}} = -2X'Y + 2X'X\hat{a} = 0 \quad \Leftrightarrow \quad \boxed{\hat{a} = (X'X)^{-1}X'Y}$$

En effet, on a $S = (Y - X\hat{a})'(Y - X\hat{a}) = (Y' - \hat{a}'X')(Y - X\hat{a}) = Y'Y - Y'X\hat{a} - \hat{a}'X'Y + \hat{a}'X'X\hat{a}$
$S = Y'Y - (\hat{a}'X'Y)' - \hat{a}'X'Y + \hat{a}'X'X\hat{a}$
$S = Y'Y - 2\hat{a}'X'Y + \hat{a}'X'X\hat{a}$

Car la transposée d'un scalaire est un scalaire ($(\hat{a}'X'Y)' = \hat{a}'X'Y$)

En effet, on a $S = \sum_{t=1} e_t^2$ qui est un scalaire, donc $S = Y'Y - Y'X\hat{a} - \hat{a}'X'Y + \hat{a}'X'X\hat{a}$ est un scalaire avec $(Y'Y)_{(1,1)}$, $(Y'X\hat{a})_{(1,1)}$, $(\hat{a}'X'Y)_{(1,1)}$ et $(\hat{a}'X'X\hat{a})_{(1,1)}$.

On voit que l'on ne peut obtenir l'estimateur â de a que si (X'X) est inversible. Lorsqu'il y a colinéarité des variables explicatives, la matrice $(X'X)^{-1}$ n'est pas inversible !

Si les hypothèses initialees sont respectées, l'estimateur des moindres carrés ordinaires possède d'excéllentes propriétés à savoir :

- Les estimateurs sont sans biais : $E(\hat{a}) = a$;

- Les estimateurs sont convergents :

$$\lim_{n\to\infty} V_{\hat{a}} = 0$$

En effet, on a : $V_{\hat{a}} = \sigma_\varepsilon^2 (X'X)^{-1}$

On a donc que $\lim_{n\to\infty} V_{\hat{a}} = 0$ si l'hypothèse H5 est vérifiée : $\frac{X'X}{n}$ tend vers une matrice finie, définie positive et inversible lorsque n tend vers ∞.

On calcule $V_{\hat{a}} = \sigma_\epsilon^2 (X'X)^{-1}$ à l'aide de l'estimateur de σ_ϵ^2 qui s'écrit comme suit :

$$\sigma_\epsilon^2 = s^2 = \frac{SCR}{n-p} = \frac{\sum_{i=1}^{n} e_i^2}{n-p-1}$$

(Pour la démonstration voir Dormont, Introduction à l'économétrie, Montchrestien).

c- <u>Estimation de la variance des résidus</u>

Pour la variance des résidus observés : $\sigma_\varepsilon^2 \equiv \mathbf{Var}[\varepsilon]$ on peut utiliser l'estimateur sans biais construit à partir de la variance des résidus observés :

$$s^2 \equiv \hat{\sigma}_\varepsilon^2 = \frac{1}{n-p-1}\sum_{i=1}^{N}\hat{\varepsilon}_i^2$$

Les $\hat{\varepsilon}$ correspondent aux résidus observés : $\hat{\varepsilon} = Y - \hat{Y}$.

On remarque que l'estimateur classique de la variance est :

$$s_{n-1}^2 \equiv \hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2$$

On n'inclut pas l'espérance des résidus, car celle-ci est supposée être de zéro (selon H2). Surtout, les résidus du modèle ont exactement une moyenne de zéro lorsqu'une constante est introduite dans le modèle.

Il existe également un autre estimateur, obtenu par la méthode du maximum de vraisemblance, qui est cependant biaisé :

$$s^2 \equiv \hat{\sigma}_\varepsilon^2 = \frac{1}{n}\sum_{i=1}^{N}\hat{\varepsilon}_i^2$$

d- **<u>Estimation de la matrice de variance-covariance de â</u>**

Il suffit de remplacer la variance théorique des résidus, σ_ϵ^2, par son estimateur sans biais des moindres carrés :

$$s^2 \equiv \hat{\sigma}_\varepsilon^2 = \frac{1}{n-p-1}\sum_{i=1}^{N}\varepsilon_i^2$$

L'estimateur de la matrice de variance-covariance des résidus devient :

$$\hat{\mathrm{Var}}[\hat{a}] \equiv \hat{\Sigma}_{\hat{a}} = \hat{\sigma}_\varepsilon^2 (X'X)^{-1}$$

La variance estimée $\hat{\sigma}^2_{\hat{a}_j}$ de l'estimation du paramètre $\hat{a}_j$ est lue sur la diagonale principale de cette matrice.

2- Etude des coefficients

Après avoir obtenu l'estimateur, son espérance et une estimation de sa variance, il ne reste plus qu'à calculer sa loi de distribution pour produire une estimation par intervalle et réaliser des tests d'hypothèses.

En partant de l'hypothèse : $\epsilon_i \sim N(0, \sigma_\epsilon)$

Nous pouvons montrer que:

$$\frac{\hat{a}_j - a_j}{\sigma_{\hat{a}_j}} \sim N(0,1) \qquad \text{ET} \qquad (n-p-1)\frac{\hat{\sigma}^2_{\hat{a}_j}}{\sigma^2_{\hat{a}_j}} \sim \chi^2(n-p-1)$$

Le rapport d'une loi normale et de la racine carrée d'une loi du Chi carré normalisée par ses degrés de liberté aboutit à une loi de Student. Nous en déduisons donc que la statistique :

$$t = \frac{\hat{a}_j - a_j}{\hat{\sigma}_{\hat{a}_j}} \sim \mathrm{T}(n-p-1)$$

Suit une loi de student à (n-p-1) degrés de liberté.

A partir de ces informations, il est possible de calculer les intervalles de confiance des estimations des coefficients.

Il est également possible de procéder à des tests d'hypothèses, notamment les tests d'hypothèses de conformité à un standard. Parmi les différents tests possibles, le test de nullité du coefficient (H_0 : $a_j = 0$, contre H1 : $a_j \neq 0$) tient un rôle particulier : il permet de déterminer si la variable x_j joue un rôle significatif dans le modèle.

3- <u>Indice de qualité de la régression</u>

L'évaluation globale de la pertinence du modèle de prédiction s'appuie sur l'équation d'analyse de variance SCT = SCE + SCR, où

SCT est la somme des carrés totaux, elle traduit la variabilité totale de la variable endogène ; SCE est la somme des carrés expliqués, elle traduit la variabilité expliquée par le modèle et SCR est la somme des carrés résiduels, elle correspond à la variabilité non expliquée par le modèle.

Toutes ces informations sont résumées dans un tableau d'analyse de variance :

Source de variation	Somme des carrés	Degrés de liberté	Carrés moyens
Expliquée	$SCE = \sum_i (\hat{y}_i - \bar{y})^2$	***p***	$CME = \frac{SCE}{p}$
Résiduelle	$SCR = \sum_i (y_i - \hat{y}_i)^2$	***n - p - 1***	$CMR = \frac{SCR}{n-p-1}$
Totale	$SCT = \sum_i (y_i - \bar{y})^2$	***n - 1***	CMT

Dans le meilleur des cas, on a SCR = 0, le modèle arrive à prédire exactement toutes les valeurs de y à partir des valeurs des x_j. Dans le pire des cas, SCE=0, le meilleur prédicateur de y est sa moyenne $\bar{y}$.

Un indicateur spécifique permet de traduire la variance expliquée par le modèle, il s'agit du coefficient de détermination. Sa formule est la suivante :

$$R^2 = \frac{SCE}{SCT} = 1 - \frac{SCR}{SCT}$$

Nous avons forcément $0 \leq R^2 \leq 1$.

Enfin, si le R^2 est certes un indicateur pertinent, il a tendance à augmenter à mesure que l'on ajoute des variables dans le modèle. De ce fait, il est important si l'on veut comparer des modèles comportant un nombre différent de variables. Il est conseillé dans ce cas d'utiliser le coefficient de détermination ajusté qui est corrigé des degrés de liberté :

$$\bar{R}^2 = 1 - \frac{SCR/(n-p-1)}{SCT/(n-1)} = 1 - \frac{n-1}{n-p-1}(1-R^2)$$

4- <u>Significativité globale du modèle</u>

Le R^2 est un indicateur simple, on comprend aisément que plus il s'approche de 1, plus le modèle est intéressant. En revanche, il ne permet pas de savoir si le modèle est statistiquement pertinent pour expliquer les valeurs de y.

Nous devons nous tourner vers les tests d'hypothèses pour vérifier si la liaison mise en évidence avec la régression n'est pas un simple artefact.

La formulation du test d'hypothèse qui permet d'évaluer globalement le modèle est la suivante :

H0 : a1 = a2 = … = ap = 0 contre H1 : au moins un des coefficients est non nul.

La statistique dédiée à ce test s'appuie (parmi les différentes formulations possibles) sur le R^2, elle s'écrit :

$$F = \frac{\frac{R^2}{p}}{\frac{1 - R^2}{n - p - 1}}$$

Suit une loi de Fisher à (p, n-p-1) degrés de liberté.

La région critique du test est donc : rejet de Ho si et seulement si :

$$F_{calc} > F_{1-\alpha}(p, n - p - 1)$$

Où α est le risque de première espèce.

ANNEXE3 : ACP sorties

Tableau53 : Contribution des variables

Colonne1	F1	F2
Complet	13,731	1,168
literacy	12,689	1,166
internet	8,701	1,697
gdp	11,955	2,029
popgrowt	10,939	2,788
Corrupt	1,879	3,032
pop	1,081	0,129
ptr	11,773	0,013
repet	7,418	1,587
female	13,024	0,309
dur_gpe	2,355	27,874
expend	0,992	22,964
response	1,224	19,274
language	1,985	1,809
private	0,253	14,162

Tableau54 : Qualité de représentation des variables

Colonne1	F1	F2
Complet	**0,776**	0,018
literacy	**0,717**	0,018
internet	**0,492**	0,026
gdp	**0,676**	0,031
popgrowt	**0,618**	0,043
Corrupt	0,106	0,047
pop	0,061	0,002
ptr	**0,666**	0,000
repet	**0,419**	0,025
female	**0,736**	0,005
dur_gpe	0,133	**0,431**
expend	0,056	**0,355**
response	0,069	**0,298**
language	0,112	0,028
private	0,014	**0,219**

Tableau55 : Contribution des observations

Colonne1	F1	F2
BDI	1,868	0,204
CAF	4,712	0,058
COM	0,399	0,171
AGO	0,546	5,049
GAB	0,025	6,179
TGO	1,650	0,464
COG	0,651	3,604
TCD	3,235	0,130
CIV	0,906	0,011
CMR	0,959	0,000
MDG	1,375	0,150
LSO	0,003	4,392
MWI	1,657	0,136
NPL	0,964	0,001
LAO	0,149	0,006
BEN	2,052	0,077
SWZ	0,161	0,091
NAM	0,173	0,011
MLI	3,247	0,041
TMP	0,544	0,372
RWA	0,913	1,229
ZAR	1,709	2,901
ERI	1,359	0,125
CPV	0,616	0,692
UGA	1,171	0,227
VUT	0,051	0,759
GTM	0,007	0,718
IRQ	0,000	2,853
MAR	0,110	0,491
KHM	0,356	0,030
BGD	0,620	0,490
BFA	2,670	6,687
GIN	1,882	0,150
SEN	0,937	0,823
SLE	2,004	0,496
NIC	0,004	0,869
BLZ	0,298	2,662
DJI	0,383	1,403
THA	1,385	0,479
MOZ	2,240	2,100
ZAF	0,710	0,222
KIR	0,633	0,499

TUN	1,194	0,528
HND	0,148	0,456
BTN	0,000	0,001
SYR	0,414	0,514
ETH	2,343	0,448
SLV	0,599	0,012
ZMB	0,438	0,036
MRT	0,783	1,068
GHA	0,355	0,587
LBR	1,571	1,565
KEN	0,100	0,660
GMB	0,548	1,243
NER	2,435	7,595
TON	1,580	0,888
YEM	0,399	0,159
PRY	0,598	0,793
BWA	1,438	0,009
MDV	1,718	0,507
TZA	0,146	0,188
MUS	1,570	0,041
SDN	0,118	0,235
PAK	0,226	0,142
IDN	0,387	0,350
IND	0,020	1,581
EGY	0,391	0,222
NGA	0,045	0,603
LVA	6,999	2,569
PHL	0,318	0,693
FJI	0,540	3,869
ALB	1,681	0,167
ECU	0,809	2,125
BOL	0,716	0,818
GUY	1,278	3,889
WSM	1,405	0,589
VNM	1,290	2,155
ARM	2,305	0,324
AZE	1,854	0,545
GEO	2,535	0,428
KGZ	1,057	0,853
MDA	2,218	6,645
TJK	0,377	0,014
UKR	2,411	0,031
UZB	0,941	0,573
JOR	1,081	0,791
LKA	1,018	0,088

WBG	0,271	0,441
MNG	0,956	0,858
CUB	3,043	3,056

Tableau56 : Qualité de représentation des Observations :

Colonne1	F1	F2
BDI	**0,357**	0,011
CAF	**0,710**	0,002
COM	0,133	0,016
AGO	0,137	**0,346**
GAB	0,007	**0,458**
TGO	**0,505**	0,039
COG	0,212	**0,321**
TCD	**0,590**	0,006
CIV	**0,387**	0,001
CMR	**0,524**	0,000
MDG	**0,605**	0,018
LSO	0,001	**0,462**
MWI	**0,457**	0,010
NPL	**0,662**	0,000
LAO	0,156	0,002
BEN	**0,820**	0,008
SWZ	0,113	0,017
NAM	0,113	0,002
MLI	**0,790**	0,003
TMP	0,185	0,035
RWA	0,220	0,081
ZAR	**0,298**	0,139
ERI	**0,638**	0,016
CPV	0,200	0,061
UGA	**0,604**	0,032
VUT	0,028	0,114
GTM	0,011	**0,293**
IRQ	0,000	**0,358**
MAR	0,056	0,068
KHM	0,207	0,005
BGD	**0,262**	0,057
BFA	**0,484**	0,332
GIN	**0,649**	0,014
SEN	**0,610**	0,147
SLE	**0,650**	0,044
NIC	0,003	0,137
BLZ	0,056	0,138

DJI	0,192	0,192
THA	**0,501**	0,047
MOZ	**0,675**	0,173
ZAF	0,310	0,027
KIR	0,217	0,047
TUN	**0,457**	0,055
HND	0,083	0,070
BTN	0,000	0,000
SYR	0,153	0,052
ETH	**0,639**	0,033
SLV	0,284	0,002
ZMB	**0,254**	0,006
MRT	**0,350**	0,131
GHA	0,243	0,110
LBR	**0,344**	0,094
KEN	0,059	0,107
GMB	**0,244**	0,152
NER	**0,433**	0,370
TON	**0,423**	0,065
YEM	0,194	0,021
PRY	**0,504**	0,183
BWA	**0,361**	0,001
MDV	**0,533**	0,043
TZA	0,099	0,035
MUS	**0,420**	0,003
SDN	0,077	0,042
PAK	0,119	0,020
IDN	0,197	0,049
IND	0,007	0,149
EGY	**0,418**	0,065
NGA	0,025	0,093
LVA	**0,581**	0,058
PHL	0,164	0,098
FJI	0,089	0,175
ALB	**0,717**	0,019
ECU	**0,373**	0,268
BOL	**0,492**	0,154
GUY	**0,297**	0,247
WSM	**0,495**	0,057
VNM	**0,465**	0,213
ARM	**0,654**	0,025
AZE	**0,615**	0,049
GEO	**0,718**	0,033
KGZ	**0,402**	0,089
MDA	**0,386**	0,316

TJK	**0,246**	0,002
UKR	**0,578**	0,002
UZB	**0,420**	0,070
JOR	**0,386**	0,077
LKA	**0,608**	0,014
WBG	**0,190**	0,085
MNG	**0,481**	0,118
CUB	**0,503**	0,138

ANNEXE4 : Tableau des données

PAYS	abr	Comple	Primary	literac	interne	gdp	popgro	Corrupt	pop	ptr	repet	female	dur_gp	expenc	respons	languag	private	dif_comp
Lesotho	LSO	69,7	32,8	89,7	3	7,2	1	6,5	7,7	37,9	20,5	77,8	7	27,1	0,7	0,3	0,6	30,8
Malawi	MWI	56	37	73,7	1,6	6,6	3	7	9,6	80	19,7	40	3	14	0,4	0,5	1,1	2,1
Swaziland	SWZ	65,5	36,3	86,9	4,9	8,4	1,3	6,9	7,1	32,5	17	72,1	0	14,6	0,6	0,1	14,1	-1,3
Namibia	NAM	84,5	32,1	88,5	4,6	8,6	1,9	5,6	7,7	30,8	16,3	66,4	0	15,8	0,6	0,8	4,3	9,2
Eritrea	ERI	48		66,6	2,8	6,3	3,4	7,5	8,5	44,1	15,3	41,9	0	9,9	0,7	0,6	8,4	7,2
Uganda	UGA	58,4	39,3	71,4	4,4	6,9	3,3	7,6	10,4	49,5	12,9	38,8	1	8,8	0,7	0,9	10,4	0
Sierra Leo	SLE	60		37,9	0,2	6,5	3	7,5	8,7	60,5	10	38	5	7,1	0,4	0,8	14,1	-1,3
South Afri	ZAF	91,1	37,4	88,7	8,3	9,1	1	5,9	10,8	31,1	8	76,5	0	14,3	0,8	0,5	2,4	10,1
Ethiopia	ETH	47,3		32,9	0,3	6,6	2,3	7,3	11,3	61,2	6,9	38,5	8	12,4	0,5	0,9	5,7	0
Zambia	ZMB	85,2	34,5	70,9	4,4	7,1	2,6	6,8	9,5	62,7	6,4	48,3	4	5,2	0,6	0,9	2,9	9
Ghana	GHA	75,2	24,4	66,6	3,3	7,2	2,4	6,1	10,1	33	6,2	34,3	8	11,9	0,7	0,8	17,5	8,6
Liberia	LBR	60,6		56,5	0,4	5,9	4,4	6,8	8,3	21,8	6,2	19,4	5	5,7	0,4	0,9	29,8	-3,8
Kenya	KEN	88,8	40,6	87	6,8	7,3	2,6	7,8	10,6	44,6	5,8	44,4	7	23,3	0,6	0,9	8,1	1,3
Gambia, T	GMB	73		46,5	5,6	7,1	2,3	6,5	6,3	36,6	5,6	33,2	9	11,4	0,7	0,8	20,2	0
Botswana	BWA	96,3	37,4	84,1	4,7	9,4	1,4	3,9	7,6	25,5	4,7	78,2	0	14,1	0,6	0,5	5,2	-1,1
Tanzania	TZA	74		72,9	1,1	7	2	7	7,7	53,8	4,5	48,6	0	22,1	0,6	0,9	1,1	-1,6
Mauritius	MUS	92	40	87,9	18,8	9,3	0,7	4,9	7,2	21,8	4,2	64,7	0	10,8	0,8	0,6	25,9	-1,2
Nigeria	NGA	80,4	38,6	60,8	10,3	7,5	2,5	7,6	11,9	39,1	3	50,1	0	14	0,4	0,9	5,2	4,3
Iraq	IRQ	70,2		78,1	0,8	8	2,9	8,2	10,3	18,7	12,4	70,5	0	13,9	0,3	0,7	14,1	1,9
Morocco	MAR	80,7	42,9	54,5	22,2	8,3	1	6,6	10,4	26,9	12,3	47,4	0	16,8	0,7	0,5	7,7	9,6
Tunisia	TUN	97,1	45,9	75,9	18	8,9	1,1	6,2	9,2	18,7	7,4	52,9	0	21,1	0,6	0	1,4	-13,1
Syrian Ara	SYR	100		82,5	12,4	8,4	2,2	9,5	9,9	17,8	7,2	66	0	18,2	0,5	0,4	4,3	-5,2
Yemen, R	YEM	60,6		58,6	1,4	7,7	2,6	7,9	7,4	27,5	5	25	9	13,9	0,5	0,6	2,5	0,5
Sudan	SDN	51,6		70,2	6,5	7,5	2,5	8,4	10,7	34,1	3,8	64,7	0	13,9	0,5	0,8	4,7	1,7
Egypt, Ara	EGY	92,5		68,9	14	8,5	1,9	7,1	7,9	25,6	3,2	54,6	0	13,9	0,5	0,5	7,7	4,1
Jordan	JOR	96,1	31,6	91,7	19,6	8,5	3	5,5	8,7	21	0,9	70,7	0	13,6	0,5	0,5	31,4	5,1
West Ban	WBG	87,6		93,7	7,6	7,8	2,1	7,2	6,4	28,5	0,6	63	0	13,9	0,5	1	9,9	-13,5
Nepal	NPL	59,1		59,1	1,3	6,9	2	7,8	10,3	36,9	19,5	33,6	3	16,4	0,6	0,7	11,8	9
Lao PDR	LAO	72,3		72,7	2,1	7,6	1,7	7,8	6,8	30,9	18,2	46,5	4	10	0,6	0,7	2,6	1,3
Timor-Les	TMP	79,8		50,6	7,4	6,6	2,5	7,6	7	37,3	16	33,1	7	27,6	0,4	0,9	13,1	0
Cambodia	KHM	82,5		75,6	0,5	7,4	1,2	7,9	9,5	50,8	12	42,7	6	6	0,7	0,9	0,9	5,5
Banglades	BGD	58,4		55,9	0,3	7,2	1,2	7,3	11,9	45,8	11,9	40,1	0	10,3	0,7	0,4	41,6	17,4
Thailand	THA	100	55,6	93,5	18,5	8,9	0,8	6,6	11,1	17,3	9,2	59,8	0	19,6	0,4	0,7	16,9	-1,8
Bhutan	BTN	77,8		52,8	5,1	8,3	2,1	4,3	6,6	29,5	7,4	40,9	3	7,2	0,5	0,9	2,3	-0,4
Pakistan	PAK	60,6		53,2	9	7,7	1,8	7,5	12	39,4	3,6	46	0	13,9	0,7	0,8	33,5	0,2
Indonesia	IDN	100	46,3	91,5	5,4	8,1	1,1	7	12,4	19	3,5	57,9	0	13,3	0,6	0,8	16,6	-10,1
India	IND	88,8	39,6	62,8	3,6	7,9	1,5	6,9	14	40,2	3,4	44	0	9	0,4	0,9	14,1	3,1
Philippines	PHL	93,2	48,1	95,4	5,8	8,1	1,8	7,4	11,4	33,9	2,3	87,8	0	9	0,6	0,9	7,8	-0,3
Vietnam	VNM	100		92,8	18,7	7,8	1,1	7,1	8	20,7	1	77,9	9	19,6	0,5	0,2	0,6	0
Sri Lanka	LKA	100		90,7	4	8,3	1,1	6,7	9,9	23,1	0,9	82,7	0	13,9	0,5	0,5	1,8	0
Mongolia	MNG	100		97,5	11,3	8,1	1,5	7,3	7,9	31,9	0,6	94,9	6	14,6	0,7	0,3	4,3	0
Burundi	BDI	41,9		66,6	0,7	5,9	3	8,1	9	51,3	31,1	54,2	0	19,5	0,7	0	1,2	7,4
Central Afi	CAF	31,8		55,2	0,4	6,5	1,8	7,8	8,4	90,3	26,4	14	4	5,7	0,7	0,9	12,6	9,9
Comoros	COM	72,4		74,2	2,9	6,9	2,7	7,6	6,6	33,4	26,1	34,2	0	27,2	0,5	0,8	12	12,8
Angola	AGO	47		70	2,1	8,4	3,1	8	9,8	45,8	25	62,1	0	4,2	0,3	0,8	2	16,4
Gabon	GAB	70	40,5	85,8	5,3	9,5	1,9	7	7,3	36	24	44,7	0	4	0,3	0,3	29,3	0
Togo	TGO	66,1	40,2	56,9	5,6	6,8	2,2	7,6	8,7	38,9	23,3	12,1	2	9,5	0,8	0,9	42	11,7

Congo, R	COG	75		78,6	2,9	8,2	2,9	7,8	6,7	63,5	22,2	46,4	0	2,9	0,6	0,9	32	14,3
Chad	TCD	32,4	35,6	31	0,9	7,2	2,9	8	9,3	62	22	12,4	0	8,6	0,6	0	8	2,3
Cote d'Iv	CIV	44,8	39	55,3	2,4	7,4	2	7,8	7,4	42,8	21,3	23,5	2	16,6	0,5	0,2	10,9	5,3
Cameroon	CMR	60,8	51,1	70,7	2,5	7,6	2,2	7,5	9,9	47	21,3	42	6	6,8	0,8	0,6	22,8	15,8
Madagas	MDG	63,3	45	64,5	0,9	6,8	3	7	9,9	48,3	21,2	57,9	7	8,8	0,8	0,7	18,9	-2,2
Benin	BEN	60,2	34,8	41,7	1,7	7,2	3,1	7	9,1	46,3	17,1	18,3	5	12,1	0,7	0,9	11,3	11,7
Mali	MLI	52,5	36,9	26,2	0,9	6,8	3,1	7,2	9,6	51,9	16	26,9	6	11,7	0,8	0,9	38,2	5,3
Rwanda	RWA	46,8		70,7	2,1	6,8	2,7	5	9,2	66,7	15,8	53,4	6	8,4	0,7	0	1,7	16
Congo, D	ZAR	52,6		66,8	0,4	5,7	2,6	8	7,9	38,2	15,5	26,3	0	11,8	0,6	0,6	82,5	-0,8
Cape Ver	CPV	89,7		82,8	13,1	8	1	4,5	6,2	25	12,9	66,6	0	17,3	0,8	0,6	14,1	9,9
Burkina F	BFA	36,5	39,5	26,1	0,7	6,9	3	7	9,7	47,8	11,4	31,5	10	31,4	0,8	0,8	13,6	10
Guinea	GIN	57,7		39,5	0,7	6,9	1,9	7,9	9,2	44,3	11,3	25,8	10	7,2	0,6	0,8	24,3	7
Senegal	SEN	51,5	37	45,8	6,2	7,4	2,7	7,1	9,4	37,5	10,2	27	6	18,1	0,7	0,6	12,4	7,2
Djibouti	DJI	35,3		78,6	1,6	7,6	1,9	7	6,8	34,4	9,3	26,7	6	23,4	0,6	0,6	13,5	-0,3
Mozambiq	MOZ	47,8	39,4	55,1	1,3	6,6	2,5	7,3	10	63,9	8,9	33,1	9	14,8	0,7	0,9	1,9	0
Mauritania	MRT	53,5	30,3	57,5	1,3	7,5	2,6	7,6	8,1	39,8	6,4	32,7	10	9,7	0,7	0,2	8,4	-7,4
Niger	NER	35,2	33,7	28,7	0,4	6,4	3,6	7,5	9,6	40,7	5,2	41,4	10	29,3	0,8	0,6	3,9	3,7
Vanuatu	VUT	81,7		80,1	6,4	8,3	2,6	6,5	5,5	21,8	12,8	54,2	0	17,6	0,5	1	26,9	0,6
Kiribati	KIR	100		78,6	2,1	7,7	1,6	6,9	4,6	25,1	7,7	77,5	0	13,9	0,4	0	14,1	-22,8
Tonga	TON	100		99	6,4	8,3	0,6	6,9	4,6	21	5	63	0	9,8	0,4	0	8,8	-6,4
Maldives	MDV	100		98,4	15,4	8,5	1,4	7,5	5,7	16,6	4,6	69,7	0	21,9	0,7	0	1,6	-9,4
Fiji	FJI	96,6		78,6	10,4	8,4	0,8	7,2	6,7	27,4	2,1	56,3	0	17,5	0,5	0,6	99	0
Samoa	WSM	96,6		98,7	4,2	8,3	0,3	6,1	5,2	29,9	1	72,6	0	12	0,4	0,5	17	0
Guatemal	GTM	75,4	44,7	74,5	11,2	8,3	2,5	7,3	9,5	30,1	12,5	64,9	0	8,5	0,6	0,7	11,4	15,5
Nicaragua	NIC	73,6	46,2	78	2,9	7,8	1,3	7,5	8,7	32	10	76,5	10	9,6	0,6	0,1	15,1	-0,3
Belize	BLZ	100		78,6	9,4	8,7	2,1	7,2	5,7	22,9	9,8	71,9	0	14,4	0,7	0,8	87,1	-2,6
Honduras	HND	85,2	44,3	83,6	8,1	8,1	2	7,4	8,9	32	7,4	74,8	10	13,9	0,5	0,1	7,5	1,6
El Salvad	SLV	86,6	47,9	83,4	8,2	8,7	0,4	6,6	8,7	38,8	6,7	71,3	0	8,3	0,7	0	10,1	5,2
Paraguay	PRY	94,2	46,3	94,6	10,1	8,3	1,8	7,8	8,8	27,1	4,8	71,9	0	11,1	0,5	0,3	17,3	3,6
Ecuador	ECU	104,5	45,3	84,2	8,3	8,8	1,5	7,3	9,6	21,2	1,6	69,1	0	3,2	0,5	0,3	28,4	4,5
Bolivia	BOL	99,6	44,4	90,7	8,1	8,3	1,7	7,2	7	24,1	1,4	61,4	0	13,7	0,4	0,7	8,1	2
Guyana	GUY	100		78,6	24,4	7,9	0,2	7,5	6,6	26,3	1,1	87,3	10	11,2	0,7	0,1	2,3	6,1
Latvia	LVA	93,7	61	99,8	53,6	9,5	-0,5	5,8	7,7	11,7	2,9	95,4	0	27,7	0,6	0,6	1,1	1,4
Albania	ALB	92,1	42,7	95,9	16,2	8,8	0,4	6,9	8,1	21,6	1,8	78,8	6	16,3	0,4	0,6	4,4	-4,9
Armenia	ARM	96,4	57,6	99,5	5,8	8,5	0,1	7,4	8	20,5	1	99,2	0	14,3	0,4	0,2	1,4	1,5
Azerbaijan	AZE	93,4	44,8	99,5	21,5	8,8	1,3	7,6	9,1	12,2	1	86,3	0	5,6	0,6	0,5	14,1	2
Georgia	GEO	93,8	53,5	99,7	13,1	8,3	-0,7	5,9	8,4	8,6	1	85,7	5	14,5	0,5	0,6	5,7	-5,5
Kyrgyz Re	KGZ	94,6		99,2	16,7	7,6	0,9	7,9	6,8	24,2	1	96,8	6	13,6	0,6	0,7	1	12
Moldova	MDA	95,1	57,5	98,5	7,4	7,8	-1	7,1	6,6	16,8	1	97,5	7	36,4	0,6	0,6	0,9	-3,6
Tajikistan	TJK	98,5		99,7	5,1	7,4	1,2	7,7	8,8	22,5	1	64,8	7	13,6	0,5	0,5	14,1	-2,2
Ukraine	UKR	100	56,8	99,7	13,5	8,7	-0,7	7,7	10,7	16,8	1	98,8	0	13,9	0,6	0,9	0,5	-12,2
Uzbekista	UZB	96,1		99,3	7,6	7,7	1,1	8,4	10,2	18,7	1	85,5	0	13,9	0,5	0,4	14,1	-3,8
Cuba	CUB	93,1	59	99,8	11,3	9	0	5,8	9,3	9,8	0,6	77	0	40,7	0,6	0,3	14,1	0