
Mémoire de Projet de Fin d'étude

Analyse de la concurrence du Groupe OCP sur le marché des phosphates et produits dérivés

Soutenu par :

Hamza SEFIANE

Ayman YACHAOUI

Membres du jury :

M. Brahim AMRANI (Encadrant interne)

M. Said ACHCHAB (Examinateur)

Mme. Lamia BEN HIBA (President)

Mme. Fadila EL HILLALI (Encadrant externe)

Soutenu le 28 juin 2016

Année Universitaire : 2015-2016

REMERCIEMENTS

RÉSUMÉ

Mots-clés :

ABSTRACT

Keywords :

LISTE DES ABREVIATIONS

OLAP : OnLine Analytical Processing

AED : Analyse exploratoire des données

BI : Business Intelligence

CART : Classification And Regression Trees

CM : Commercial-Marketing

CRISP-DM : Cross-Industry Standard Process for Data Mining

DAP : DiAmmonium Phosphate

ENSIAS : École Nationale Supérieure d'Informatique et d'Analyse des Systèmes

FA : Forêts aléatoires

FAO : Food and Agriculture Organization of the United Nations

FSU : Former Soviet Union

IFA : International Fertilizer industry Association

MAP : MonoAmmonium Phosphate

NPK : Nitrogen-Phosphorus-Potassium

OCP : Office Chérifien des Phosphates

OOB : Out-of-Bag

PA : Acide phosphorique

PIB : Produit intérieur brut

TSP : Triple Super Phosphate

USA : United States of America

TABLE DES FIGURES

1.1	Filiales et coentreprises de l'OCP [19]	14
1.2	Le management senior de l'OCP [19]	16
1.3	Les six phases de la méthodologie CRISP-DM	20
1.4	Diagramme de GANTT	21
2.1	Description du processus d'analyse du marché	24
2.2	Architecture applicative de la solution proposée	31
2.3	Logos des outils utilisés	32
3.1	Lecture "machine" du .pdf de la "Trade Matrix" de la figure 3.2	34
3.2	Exemple d'une "Trade Matrix" ¹ dans le rapport trimestriel de l'IFA	34
3.3	Exemple d'une sortie de requête sur le Datamart OCP-CM	35
3.4	Exemple d'en-tête des tables du format DET des rapports IFA.	36
3.5	Exemple d'en-tête des tables du format AGG des rapports IFA.	36
3.6	Représentation de l'information transportée par la page .pdf de la figure 3.2 pour une résolution d'image de 700x800.	37
3.7	Classification de l'information transportée par la page .pdf de la figure 3.2 selon (\in zone de texte, \notin zone de texte)	38
3.8	Inférence des frontières de la table de la figure 3.2.	38
3.9	Flux de données du processus de consolidation.	39
3.10	Audit : Dimensions et données manquantes.	40
3.11	Audit : Interprétation des chaînes de caractères.	41
3.12	Audit : Conversion des unités de mesures.	42
3.13	Audit : Correction des unités de mesures.	42
3.14	Boites à moustaches des quantités logarithmiques exportées par produit phosphatés. Gauche : Quantités mondiales. Droite : Exports Marocains.	43
3.15	Histogramme des quantités logarithmiques exportées par produit.	44

1. Matrice des imports/exports entre les pays du monde, deux-à-deux, des phosphates et produits dérivés

3.16	Distributions des quantités logarithmiques mondiales exportées par produit et par exportateur majeur.	45
3.17	Parts de marché des produits phosphatés par région de destination et exportateur majeur	46
3.18	Sommes trimestrielles des exports par région de destination	47
3.19	Recherche aléatoire du paramètre k (<i>mtry</i>)	52
3.20	Recherche séquentielle du paramètre k (<i>mtry</i>)	53
3.21	Variation de l'importance des variables exogènes.	53
3.22	Classement par ordre d'importance des 24 premières variables	54
3.23	Graphe de l'évolution de l'erreur OOB dépendamment du nombre de variables exogènes intégrées.	54
4.1	Diagrammes des dispersions croisées, courbes splines et corrélations des variables exogènes	57
4.2	Carte coloriée des regroupement des variables corrélées	58
4.3	Valeurs propres de la décomposition spectrale de la matrice de corrélation	59
4.4	Graphe des éboulis de la décomposition spectrale de la matrice de corrélation	59
4.5	Inertie expliquée par les 5 premières CP	60
4.6	Corrélations variables-facteurs	60
4.7	Carte des individus	62
4.8	Diagrammes des dispersions croisées, courbes splines et corrélations des CPs	62
4.9	Régression linéaire de Cons.Fert sur les CP	63
4.10	Graphiques des résidus de la régression linéaire de Cons.Fert sur les CP	63
4.11	Validation du modèle et calcul du RMSE	64
A.1	Arbre CART	67

TABLE DES MATIÈRES

Résumé	1
Abstract	2
Introduction	9
1 Contexte général du projet	12
1.1 Présentation de l'OCP	13
1.1.1 Historique	13
1.1.2 Fiche signalétique	13
1.1.3 Chronologie des événements marquants de l'histoire du Groupe OCP. .	13
1.1.4 Filiales et partenariats	14
1.1.5 Les principales activités du Groupe OCP	14
1.1.6 Organisation du groupe OCP	15
1.1.6.1 Organigramme institutionnel	15
1.1.6.2 La direction CM : hôte de notre stage	16
1.2 Présentation du projet	16
1.2.1 Définitions et terminologie	16
1.2.2 Cadre général du projet	17
1.2.3 Motivation et problématique	18
1.2.3.1 Motivation	18
1.2.3.2 Problématique	18
1.3 Planification du projet	19
1.3.1 Les étapes CRISP-DM d'un projet Data Mining	19
1.3.2 Le planning du projet	21
1.3.2.1 Diagramme de GANTT	21
1.3.2.2 Démarche suivie dans notre projet :	22

2 Analyse et spécification	23
2.1 Analyse de l'existant	24
2.1.1 Description du processus d'analyse du marché	24
2.1.2 Critique de l'existant	24
2.2 Revue de littérature	25
2.2.1 Qu'est ce que le Data Mining ?	25
2.2.2 Les études de prévisions en matière de fertilisants	26
2.3 Compréhension du problème	28
2.3.1 Cadrage fonctionnel du projet	28
2.3.1.1 Les étapes d'une prévision en matière de fertilisants	28
2.3.1.2 Les modèles causaux pour la prévision	29
2.3.2 Cadrage technique du projet	30
2.3.2.1 Architecture globale de notre solution	30
2.3.2.2 Outils de réalisation	31
3 Collecte, compréhension et préparation des données.	33
3.1 Compréhension et préparation des données locales ²	34
3.1.1 Compréhension des données locales	34
3.1.2 Consolidation des données locales : Développement spécifique d'un ETL .	35
3.1.2.1 Conception fonctionnelle du processus de consolidation	35
3.1.2.2 Conception technique du processus de consolidation	37
3.1.2.3 Conception du flux de données du processus de consolidation .	39
3.1.3 Audit et description des données locales consolidées	40
3.1.3.1 Audit des données de consolidation	40
3.1.3.2 Description des données de consolidation	43
3.2 Collecte et préparation des données externes	47
3.2.1 Énumération et collecte des données externes	47
3.2.2 Préparation des données externes	50
3.2.2.1 Introduction à la selection de variables via forêts de décision aléatoires	50
3.2.2.2 Procédure de sélection des variables et élagage des données externes	51
3.2.2.3 Mise en œuvre de la sélection des variables et élagage	52
4 Analyse causale	55
4.1 Introduction à la régression en composantes principales	56
4.2 Mise en œuvre : corrélations croisées des variables exogènes	57
4.3 Mise en œuvre : calcul des composantes principales des données élaguées	59
4.4 Mise en œuvre : régression en composantes principales	62

2. Données disponibles au sein du portail Business Intelligence de l'OCP

Conclusion	64
A Sélection de variables via Forets Aléatoire	66
A.1 Introduction aux forêts de décision aléatoires	66
A.2 Utilisation de CART dans la construction des Forêts aléatoires	67
A.3 L'erreur Out-Of-Bag et le score FA d'importance des variables	68
A.4 Étude du biais et de variance d'un estimateur agrégé [14]	69
Bibliographie	70

INTRODUCTION

Agriculture is not crop production as popular belief holds - it's the production of food and fiber from the world's land and waters. Without agriculture it is not possible to have a city, stock market, banks, university, church or army. Agriculture is the foundation of civilization and any stable economy.

Allan Savory

Presque toutes les décisions prises par un gestionnaire ont besoin d'une prévision. S'il a une idée de ce qui se passera dans l'avenir, celui-ci peut prendre des décisions de gestion appropriées. Il a également besoin d'évaluer l'effet de ses décisions actuelles sur l'avenir afin que les bonnes décisions soient prises aujourd'hui pour créer une condition souhaitée demain.

Pour une organisation commercialisant et produisant les engrains, les estimations de la demande et des parts de marché sont indispensables pour les décisions stratégiques et celles concernant l'allocation des ressources. Les pays en développement, confrontés à des problèmes de faible productivité agricole, la croissance démographique et les besoins alimentaires, reconnaissent le rôle crucial des engrains au sein de leur politique de sécurité alimentaire. Les prévisions de la demande d'engrais à court terme et de la consommation potentielle à long terme sont essentielles pour la détermination des politiques appropriées en matière de production alimentaire et l'utilisation des engrais. De même, le mouvement mondial des engrais, leurs tendances des prix et des investissements dans de nouvelles installations de production sont influencés par les attentes de la demande.

La puissance de calcul mondiale augmentant de façon exponentielle, notre capacité à rassembler, stocker et analyser les données augmente également de façon spectaculaire. Les données

sont plus abondantes dans presque toutes les industries et applications académiques. L'industrie agricole ne fait pas l'exception. Dans un certain sens, il y a plus de données au sein de l'industrie agricole que dans la plupart des autres.[9].

Les informations et les connaissances acquises par les entreprises de la technologie agricoles sont généralement brevetées, surveillées, et utilisées à des fins de concurrence dans le marché. Certains pensent que de meilleures décisions en matière de chaîne d'exploitation et d'approvisionnement pourraient être établies par la compilation minutieuse et l'analyse de segments particuliers de ces données. L'OCP³ rejoint ce sentiment.

Dans le cadre de notre projet de fin d'études la mission - en un premier lieu - de mettre en place un mécanisme d'extraction, de transformation et de chargements de données non structurées. La nature non-structurée de celles-ci interdit tout approche BI classique en matière d'intégration des données et a nécessité de notre part un développement spécifique d'un moteur de Parsing qui automatisera les tâches de structuration et d'extraction. En un second lieu, notre encadrement à l'OCP nous a confié l'accès à l'ensemble des données en la possession du département hôte de notre stage avec la charge de réunir des éléments d'informations décisifs à la concurrence du groupe OCP. Nous nous sommes fixés le but de construire des modèles statistiques de prévision quant à la demande des produits dérivés phosphatés. La perspective finale est d'intégrer ces modèles au sein d'un module logiciel de prévisions dispensant des projections de marché à la demande, réalisant ainsi un moteur de prévision.

Ce mémoire présente les différentes étapes de réalisation de notre projet. Il se compose de quatre chapitres dont la description est comme suit. Le premier chapitre aborde le contexte général du projet. Il présente l'organisme d'accueil, la direction commerciale et dévoile la motivation et les objectifs du projet. Ensuite, on passe à la démarche suivie avant de terminer avec le planning. Le deuxième chapitre est dédié à l'analyse et spécifications des besoins. Ce qui se traduit par une étude de l'existant, ses limites et le recensement des différents besoins de la direction commerciale, pour aboutir à un cadrage fonctionnel et technique de la solution proposée. Nous procédons lors du troisième chapitre à concevoir le moteur d'extraction (notre ETL développé par nous mêmes), testons sa robustesse à travers un audit de la qualité des données qu'il transforme. Le chapitre 3 sera également pour nous l'occasion de nous placer dans la perspective de notre problématique et enrichir les données disponibles à l'OCP pour permettre une analyse causale de la demande en fertilisants dont nous construisant le modèle au chapitre quatre. Pour clore ce mémoire, nous résumons les résultats et les acquis réalisés lors de notre stage, les difficultés rencontrées et les perspectives de ce travail.

3. Office Chérifien des Phosphates

CHAPITRE 1

CONTEXTE GÉNÉRAL DU PROJET

SpaceX is a flat organization. Anyone gets to talk to anyone, and the best idea wins - even if it comes from an intern.

Gwynne Shotwell

Ce chapitre contextualise le projet dans son environnement, en présentant en premier lieu l'organisme hôte de celui-ci – L'Office Chérifien des Phosphates – avant de motiver les raisons inhérentes à son implémentation pour finir sur la planification du déroulement de l'analyse, la conception et la réalisation du projet.

1.1 Présentation de l'OCP

1.1.1 Historique

En 1920, alors que partout dans le monde, les compagnies minières fouillent fébrilement le sous-sol à la recherche du phosphate, minerai aux précieuses vertus fertilisantes, l'Office Chérifien des Phosphates (OCP S.A. depuis 2008) voit le jour.

En 1965, avec la mise en service du Maroc Chimie à Safi, le groupe devient également exportateur de produits dérivés. En 1998, il franchit une nouvelle étape en lançant la fabrication et l'exportation d'acide phosphorique purifié.

Parallèlement, de nombreux partenariats sont développés avec des opérateurs industriels du secteur, au Maroc et à l'étranger.

1.1.2 Fiche signalétique

Nomination sociale : Groupe Office Chérifien des Phosphates

Date de création : 1920

Siège social : 2-4, rue Al Abtal, Hay Erraha, 20200 Casablanca

Capital social : 8287 M MAD (2013)

Effectif employé : 23,000 (2013)

Site web : www.ocpgroup.ma

1.1.3 Chronologie des événements marquants de l'histoire du Groupe OCP.

1920 : Création, le 7 août, de l'office chérifien des Phosphates (OCP).

1959 : Création de la société marocaine d'études spécialisées et industrielles (SMESI).

1965 : Création de la société Maroc Chimie.

1974 : Lancement des travaux pour la réalisation du centre minier de Benguérir, en mai.

1975 : Création du Groupe OCP avec l'intégration des industries chimiques aux

1998 : Le Groupe OCP obtient le Prix National de la Qualité.

2003 : L'OCP est devenu le seul actionnaire de Phosboucraâ.

2008 : La société anonyme OCP SA est née le 22 janvier - Démarrage de Pakistan Maroc

2009 : Démarrage de Bunge Maroc Phosphore à Jorf Lasfar (BMP).

2010 : Création de JESA, joint-venture sous forme de partenariat en ingénierie

2012 : Creation de la JV BSFT (Black Sea Fertilizer Trading Company)

2013 : Signature d'une joint-venture avec Dupont

2014 : Inauguration du SLURRY PIPELINE entre Khouribga et Jorf Lasfar

1.1.4 Filiales et partenariats

L'OCP se structure en quatre filières chacune se focalisant sur un segment du groupe OCP et ayant co-créé plusieurs coentreprises. Ces partenariats ont été établis avec des clients du Groupe OCP. Ceci est le fruit d'une coopération qui touche aussi bien les accords de livraison à moyen et long terme que la construction d'unités de production. Dans cette optique, des unités basées au Maroc et à l'étranger sont en exploitation en joint-venture avec plusieurs partenaires dont la figure 1.1 fait la liste :

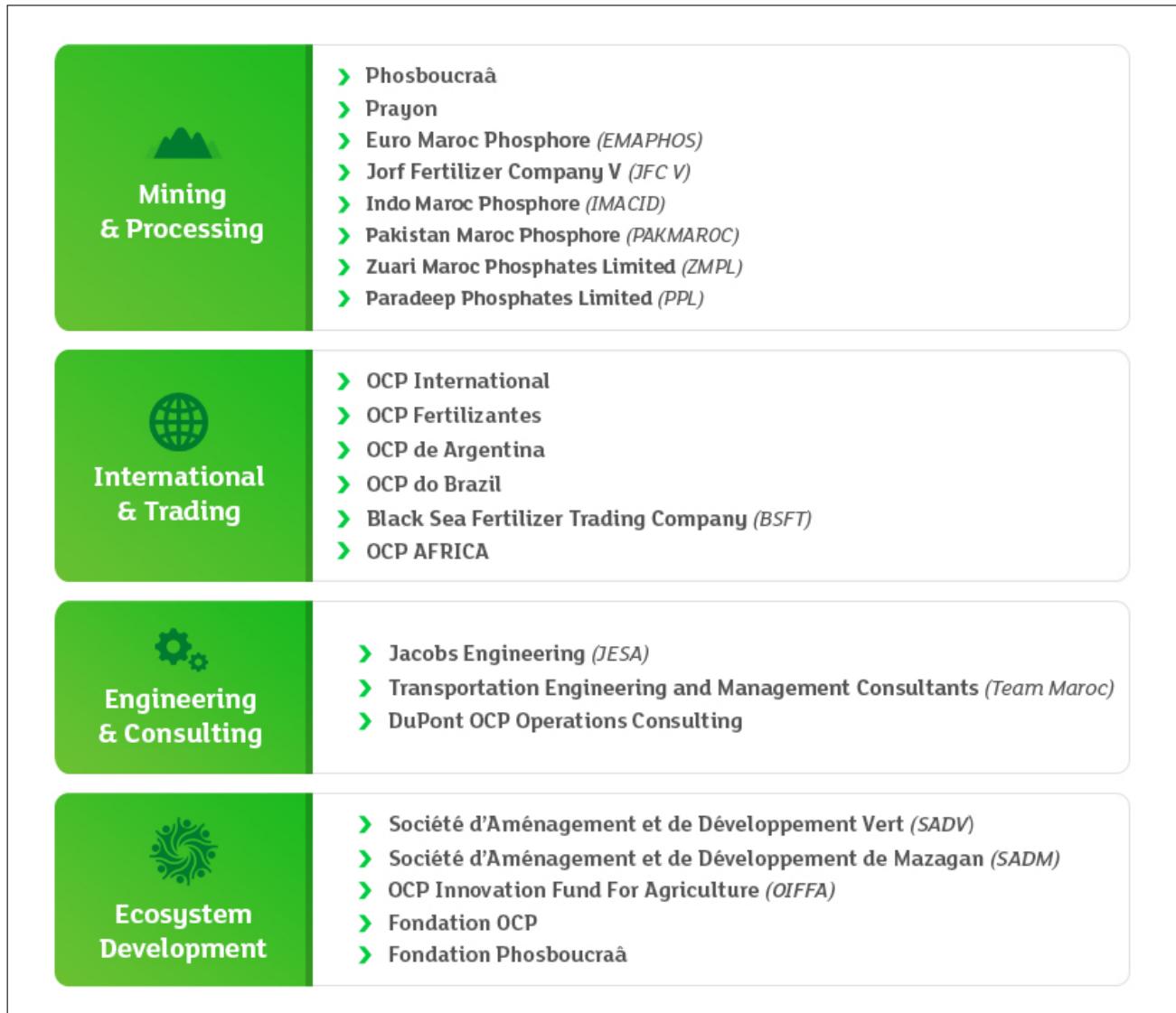


FIGURE 1.1 – Filiales et coentreprises de l'OCP [19]

1.1.5 Les principales activités du Groupe OCP

Les phosphates marocains sont exploités dans le cadre d'un monopole d'État confié à l'OCP. Le Groupe OCP a pour mission l'extraction, la valorisation et la commercialisation de phosphate et de ses produits dérivés. Chaque année, plus de 24 millions de tonnes de minéraux sont

extraites du sous-sol marocain qui recèle les trois quarts des réserves mondiales¹. Le phosphate brut provient des sites de Khouribga, Benguerir, Youssoufia et Boucraâ-Laâyoune. Généralement, le minerai subit une ou plusieurs opérations de traitement (criblage, séchage, calcination, flottation, enrichissement à sec ...). Celui-ci est ensuite exporté à l'état brut soit livré aux industries chimiques du groupe (à Jorf Lasfar ou à Safi) pour être transformé en produits dérivés commercialisables : acide phosphorique de base, acide phosphorique purifié et autres engrains solides, une large gamme de produits répondant à différents besoins, lui permettant de diversifier son portefeuille de clients et de faire face aux évolutions du marché [1].

Minerai de phosphate

OCP est le premier exportateur mondial de phosphate brut avec 33 % de parts de marché.

Acide phosphorique

Produit intermédiaire entre le minerai et les engrains, l'acide phosphorique est en fait le fruit d'un enrichissement de la roche obtenu par réaction avec un ensemble d'acides différents à concentrations distinctes. L'acide phosphorique purifié est produit en moindres quantités destinées à des applications alimentaires et industrielles.

OCP est le premier exportateur mondial d'acide phosphorique avec 46 % de parts de marché.

Engrais phosphatés

- Le MAP est un engrais binaire composé de deux éléments fertilisants : le phosphore et l'azote.
- Le DAP est un engrais tertiaire composé de deux éléments fertilisants : le phosphore et l'azote, engrais le plus répandu.
- Le TSP est un engrais entièrement phosphaté.
- Le NPK est un engrais ternaire composé de trois éléments : phosphore, azote et potassium.

L'influence du Groupe OCP vient de l'exportation de plus de 95% de sa production en dehors des frontières nationales. Opérateur international, il est présent sur les cinq continents. Il occupe une place très importante dans l'économie nationale puisque ses exportations ont totalisé plus que 20% des exportations marocaines. Il a ainsi contribué en 2013 à environ 4,3% du PIB [20].

1.1.6 Organisation du groupe OCP

1.1.6.1 Organigramme institutionnel

Présidé par le directeur général, le comité de direction exécutif compte sept directeurs exécutifs. Ils sont respectivement en charge de la direction du capital humain, du pôle industriel axe nord, du pôle industriel axe centre, du pôle commercial, du pôle finances et contrôle de gestion, du pôle stratégie & corporate development² et enfin du pôle juridique.

1. Connue à ce jour (rapport US Geological Survey, 2011)

2. Développement entreprise

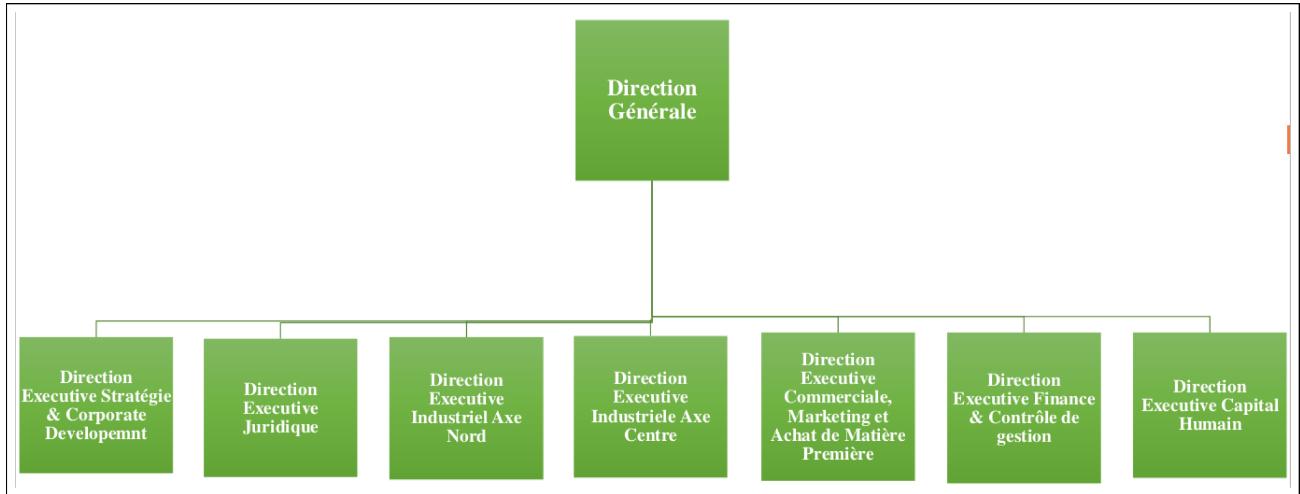


FIGURE 1.2 – Le management senior de l’OCP [19]

1.1.6.2 La direction CM : hôte de notre stage

La direction commerciale, où nous avons effectué notre stage, joue un rôle central dans le dispositif mis en avant par la nouvelle stratégie de l’OCP. Elle est composée de six directions : direction des ventes (découpée en quatre régions), direction des opérations, direction marketing, direction logistique et maritime, direction "Business Developement"³ et la direction "Procurement"⁴.

1.2 Présentation du projet

1.2.1 Définitions et terminologie

Avant de procéder plus en avant, il est nécessaire de définir certains termes couramment utilisés, mais ne possédant pas la même signification pour tout le monde. Dans la pratique, des termes comme «potentiel», «marché», «demande», et «ventes», sont utilisés de façon interchangeable. Aux besoins du présent rapport, nous allons en premier lieu définir ces termes pour nous assurer de la clarté de leurs utilisations ci-après.

Le terme «potentiel» réfère à la quantité d’engrais qu’un pays ou une région de pays peut consommer dans des conditions optimales. L’ensemble de la superficie cultivée dans diverses cultures, multiplié par la dose d’engrais recommandée pour chaque culture est assimilable à la consommation d’engrais potentiel pour un pays.

Il est évident que l’intérêt pour un engrais ou la possibilité de l’utiliser ne suffit pas d’elle-même pour expliquer la consommation. Les agriculteurs doivent d’abord avoir la conviction que par l’utilisation d’engrais, ils peuvent augmenter leurs revenus. Même si tel est le cas, ceux-ci

3. Développement des affaires

4. Achat de matières premières

ne seront pas forcément en mesure d'acheter des engrais. Cela dépend des liquidités financières dont ils disposent lors du besoin de l'engrais.

Les agriculteurs s'attendent à ce que l'engrais dont ils ont besoin soit disponible en temps opportun. «L'accès» fait référence à la capacité du réseau de distribution de rendre le bon type d'engrais à disposition de l'agriculteur dans le temps et aussi facilement que possible. En fonction de ces facteurs, il existe un «marché» disponibles ou «demande». **La prévision de la demande est la quantité d'engrais susceptible d'être vendue sur une certaine période définie pour tout un pays ou région de pays.** Les termes «marché» et «demande» sont utilisés de manière interchangeable. Alors que le «marché» ou «demande» se réfèrent aux ventes de toutes les entreprises dans le pays, le terme «ventes» fait référence à une seule entreprise. Les ventes d'engrais d'une entreprise dépend ainsi de sa part de marché.

Les prévisions à court et à long terme servent à des fins différentes. Les prévisions à court terme sont concernées par la prochaine saison ou année. Les projections à court terme sont nécessaires pour organiser la production, l'approvisionnement en matériaux brutes, finis et d'emballage, le stockage, le transport et les fonds de roulement de sorte que la demande prévue est assurée avec succès. Les projections à long terme de plus de quatre à cinq ans sont utilisés pour les décisions de politique et d'investissement. L'investissement dans de nouvelles installations de production et de recherche, la formulation des plans de développement, le renforcement de l'établissement de crédit, etc., sont des exemples de décisions régies par les prévisions à long terme.

1.2.2 Cadre général du projet

L'OCP a itérativement modernisé ces processus d'organisation et de management dans la portée de promouvoir sa productivité et assurer le développement et la croissance continue par rapport à la concurrence. Une modernisation qui ne peut se passer d'un accompagnement informationnel performant.

De grandes avancées sont à noter, principalement dans le domaine de la Business Intelligence. Ceci a débuté par l'informatisation de ces procédés pour garantir une optimisation au niveau des durées de traitements que subit l'information et les fiabiliser. D'abord par la mise en place d'un portail décisionnel, à base de fichiers .pdf dédié à la direction commerciale avant de se poursuivre en l'optimisation du temps d'accès à travers l'automatisation des processus de gestion de données, de réception des mails et extraction de données, mais aussi par le type des rapports (paramétrables interactifs, structures de coût) et leur publication dans un portail Web dédié.

Ces données proviennent des organismes indépendants spécialisés dans l'analyse de marché. Hebdomadairement, la direction commerciale reçoit plus de 100 publications de ces organismes qui contiennent des guides de prix, des évaluations ou même des prévisions concernant le marché de phosphates, ses dérivées et les matières premières.

Durant notre familiarisation préliminaire avec l'existant, nous avons relevé plusieurs discrepancies entre cet idéal informationnel voulu et la réalisation sur le terrain de cette vision, que nous détaillerons dans le second chapitre de ce rapport.

1.2.3 Motivation et problématique

1.2.3.1 Motivation

Parce que la demande d'engrais dépend d'une variété de facteurs agronomiques, celle-ci n'est pas stable, ni est-elle sujette à des prédictions exactes. Le choix des méthodes de prévision est donc particulièrement important, tant pour la gestion efficiente des compagnies productrices des engrais que pour la formulation de politiques appropriées par les gouvernements. La prévision efficace de la demande peut permettre aux exportateurs de tirer pleinement parti des fluctuations des cours mondiaux du marché. Le stockage requis, le transport, les ressources humaines à mobiliser, les arrangements financiers de crédits et de devises étrangères sont tributaires de la demande.

Considérant que les engrais produits mais non vendus peuvent être conservés pendant un an avant de trouver un acheteur et qu'une durée de stockage d'une année peut causer des pertes en quantité et en qualité conséquentes, l'importance de la prévision de la demande peut être facilement appréciée. Si la demande réelle est plus grande que prévue, ce ci conduit à des pénuries, une production agricole inférieure et, souvent, à des implications politiques pour les pays importateurs.

Tous les plans des entreprises fabriquant ou commercialisant des engrais devraient être dérivés directement ou indirectement de la prévision de la demande. A partir d'une prévision internationale de la demande, les ventes attendues d'une entreprise peuvent être estimées en évaluant sa part de marché dans chaque région du pays.

La prévision des ventes servira de consigne à l'égard du département de production quant à quoi, quand et combien produire. Le département financier de l'OCP, lui, est ainsi en mesure de préparer, sur la base des prévisions de ventes, un plan d'entrées et sorties de fonds, évaluer l'écart entre les fonds de roulement et d'organiser le soutien nécessaire de la banque. Le département marketing est guidé par la prévision dans le déploiement du personnel de vente, alors que le logistique sera éclairé quant à l'organisation du stockage à des endroits appropriés, les contrats de transport de marchandises pour faire face au volume prévu de l'entreprise.

1.2.3.2 Problématique

Au fil des années, la direction commerciale-marketing où nous avons effectué notre stage de fin d'études, a itérativement modernisé ses outils de traitement d'information. Dernière contribution en date de l'ENSIAS à cet essor informationnel, celle de Mr NACER [1], a consisté en la mise en place d'une solution adéquate basée sur les concepts et technologies du BI ayant

permis l'automatisation de la réception, de la normalisation et l'exploitation de tableau de bord post-analyse des données recueillies. Celle-ci se définissant comme une compléction du travail précurseur de Mr CHEMLAL [20]. La plate-forme ainsi réalisée par nos prédecesseurs est ainsi un socle pour aider les décisionnaires CM à mieux appréhender les évolution et historiques des marchés mondiaux.

Nous proposons d'emmener les perspectives informationnelles du département CM au-delà du reporting⁵ continu vers des mécanismes décisionnels et prévisionnels mettant en œuvre les techniques à l'état de l'art de fouilles de données (cf def data mining) . Nous souhaitons ainsi 'miner' des relations intéressantes régissant le marché international des phosphates qui ne sauraient être mises en relief par les techniques de la BI à savoir la segmentation en faits, dimensions et mesures. Notre problématique s'articule ainsi :

Quelles mécanismes prévisionnels de la demande en fertilisants phosphatés peuvent être établis à la suite d'une recherche de structure précédemment inconnue dans le marché international des phosphates ?

Répondre d'une manière exhaustive à cette question constituerait un atout clé en faveur d'une meilleure compétitivité de l'OCP et nous nous proposons d'explorer divers piste dont le présent rapport témoigne.

1.3 Planification du projet

1.3.1 Les étapes CRISP-DM d'un projet Data Mining

CRISP-DM, qui signifie Cross-Industry Standard Process for Data Mining⁶, est un modèle de processus d'exploration de données qui décrit les approches couramment utilisées par les experts pour résoudre des problèmes d'exploration de données [4].

Des sondages effectués en 2002, 2004, 2007 et 2014 montrent qu'il s'agit de la méthode principale utilisée par les data miners [18].

CRISP-DM découpe le processus de data mining en six phases principales. La séquence des phases n'est pas stricte et un va et vient entre les différentes phases est toujours nécessaire. Les flèches dans le diagramme du processus indiquent les dépendances les plus importantes et fréquentes entre les phases. Le cercle extérieur dans le diagramme symbolise la nature cyclique de l'exploration de données. Un projet de fouille de données se poursuit après qu'une solution ait été déployée. Les leçons apprises au cours du processus peuvent déclencher de nouvelles questions métier, souvent plus ciblées et font bénéficier les processus d'extraction de données ultérieures de l'expérience des précédents. Les six phases sont résumés dans la figure 1.3.

5. Tableaux de bord et rapports Business Intelligence

6. Standard inter-entreprises du processus de fouille de données.

Compréhension du problème : Cette première phase se concentre sur la compréhension des objectifs et des exigences du projet à partir d'un point de vue commercial, puis convertir ces connaissances en une définition de problème d'extraction de données, et un plan préliminaire conçu pour atteindre les objectifs désirés.

Compréhension des données : La phase de compréhension des données commence par une collecte de données initiale et procède à des activités dans le but de se familiariser avec les données, pour identifier les problèmes de qualité des données, pour découvrir un premier aperçu de celles-ci, ou pour détecter des sous-ensembles intéressants pour former des hypothèses pour des informations cachées.

Préparation des données : La phase de préparation des données couvre toutes les activités pour construire l'ensemble de données final (celui qui sera fourni en entrée aux algorithmes de modélisation) à partir des données brutes initiales. Cette tâche est susceptible d'être effectuée plusieurs fois, et pas dans un ordre prescrit. Celle-ci comprend la sélection des tables, des enregistrements, et des attributs d'intérêt ainsi que leur transformation et nettoyage.

Modélisation des Données : Dans cette phase, les différentes techniques de modélisation sont choisies et appliquées, leurs paramètres sont étalonnés à des valeurs optimales. En règle générale, il existe plusieurs techniques pour le même type de problème d'exploration de données. Certaines techniques ont des exigences spécifiques sur la forme de données. Par conséquent, un retour à la phase de préparation des données est souvent nécessaire.

Évaluation : A ce stade du projet, un modèle qui semble avoir de bonnes propriétés, du point de vue de l'analyse des données a été construit. Avant de procéder au déploiement final du modèle, il est important d'évaluer de manière plus approfondie le modèle, et d'examiner les étapes exécutées pour la construction de celui-ci, pour être certain qu'il répond correctement aux objectifs métiers précédemment arrêtés.

Déploiement : La création du modèle est généralement pas la fin du projet. Même si le but est d'accroître les connaissances des données, l'acquis doit être organisé et présenté d'une manière qui est utile pour le commanditaire. Selon les besoins, la phase de déploiement peut être une simple génération de rapport aussi bien qu'une mise en œuvre de mécanismes prévisionnels.

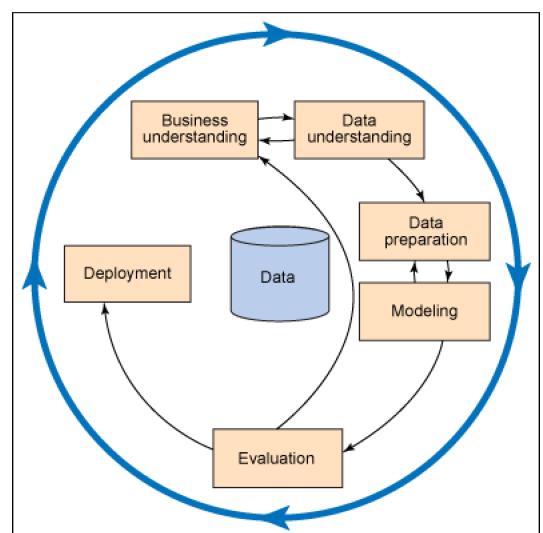


FIGURE 1.3 – Les six phases de la méthodologie CRISP-DM

1.3.2 Le planning du projet

1.3.2.1 Diagramme de GANTT

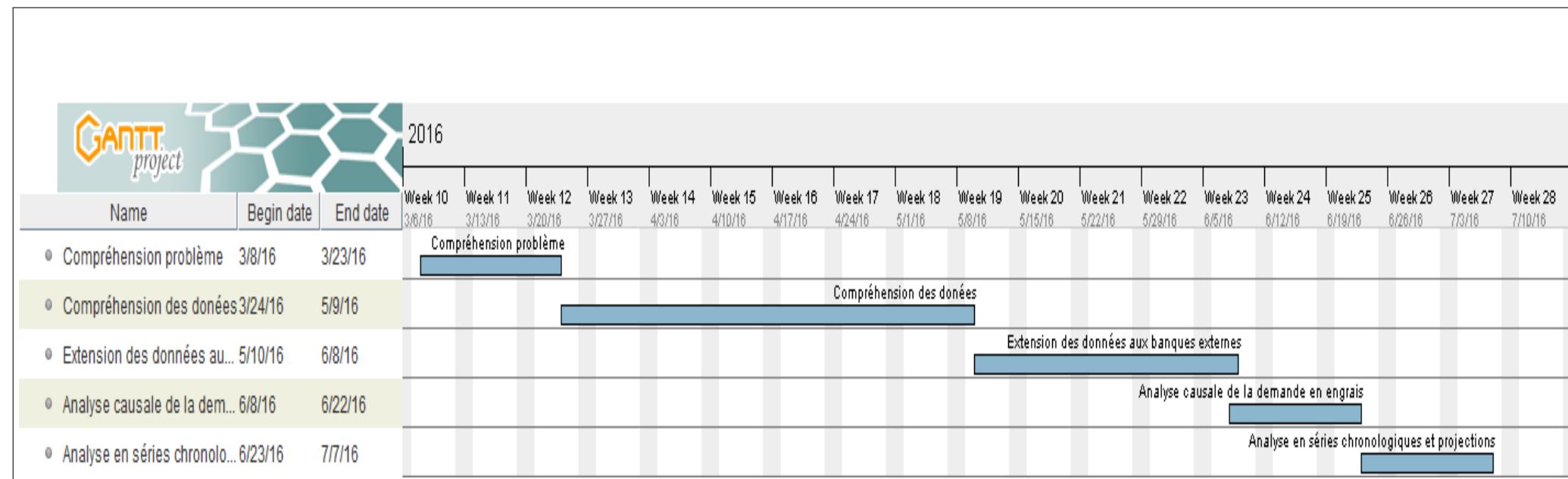


FIGURE 1.4 – Diagramme de GANTT

1.3.2.2 Démarche suivie dans notre projet :

Le projet est décomposé en plusieurs lots :

Lot I : Compréhension du problème.

- Comprendre le métier de la cellule Business intelligence de la direction commerciale au sein du groupe OCP.
- Détermination des objectives du projet Data Mining.
- Détermination des métriques pour l'évaluation de la réussite du projet.

Lot II : Compréhension des données.

- Description et familiarisation avec les différentes sources de données au sein de l'OCP.
- Parsing⁷ des données non structurées concernant les imports et exports des différents produits phosphatés.
- Audit de la qualité du résultat du Parsing et analyse préliminaire.

Lot III : Extension des données aux banques externes.

- Recherche et énumération des différents banques de données Web.
- Extraction et formatage des données d'extension.
- Classement des variables par ordre d'importance et sélection.

Lot IV : Analyse causale de la demande en engrais.

- L'analyse en composantes principales des différentes données enrichies du lot III.
- Sélection des premières composantes principales les plus représentatives.
- Régression sur les composantes retenues et interprétation.

7. Anglais pour "analyse syntaxique" : processus d'analyse d'une chaîne de symboles soit en langage naturel ou machine suivant les règles d'une grammaire formelle.

CHAPITRE 2

ANALYSE ET SPÉCIFICATION

If a student takes the whole series of my folklore courses including the graduate seminars, he or she should learn something about fieldwork, something about bibliography, something about how to carry out library research, and something about how to publish that research.

Alan Dundes

Ce chapitre procède à une analyse de l'existant au sein du département commercial-marketing au début de notre stage et y relève des points de critiques auxquels nous proposons une architecture applicative en guise de solution. Ceci après avoir fait un tour d'horizon des recherches scientifiques ayant adressé des problématiques similaires pour inspirer le cadrage fonctionnel et technique de notre projet.

2.1 Analyse de l'existant

2.1.1 Description du processus d'analyse du marché

L'utilisation des données au sein de la direction commerciale est basée sur le partage d'information concernant les marchés cibles de l'OCP. La répartition des chargés d'études de marché selon les continents du globe crée un besoin spécifique pour chaque analyste marché. Pour déterminer les prix de vente ou négocier les prix d'achat des matières premières plusieurs réunions quotidiennes sont indispensables. Lors de ces réunions des perspectives sur le marché sont partagés entre les analystes OCP à la base de leur revue des données qui leur parviennent par les prestataires d'informations et les solution de Business Intelligence mises en place par nos prédecesseurs([1, 20]). L'objectif de notre stage est d'emmener les capacités Business Intelligence du simple reporting à la proposition de mécanismes prévisionnels fluidifiant le processus d'arbitrage quant aux volumes et prix de production. Nous résumons ce processus par la figure 2.1 ci-dessous.

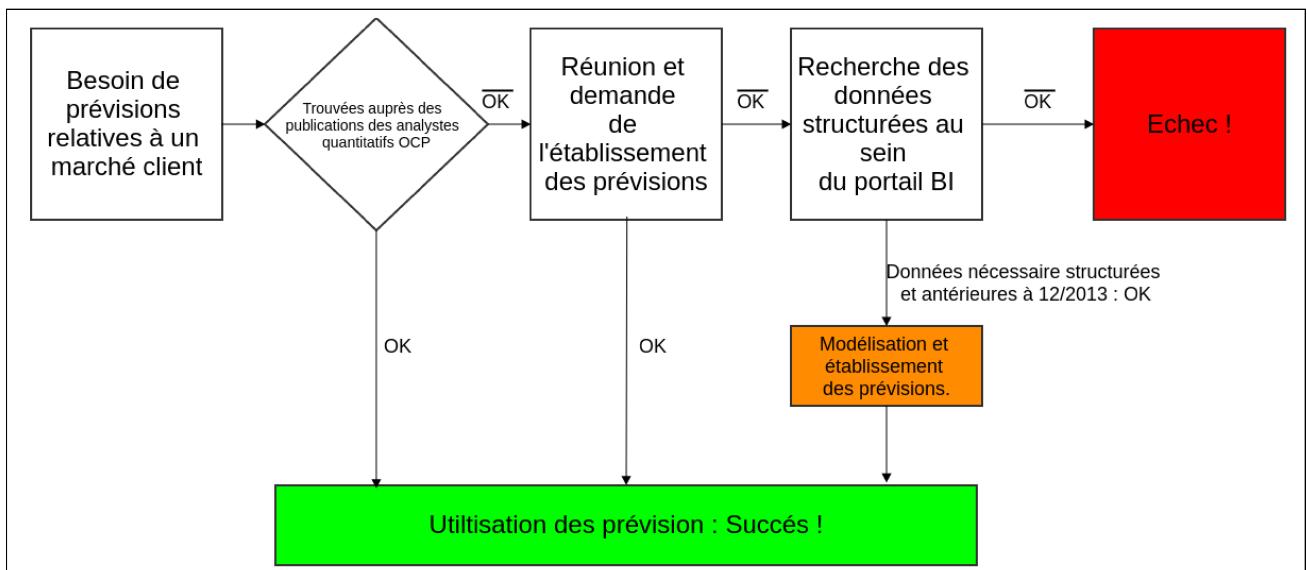


FIGURE 2.1 – Description du processus d'analyse du marché

2.1.2 Critique de l'existant

En se basant sur l'utilisation actuelle des fichiers, nous avons constaté que les analystes passent un temps important à structurer les données avant leur utilisation pour émettre des perspectives du marché en cas de non existence de prévision souhaitées.

Après avoir structuré les données, l'analyste les transforme en information en les éditant sous Excel. Cette transformation, dès fois manuelle, peut engendrer plusieurs erreurs qui sont liées à la protection de fichiers PDF ou leur mise en page sans compter le temps gaspillé par ces routines non automatisées.

Les données ainsi compilées, elles sont utilisées aux fins de modélisation. À la réapparition d'un scénario similaire concernant un client différent, le processus est à nouveau repris avec les

mêmes hypothèses.

Au vu de ce processus, nous avons pu identifier plusieurs problèmes parmi lesquels nous citons :

- Inexistence d'outils d'aide à la décision intégrant des mécanismes prévisionnels.
- Les données rapportés par le système BI en place datte parfois de fin décembre 2013.
- Le traitement des données prend un temps important ce qui gêne la compréhension mutuelle et la collaboration des analystes.
- Le niveau d'exploitabilité des données est très bas.
- L'absence d'un système d'historisation des données brutes ainsi que les résultats des analyses.
- L'absence de la structuration automatique des données brutes, notamment les .pdf.
- Faible fiabilité des données saisies manuellement.
- Risque de perte d'information à toute étape du processus.

Cette situation est principalement imputable à l'inexistence d'un traitement efficient des données non structurées à la disposition de l'OCP. Nous détaillerons dans la section 3.1.1 les technicités inhérentes à cette situation.

2.2 Revue de littérature

2.2.1 Qu'est ce que le Data Mining ?

La fouille de données, français pour le "Data Mining" est le processus de recherche et de découverte d'auparavant inconnus et potentiellement intéressants modèles dans les grands ensembles de données [10]. L'information 'minée' est typiquement représentée par un modèle de la structure sémantique de l'ensemble de données, où le modèle peut être utilisé sur de nouvelles données à des fins de prédiction ou de classification. Alternativement, des experts humains du métier en question peuvent choisir d'examiner manuellement le modèle, à la recherche d'éléments qui expliqueraient des caractéristiques précédemment mal comprises ou inconnues du domaine d'étude.

Brynjolfsson, Hitt, et Kim[12] ont mené des recherches empiriques et ont conclu que la performance des organisations est directement liée à leur capacité à prendre des décisions dirigées par les données. Il est donc important de comprendre les facteurs de succès nécessaires pour adopter des techniques de fouilles de données dans les organisations. En plus des défis techniques, le data mining présente également un ensemble de défis de gestion. Le plus important parmi ceux-ci est la nouvellement acquise capacité de l'organisation à prendre des décisions basées sur les données et donc l'éloigne de la prise de décision basée sur l'intuition. Le leadership, la gestion des talents, de la technologie, la prise de décisions et de la culture de l'entreprise sont des facteurs essentiels pour réussir à tirer parti des grandes données pour la performance organisationnelle[7].

Du côté agricole, Ashlee Vance de rapporte que l'entreprise innovante, *Climate Corp*, qui utilise des données météorologiques localisées pour prédire les rendements des cultures et utilise ces données pour échafauder des primes d'assurance sur les récoltes sur mesure aux agriculteurs individuels. *Climate Corp* enregistre et analyse plus de 60 ans de données météorologiques, comprenant les précipitations, la température et les conditions du sol avec une granularité de 4 kilomètres de rayon afin d'arriver à des prévisions de rendement des cultures[2].

2.2.2 Les études de prévisions en matière de fertilisants

L'agriculture est étroitement liée à la sécurité alimentaire des nations, un grand nombre de recherches ont été menées par des organismes gouvernementaux pour assurer un avenir agricole durable et productif. La majorité des prévisions dans les utilisations agricoles utilisent des modélisations économétriques. Cela implique l'utilisation de variables exogènes pour prévoir les rendements des cultures, la consommation des engrains, etc. Nous sommes particulièrement intéressés par les recherches liées à la prévision de la consommation d'engrais puisqu'elles se rapportent à notre problématique.

Les premiers travaux connus en matière de prévision de la consommation des engrains ont été réalisés par Vail [6], puis par Mehring et Shaw [13]. Ils ont essayé d'étudier la relation entre la consommation par hectare d'azote et entre la valeur des récoltes, la teneur du sol en azote, et la proportion de trésorerie générée du bétail. Ils ont tous étudié la relation entre les dépenses en engrains avec le revenu agricole retardé d'une année. Mehring et Shaw ont conclu que les agriculteurs ont toujours tendance à dépenser une proportion constante de leur revenu sur les engrains. Vail n'a trouvé aucune relation significative entre la consommation d'engrais et les prix des engrais. Zvi Griliches[27] a réalisé une autre étude de la consommation d'engrais avec l'objectif d'estimer l'élasticité à court et long terme de la demande. Il a constaté que les changements technologiques qui ont abouti à de nouvelles technologies de production ont conduit à une réduction du prix réel des engrais, ce qui conduit à l'adoption à grande échelle et une utilisation accrue des engrais. L'hypothèse de Griliches était qu'une augmentation de l'utilisation d'engrais a été principalement entraînée par une baisse du prix de l'engrais par rapport à d'autres intrants agricoles. Il a modelé la consommation d'engrais dans la saison en cours en fonction du prix de l'engrais et la consommation d'engrais de la saison précédente en utilisant les données 1911-1956.

La FAO¹ publie une perspective mondiale sur la demande d'engrais et les tendances des autres produits agricoles telles que les rendements des cultures et la production de céréales. Parthasarathy[24] de la FAO a analysé divers facteurs qui influent sur la consommation d'engrais à long et court terme. En plus de facteurs exogène tels que le pouvoir d'achat des agriculteurs, qui est déterminé par le caractère abordable des engrais, et la trésorerie de liquidités des agriculteurs, la disponibilité de l'engrais influence également la consommation. L'infrastructure et une meilleure gestion de la logistique résultent en une plus grande disponibilité du produit au moment du besoin des agriculteurs, ainsi la demande en assurant est convertie en vente.

1. Organisation des nations unies pour l'alimentation et l'agriculture

Selon Parthasarathy, en outre à la terre (superficie plantée et récoltée) et le rendement, d'autres facteurs tels que la quantité et la répartition des précipitations, les modes de culture, et la taille des exploitations influent également sur la consommation d'engrais.

Dans un rapport plus récent, la FAO a émis ses prévisions de la consommation d'engrais pour 2015 et 2030 et conclue que les intrants agricoles sont étroitement liés au rendement des cultures, et que la croissance de la production est directement gouvernée par les facteurs macro tels que l'augmentation de la population et le revenu par habitant. La relation positive entre la consommation d'engrais et la production agricole est bien établie dans les deux pays en développement et développés. Le scénario de base pour la projection des utilisations d'engrais suppose l'utilisation d'engrais liée à la superficie plantée et le rendement. Le modèle était un modèle de régression logistique, où les transformations logarithmiques des variables indépendantes ont été utilisés comme intrants[17].

Enfin, alors que les travaux de Griliches [27] ont porté uniquement sur l'utilisation des engrais, Tenkorang [16] a tenté de prévoir la demande d'engrais global à long terme, similairement au travail effectué par la FAO. En plus de la prévision de la consommation d'engrais, que la FAO a achevée, Tenkorang a également tenté d'estimer l'équilibre des éléments nutritifs du sol. Puisque notre projet ne portera pas sur l'estimation des éléments nutritifs du sol, nous nous concentrerons sur sa méthode d'estimation des engrais. Il a prévu la demande à travers 182 pays divisés en neuf régions, en utilisant les données de 1962 à 2005. Un intéressant constat est la relation entre les usages d'engrais dans les années suite à une récolte exceptionnelle. Les agriculteurs ont tendance à avoir le syndrome de *Good Year/Bad Year* où ils sentent qu'une saison à haut rendement est généralement suivie par un rendement plus faible. Par conséquent, ils manquent peut-être de motivation à prendre des mesures pour augmenter les consommations de fertilisants suite à une bonne année. En conséquence, Tenkorang a modélisé l'utilisation des engrais dans une région et période de temps données en fonction d'une régression linéaire entre la production de la campagne agricole actuelle, la récolte de l'année précédente, la superficie totale cultivée, et une variable fictive² pour tenir compte de tout changement structurel des utilisation d'engrais. Cependant, puisque que les terres cultivées et le rendement des cultures sont fondamentalement liés, son modèle souffre de multi-colinéarité. Ceci a été corrigé en supprimant ces variables indépendantes de l'équation en se basant sur le VIF³. Le modèle final, ajusté pour multi-colinéarité, a réalisé un R^2 élevé⁴, avec pour la majorité des régions, la terre cultivée comme facteur le plus important.

Ces études mettent en avant deux concepts importants dans les prévisions agricoles. Tout d'abord, la plupart des modèles de prévisions utilisés sont des modèles causaux qui tentent

2. *Dummy Variable* pour les anglophones.

3. Le Facteur de l'Inflation de la Variance quantifie la gravité de multi-colinéarité dans une régressions par méthodes des moindres carrés. Il fournit un indice qui mesure combien la variance d'un coefficient de régression estimé est augmenté en raison de la colinéarité.

4. En statistique, le coefficient de détermination R^2 est une mesure de la qualité de la prédiction d'une régression linéaire. Il est défini comme 1 moins le ratio entre l'erreur avec les valeurs prédictives et la variance des données

de relier une variable dépendante telle que la demande d'engrais à un ensemble de variables exogènes. Deuxièmement, ces études nous indiquent quelques uns des facteurs de base ou variables indépendantes que nous devrions prendre en compte dans la construction de notre modèle : Des facteurs agraires, climatiques mais aussi socio-économiques.

2.3 Compréhension du problème

2.3.1 Cadrage fonctionnel du projet

À la lumière de notre lecture bibliographique, nous procédons à l'établissement de la liste des divers méthodologies de prévisions en nous intéressant particulièrement aux forces et faiblesses de chacune ainsi qu'aux exigences techniques et informationnelles que celles-ci nécessitent. Nous présenterons l'approche retenue : La méthode causale.

2.3.1.1 Les étapes d'une prévision en matière de fertilisants

La prévision des engrais peut être divisée en trois étapes complémentaires les unes aux autres, à savoir :

1. l'évaluation du potentiel,
2. les prévisions de la demande
3. les prévisions des ventes

Pour le gouvernement les deux premières étapes sont importantes. Pour les organisations de commercialisation, les deuxième et troisième étapes sont pertinentes. Le gouvernement souhaite savoir l'écart entre le potentiel et la demande afin de déterminer ce qu'il doit faire pour transformer une partie du potentiel en demande effective. Une entreprise souhaite savoir ce que la demande effective est et quelle part elle peut être satisfaite par les ventes de l'entreprise.

En ce qui concerne la méthodologie de prévision employée, les méthodes de prévision entrent dans l'une des quatre approches de base :

- Mesure du potentiel par des méthodes agronomiques ou orientées besoins,
- L'analyse et la projection de séries chronologiques,
- Modèles causaux,
- Approche qualitative.

De par sa nature, la première approche s'intéresse au long terme et est essentiellement idéaliste. Elle nous dit ce que la demande d'engrais pourrait être et non ce que la demande d'engrais est susceptible d'être.

La seconde approche repose sur des données historiques pour analyser et discerner des schémas de la demande pour prévoir l'avenir, l'hypothèse étant que l'avenir est une continuation du passé. Des données fiables sur plusieurs années sont essentielles pour cette approche.

La troisième approche cherche à établir une relation de cause à effet entre la demande d'engrais et d'autres variables indépendantes dans l'environnement du marché des fertilisants.

Comme dans le cas de la méthode des séries chronologiques, les données passées sont importantes pour évaluer l'effet de ces facteurs sur la demande. Les deuxième et troisième méthodes ne peuvent pas être utilisées lorsque l'on veut prévoir l'avenir d'un produit qui n'a pas d'historique ou est dans un stade de développement tel qu'aucun modèle ni tendance sont perceptibles. La consommation d'engrais dans de nombreux pays en développement est dans une telle situation.

La quatrième approche utilise des informations qualitatives, y compris des avis d'experts, pour prévoir la demande. Cette approche peut ou peut ne pas considérer le passé. Les approches qualitatives sont utilisées lorsque les informations sont rares ou ne sont pas fiables, comme dans le stade élémentaire du cycle de vie du produit. Ces conditions étant caractéristiques des marchés des engrais dans les pays en développement, les techniques qualitatives sont très utiles. L'utilisation du jugement humain associé à des systèmes de notation transforment les opinions qualitatives en moyens quantitatifs pour les prévisions. Une méthode couramment utilisée pour la prévision est l'analyse des données historiques pour discerner l'évolution de la croissance de la demande et de l'étendre à l'avenir pour prévoir la demande. Si les données de plusieurs années de ventes d'engrais sont disponibles et la tendance est relativement stable, il est possible de lire les données antérieures de la "vitesse" actuelle de la croissance de la demande et la mesure dans laquelle la vitesse augmente ou diminue.

L'extension de tendance est effectuée comme suit :

- Les données chronologiques de la demande d'engrais sont répertoriées.
- Ces données sont susceptibles d'avoir quatre composantes majeurs. La première reflète la *tendance* qui se réfère à la variation se produisant constamment sur une longue période. La deuxième composante saisit un mouvement cyclique de la demande dont la plupart des produits, y compris les engrais, sont soumis. Comprendre la variation cyclique est utile pour les prévisions à court et moyen terme. Le troisième élément est la *variation saisonnière* au sein de chaque année. La quatrième composante est la perturbation causée par des événements erratiques et aléatoires. Grâce à des méthodes statistiques les données de séries chronologiques sont analysées et décomposées en ces quatre composantes et recombinées pour fournir la formule de prévision.
- L'extension de la tendance est déterminée par un ajustement d'une équation mathématique appropriée qui se rapproche de la tendance historique. Par exemple, la tendance peut être linéaire, à savoir une ligne droite, la pente de degré de la ligne droite indiquant la quantité d'augmentation annuelle de la demande. Les tendances peuvent être quadratique ou exponentielle. Les équations mathématiques décrivent le taux représenté par chacune de ces formes de croissance et nous choisissons l'équation minimisant l'erreur à la tendance historique pour calculer la demande future.

2.3.1.2 Les modèles causaux pour la prévision

Le modèle causal est ainsi appelé parce qu'il emploie la relation de cause à effet entre la demande d'engrais et les facteurs qui l'affectent. Le modèle ne représente pas la demande d'engrais au fil du temps ou pour un moment particulier, mais présente la demande par rapport

à un ensemble de circonstances. Bien que la méthode d'extension de tendance suppose que le temps reflète tous les facteurs, la méthode causale cherche à établir des relations directes entre la demande d'engrais et les facteurs qui l'influencent. Les facteurs influant sur la demande, comme nous l'avons vu dans la section 2.2.2, comprennent les prix des cultures, les prix des engrais, la disponibilité du crédit, la superficie irriguée, les précipitations, la superficie cultivée en variétés à haut rendement, le modèle de culture,etc. En analysant les données passées, un ensemble de facteurs essentiels qui ont l'effet le plus profond peuvent être sélectionnés et l'effet des facteurs sélectionnés quantifiés et exprimés sous la forme d'équations mathématiques. Pour projeter la demande pour les années à venir, l'état probable de chaque facteur critique sélectionné à ce point de temps doit d'abord être évalué.

Cette méthode est extrêmement complexe, impliquant des équations mathématiques pour exprimer les relations et les inter-relations entre les variables. En outre, la fiabilité n'est pas garantie car elle dépend des prévisions des valeurs des facteurs critiques choisis. Au mieux, il nous dit ce que la demande est susceptible d'être dans un ensemble donné de circonstances, mais il n'y a aucune certitude que cet ensemble de circonstances prévaudra durant les années sous prévisions.

Une condition sine qua non pour l'utilisation de cette méthode est la disponibilité des données, selon la région et la saison, de divers facteurs critiques en plus des données sur la demande d'engrais pendant plusieurs années. Comme mentionné, une sérieuse limitation de la méthode de causalité est que la prévision, à son tour, dépend des indicateurs qui doivent être eux-mêmes projetés.

2.3.2 Cadrage technique du projet

2.3.2.1 Architecture globale de notre solution

La solution que nous proposons à la problématique que nous nous sommes fixée en section 1.2.3.2 intègre les réalisations suivantes :

- Un noyau de prévisions générant des modèles causaux stockable en des procédures automatisables.
- Un noyau de Parsing asservi au noyau de prévisions, servant à structurer les données nécessaires à l'établissement des modèles de prévision.
- Un système de fichiers indexant les données avant et après leur structuration, ainsi que les procédures automatisables produites par le noyau de prévisions.

La solution mise en place devrait permettre de :

- Automatiser les processus de collecte, de structuration et de modélisation des données.
- Minimiser le temps de structurations des données complexes.
- Centraliser et historiser les données dérivées aux fins d'analyse et de modélisation.
- Augmenter la fiabilité des données non structurées transformées.
- Proposer des prévisions à la demande en matières de demande en fertilisants.

- La Génération de rapports et de graphiques de synthèses des conditions du marché.
L'architecture applicative d'une telle solution est résumée par la figure 2.2 :

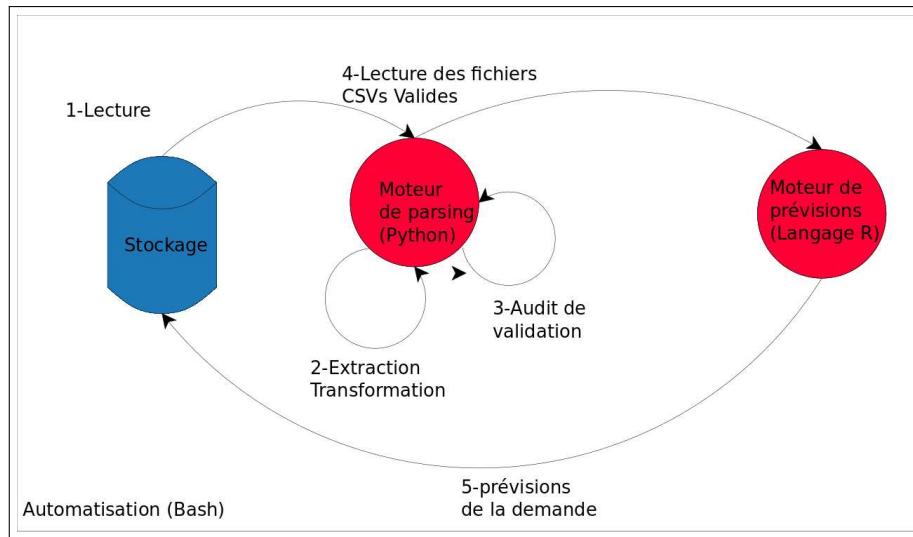


FIGURE 2.2 – Architecture applicative de la solution proposée

2.3.2.2 Outils de réalisation

- **Le langage de programmation Python :**

- * Python est un langage de programmation objet, multi-paradigme et multiplate-forme. Il favorise la programmation impérative structurée, fonctionnelle et orientée objet.
- * Il est conçu pour optimiser la productivité des programmeurs en offrant des outils de haut niveau et une syntaxe simple à utiliser.

- * **Pourquoi Python dans notre projet ?**

- + Calculatrice vectorielle évoluée.
- + Traitement de fichier texte.
- + Scripts, ou commandes Unix pour traitements de fichiers par lots.
- + Langage "Glue" pour enchaîner les traitements par différents programmes.
Cas d'utilisation dans notre projet : Parsing PDF et enregistrement sur fichiers, repris par un code pour structuration des données et traitement.

- **Le langage statistique R :**

- * R est un environnement permettant de faire des analyses statistiques et de produire des graphiques évolués.
- * C'est également un langage de programmation complet et mature. Sa licence est open-source, son utilisation est gratuite, même dans le contexte de l'entreprise ou de la formation.
- * L'environnement R intègre de nombreuses fonctionnalités pour l'acquisition, le nettoyage et la modélisation de données.

* **Pourquoi R dans notre projet ?**

- + C'est un outil très puissant et très complet, particulièrement bien adapté pour la mise en œuvre informatique de méthodes statistiques. Il est plus difficile d'accès que certains autres logiciels du marché (comme SPSS ou Matlab par exemple), car il n'est pas conçu pour être utilisé à l'aide de «clics» de souris dans des menus. Mais son approche par l'écriture de code informatique pour l'analyse statistique lui confère la flexibilité désiré pour un projet de Data Science.
- + L'approche est pédagogique puisqu'il faut maîtriser les méthodes statistiques pour parvenir à les mettre en œuvre.
- + l'outil est très efficace lorsque l'on domine le langage puisque l'on devient alors capable de créer ses propres outils, ce qui permet ainsi d'opérer des analyses très sophistiquées sur les données en retenant tous les avantages de la programmation modulaire dont la ré-utilisabilité et la générnicité du code produit.

— **Le langage de scripting système BASH :**

- * Un script BASH est une suite d'instructions, de commandes qui constituent un scénario d'actions. C'est un fichier texte que l'on peut exécuter, c'est à dire, lancer comme une commande.

* **Pourquoi BASH dans notre projet ?**

- + le shell est l'interface de tous les jours en UNIX. Bien connaître son shell permet d'économiser beaucoup d'efforts.
- + le shell est universel : peu importe le système UNIX, les tâches d'automatisation des appels systèmes privilégient les scripts BASH.
- + Il est plus aisés de programmer en BASH, par rapport par exemple à C ; le Shell n'a pas été conçu pour être minimal ou théoriquement élégant ; il a été conçu pour être flexible et pratique. Ainsi, dans bien des cas on va s'en servir pour automatiser les tâches routinières du système.

Cas d'utilisation dans notre projet : Automatisation de la création des arborescences de structuration des données à consolider ainsi que les appels systèmes de leur labellisation.



(a) Logo Python



(b) Logo R



(c) Logo Bash

FIGURE 2.3 – Logos des outils utilisés

CHAPITRE 3

COLLECTE, COMPRÉHENSION ET PRÉPARATION DES DONNÉES.

“Data ! Data ! Data !” he cried impatiently. “I can’t make bricks without clay.”

Sherlock Holmes

Ce chapitre présente le processus d’acquisition des données et leur préparation pour les besoins d’analyse et de modélisation. Nous décrivons en premier lieu le paysage des données au moment du début de notre stage pour ensuite présenter les opérations de consolidation, d’audit et d’extension que nous lui faisons subir.

3.1 Compréhension et préparation des données locales¹

3.1.1 Compréhension des données locales

Comme rappelé dans la section 1.2.2. Les travaux de nos prédécesseurs dans leurs efforts de moderniser le portail *Business Intelligence* de l'OCP se sont arrêtés à automatiser l'archivage des données se rapportant aux historiques de ventes en terme de prix et volumes en un premier lieu[20] avant de concevoir le socle OLAP² dans la vue de générer des rapports synthétiques concernant les historiques des échanges du marché des phosphates ensuite[1].

Seules sont ainsi présentes les données concernant les échanges internationaux en terme de produits phosphatés et ceux-ci sont présents sous deux différentes formes de données. La première est notre format de fichier source : des fichiers .pdf non structurés vis-à-vis de notre besoin (figure 3.2), et qui présentent :

- L'avantage d'être à jour, exhaustifs et dont la véracité est certifiée par un organisme international (IFA)
- L'inconvénient d'être flexible pour la lecture humaine mais ne présentant pas une grammaire machine formelle rendant possible une analyse syntaxique, comme en témoigne la figure 3.1.

		DAP Exports by Destination					PIT/2015/3Q/P/7		
		January - September 2015 ('000 metric tonnes P2O5)							
3Q 2015 Importing countries	Exporting countries	USA, Africa & West Asia		Brazil	China	Others	TOTAL 2015	TOTAL 2014	TOTAL 2013
Africa									
Angola	-	-	-	0.2	-	-	0.2	0.3	0.4
Benin	-	-	-	-	-	-	-	1.0	-
Cameroun	3.0	-	0.1	2.7	-	-	5.8	3.2	1.7
Congo	-	-	-	-	-	-	-	1.0	-
Côte d'Ivoire	7.8	-	-	-	-	-	7.8	14.1	23.8
Egypt	-	-	-	-	-	-	-	-	-
Ethiopia	-	-	-	-	-	-	-	-	-
Ghana	-	-	-	-	-	-	-	-	-
Kenya	-	-	-	-	-	-	-	-	-
Liberia	6.9	-	-	-	-	-	6.9	11.4	19.6
Madagascar	0.1	-	0.5	-	-	-	0.7	0.2	0.4
Malawi	-	-	-	-	-	-	-	-	7.1
Mali	7.6	-	-	-	-	-	7.6	-	2.5
Mauritanie	-	-	-	-	-	-	-	-	1.5
Maurice	-	-	-	1.5	-	-	1.5	1.5	2.0
Mozambique	-	-	7.2	0.2	-	-	7.5	9.0	2.2
Nigeria	2.5	-	-	-	-	-	-	2.5	-
Senegal	15.2	-	-	0.2	-	-	15.4	17.9	25.4
Sierra Leone	-	-	-	-	-	-	-	-	0.1
South Africa	3.7	-	7.2	-	-	-	10.9	2.8	1.8
Sudan	0.8	-	-	-	-	-	0.8	2.0	1.4
Tanzania	5.0	-	2.9	-	-	-	7.9	14.2	14.2
Togo	7.2	-	-	0.7	-	-	7.9	6.2	0.4
Zambie	-	-	-	-	-	-	-	-	0.8
Zimbabwe	-	-	-	-	-	-	-	-	1.3
Subtotal	115.3	-	19.5	3.8	-	-	138.7	231.5	251.6

International Fertilizer Industry Association Page: 26

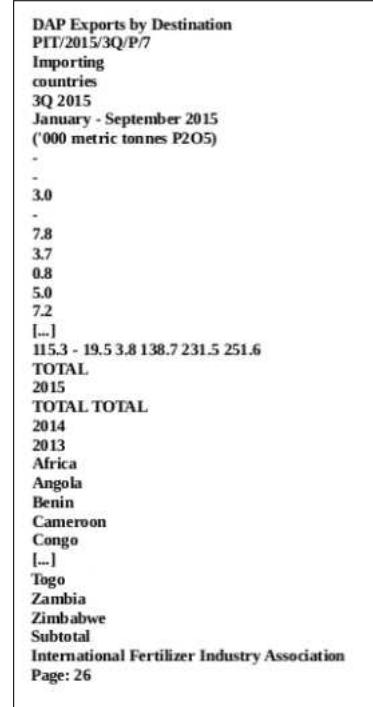


FIGURE 3.2 – Exemple d'une "Trade Matrix"³ dans le rapport trimestriel de l'IFA

La seconde est structurée dans des datamarts dont une sortie de requête est présentée dans la figure 3.3. Ceci est notre format de données cible et présente :

1. Données disponibles au sein du portail *Business Intelligence* de l'OCP
2. le traitement analytique en ligne (OnLine Analytical Processing, OLAP) est un type d'application informatique orienté vers l'analyse sur-le-champ d'informations selon plusieurs axes, dans le but d'obtenir des rapports de synthèse

- L'avantage d'offrir le maximum de flexibilité pour le requêtage, et d'être de grande qualité en terme de disponibilité et de véracité, puisque celui-ci a été soigneusement introduit à la main⁴.
- L'inconvénient d'être prohibitif en temps et en ressources humaines. En effet, nous avons constaté un retard datant de fin décembre 2013 par rapport aux derniers .pdf reçus par l'OCP.

Importing country	Region (IFA)	Region (OCP)	Exporting country	Product	Year	Quarter	Code	kT	P205/Product	Cumulated/
Austria	West Europe	North Europe & FSU	West Asia	Rock	2015	Q4	AustriaWest EuropeNorth Europe & FSUWest AsiaRock2015Q4P205	2 P205	Cumulated	
Austria	West Europe	North Europe & FSU	West Asia	Rock	2015	Q4	AustriaWest EuropeNorth Europe & FSUWest AsiaRock2015Q4Produ	7 Product	Cumulated	
Austria	West Europe	North Europe & FSU	Africa	Rock	2015	Q4	AustriaWest EuropeNorth Europe & FSUAfricaRock2015Q4P205	192 P205	Cumulated	
Austria	West Europe	North Europe & FSU	Africa	Rock	2015	Q4	AustriaWest EuropeNorth Europe & FSUAfricaRock2015Q4Product	640 Product	Cumulated	
Austria	West Europe	North Europe & FSU	China	Rock	2015	Q4	AustriaWest EuropeNorth Europe & FSUCHinaRock2015Q4P205	0 P205	Cumulated	
Austria	West Europe	North Europe & FSU	China	Rock	2015	Q4	AustriaWest EuropeNorth Europe & FSUCHinaRock2015Q4Product	0 Product	Cumulated	
Austria	West Europe	North Europe & FSU	Others	Rock	2015	Q4	AustriaWest EuropeNorth Europe & FSUOthersRock2015Q4P205	0 P205	Cumulated	
Austria	West Europe	North Europe & FSU	Others	Rock	2015	Q4	AustriaWest EuropeNorth Europe & FSUOthersRock2015Q4Product	0 Product	Cumulated	
Austria	West Europe	North Europe & FSU	America	Rock	2015	Q4	AustriaWest EuropeNorth Europe & FSUAmericaRock2015Q4P205	0 P205	Cumulated	
Austria	West Europe	North Europe & FSU	America	Rock	2015	Q4	AustriaWest EuropeNorth Europe & FSUAmericaRock2015Q4Product	0 Product	Cumulated	
Belgium	West Europe	North Europe & FSU	West Asia	Rock	2015	Q4	BelgiumWest EuropeNorth Europe & FSUWest AsiaRock2015Q4P205	23 P205	Cumulated	
Belgium	West Europe	North Europe & FSU	West Asia	Rock	2015	Q4	BelgiumWest EuropeNorth Europe & FSUWest AsiaRock2015Q4Produ	77 Product	Cumulated	
Belgium	West Europe	North Europe & FSU	Africa	Rock	2015	Q4	BelgiumWest EuropeNorth Europe & FSUAfricaRock2015Q4P205	172 P205	Cumulated	
Belgium	West Europe	North Europe & FSU	Africa	Rock	2015	Q4	BelgiumWest EuropeNorth Europe & FSUAfricaRock2015Q4Product	573 Product	Cumulated	
Belgium	West Europe	North Europe & FSU	China	Rock	2015	Q4	BelgiumWest EuropeNorth Europe & FSUCHinaRock2015Q4P205	0 P205	Cumulated	
Belgium	West Europe	North Europe & FSU	China	Rock	2015	Q4	BelgiumWest EuropeNorth Europe & FSUCHinaRock2015Q4Product	0 Product	Cumulated	
Belgium	West Europe	North Europe & FSU	Others	Rock	2015	Q4	BelgiumWest EuropeNorth Europe & FSUOthersRock2015Q4P205	633 P205	Cumulated	
Belgium	West Europe	North Europe & FSU	Others	Rock	2015	Q4	BelgiumWest EuropeNorth Europe & FSUOthersRock2015Q4Product	2110 Product	Cumulated	
Belgium	West Europe	North Europe & FSU	America	Rock	2015	Q4	BelgiumWest EuropeNorth Europe & FSUAmericaRock2015Q4P205	0 P205	Cumulated	
Belgium	West Europe	North Europe & FSU	America	Rock	2015	Q4	BelgiumWest EuropeNorth Europe & FSUAmericaRock2015Q4Produ	0 Product	Cumulated	
Finland	West Europe	North Europe & FSU	West Asia	Rock	2015	Q4	FinlandWest EuropeNorth Europe & FSUWest AsiaRock2015Q4P205	0 P205	Cumulated	
Finland	West Europe	North Europe & FSU	West Asia	Rock	2015	Q4	FinlandWest EuropeNorth Europe & FSUWest AsiaRock2015Q4Produ	0 Product	Cumulated	
Finland	West Europe	North Europe & FSU	Africa	Rock	2015	Q4	FinlandWest EuropeNorth Europe & FSUAfricaRock2015Q4P205	0 P205	Cumulated	
Finland	West Europe	North Europe & FSU	Africa	Rock	2015	Q4	FinlandWest EuropeNorth Europe & FSUAfricaRock2015Q4Product	0 Product	Cumulated	
Finland	West Europe	North Europe & FSU	China	Rock	2015	Q4	FinlandWest EuropeNorth Europe & FSUCHinaRock2015Q4P205	0 P205	Cumulated	
Finland	West Europe	North Europe & FSU	China	Rock	2015	Q4	FinlandWest EuropeNorth Europe & FSUCHinaRock2015Q4Product	0 Product	Cumulated	
Finland	West Europe	North Europe & FSU	Others	Rock	2015	Q4	FinlandWest EuropeNorth Europe & FSUOthersRock2015Q4P205	0 P205	Cumulated	
Finland	West Europe	North Europe & FSU	Others	Rock	2015	Q4	FinlandWest EuropeNorth Europe & FSUOthersRock2015Q4Product	0 Product	Cumulated	
Finland	West Europe	North Europe & FSU	America	Rock	2015	Q4	FinlandWest EuropeNorth Europe & FSUAmericaRock2015Q4P205	0 P205	Cumulated	
Finland	West Europe	North Europe & FSU	America	Rock	2015	Q4	FinlandWest EuropeNorth Europe & FSUAmericaRock2015Q4Produ	0 Product	Cumulated	
France	West Europe	North Europe & FSU	West Asia	Rock	2015	Q4	FranceWest EuropeNorth Europe & FSUWest AsiaRock2015Q4P205	21 P205	Cumulated	
France	West Europe	North Europe & FSU	West Asia	Rock	2015	Q4	FranceWest EuropeNorth Europe & FSUWest AsiaRock2015Q4Produ	70 Product	Cumulated	

FIGURE 3.3 – Exemple d'une sortie de requête sur le Datamart OCP-CM

3.1.2 Consolidation des données locales : Développement spécifique d'un ETL

La consolidation désigne la collection et l'intégration des données de sources multiples en une unique destination. Durant ce processus, nous unifierons les deux types de formats de données. Ceci nous permettra de présenter les données de manière plus flexible, tout en facilitant leur analyse effective. Ceci nous amène à considérer le format adopté par le datamart OCP-CM comme format cible de consolidation et les fichiers .pdf comme format source.

La nature non structurées des données interdit toute approche BI de l'intégration de données via ETL préfabriqué. Nous décrivons dans ce qui suit L'ETL que nous développons.⁵

3.1.2.1 Conception fonctionnelle du processus de consolidation

L'OCP reçoit régulièrement des rapports trimestriels présentant la situation du marché accompagnée des mouvements observés des produits fertilisants. Ces mouvements sont reportés sur des tableaux tel que celui présenté dans la figure 3.2. Notre solution doit ainsi considérer les nuances fonctionnelles suivantes :

4. À travers une lecture "humaine" des .pdf présentés par la figure 3.2

5. Le code-source est disponible dans le dossier *Sanzoriarty IFA Parser* du dépôt **GitHub** de notre mémoire de projet de fin d'études[15]

- Les rapports diffèrent par leur granularité. En effet ceux-ci peuvent être :
 - DET : Détailés. Présentant l'historique des mouvements de produits entre les pays du monde deux-à deux.

3Q 2014	Importing countries	Exporting countries	Spain	Lithuania	Russia	USA	Morocco	Tunisia	Jordan	Saudi Arabia	Turkey	China	Australia	Various	TOTAL 2014	TOTAL 2013	TOTAL 2012
Latin America																	

FIGURE 3.4 – Exemple d'en-tête des tables du format DET des rapports IFA.

- AGG : Agrégés. Présentant l'historique des mouvements entre les pays agrégés selon la région du monde à laquelle ceux-ci appartiennent.

1Q 2015	Importing countries	Exporting countries	Africa, West Asia & USA	Brazil	China	Others	TOTAL 2015	TOTAL 2014	TOTAL 2013
West Europe									

FIGURE 3.5 – Exemple d'en-tête des tables du format AGG des rapports IFA.

- Les rapports diffèrent par les normes des chiffres rapportés. En effet ceux-ci peuvent être :
 - NOT CUMULATED : Les chiffres rapportés au trimestre Q_i sont bruts et représentent uniquement les ventes ayant effectivement eu lieu durant ce trimestre.
 - CUMULATED : Les chiffres rapportés au trimestre Q_i sont cumulés, i.e $Q_i = \sum_{j=1}^{j=i} Q_j$
 - ANN : Les chiffres rapportés représentent toutes les ventes de l'année, i.e $Q_i = \sum_{j=1}^{j=4} Q_j$
- Les attributs des enregistrements du datamart OCP-CM, sont des champs obligatoirement **NOT NULL** et sont les suivants :
 - "Importing.countries" : Pays de destination de l'enregistrement-vente.
 - "Region..IFA." : Région à laquelle appartient le pays de destination selon le découpage IFA.
 - "Region..OCP." : Région à laquelle appartient le pays de destination selon le découpage OCP.
 - "Exporting.countries" : Pays d'origine de l'enregistrement-vente.
 - "Product" : Produit de l'enregistrement-vente. (MAP, DAP, PA, TSP ,Rock⁶)
 - "Year" : Année de l'enregistrement-vente.
 - "Quarter" Trimestre de l'enregistrement-vente.
 - "Code" : Code de Synthèse des champs précédents.
 - "kT" : Poids de l'enregistrement-vente en kT⁷.

6. Minerai du phosphate brut en roche.

7. Kilotonne = 10^6 kilogramme.

- "P2O5.Product" : Drapeau indiquant si le poids indiqué à la colonne *kTest* net en P2O₅⁸ ou en poids brut.
 - "Cumulated.Not.cumulated" : Drapeau indiquant la norme trimestrielle de l'enregistrement-vente.
 - "AGG.DET.ANN" : Drapeau indiquant la granularité de l'enregistrement-vente
- Ainsi au-delà de la lecture des données contenues au sein des documents .pdf, ceux-ci doivent être :
- **Décumulés** : Des enregistrement des volumes unitaires par trimestre doivent être créés.
 - **Convertis en P2O5** : Les données contenues dans les .pdf représentent les volumes échangés par kT qu'ils faut convertir selon la concentration du produit de l'enregistrement-vente en P2O5.
 - **Normés** : La nature de l'agrégation trimestrielle de l'enregistrement-vente doit être spécifiée.

3.1.2.2 Conception technique du processus de consolidation

La collecte et la préparation des données est un processus extrêmement lent et complexe, traditionnellement fait à la main et dont l'automatisation est souvent très difficile. Le constat montre que 80 % de la durée d'un projet Data Mining est consacrée à la récolte et la préparation des données[5]. Pour des données non structurées telles que des fichiers .pdf, la tâche est d'autant plus complexe. À la question : "Comment réaliser un *Parsing* de fichiers PDFs", la réponse est souvent : "avec beaucoup de difficultés".

PDF est un format de description de page et ne contient aucune information sur la structure logique d'un document telle que :

- L'emplacement du titre,
- Le début d'un paragraphe,
- Si la page est en une seule colonne ou plusieurs, etc.

Tout ce qu'un fichier .pdf indique est l'emplacement des caractères. La figure 3.6 ci-dessous montre comment les caractères sont disposés sur la page de la figure 3.2.

Une solution existe : **Tabula** de Mozilla écrite en **Ruby** et précédemment utilisée par nos

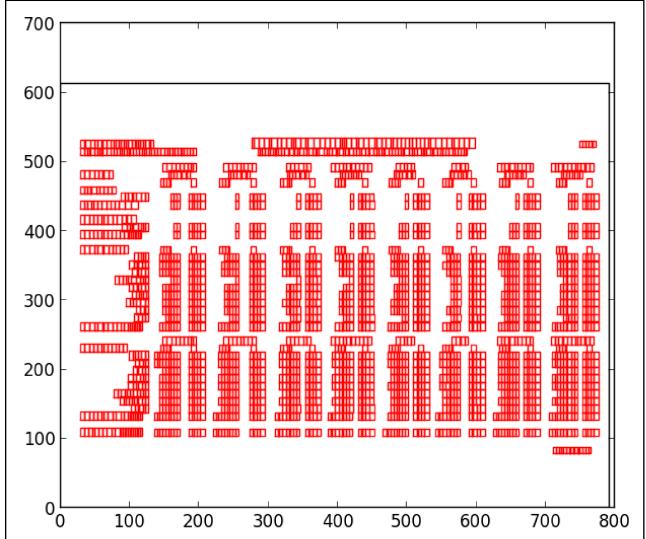


FIGURE 3.6 – Représentation de l'information transportée par la page .pdf de la figure 3.2 pour une résolution d'image de 700x800.

8. Pentoxyde de phosphore, la molécule de base des engrais phosphatés

prédecesseurs ([20],[1]). Mais celle-ci présente l'inconvénient de nécessiter de l'utilisateur de dessiner des rectangles autour des tables cibles, ce qui n'offre aucun avantage d'automatisation.

Notre solution utilise le package **pdfminer**[26] du langage **Python** qui extrait les objets non-texte et les mots de liaison en tant que blocs cohérents. Cette classification est montrée dans la figure 3.7 ci-contre. Les boîtes bleues indiquent où **pdfminer** a rassemblé un ensemble de caractères pour en faire des *text boxes*⁹ (cet ensemble de caractères peut être des mots ou des phrases) et les boîtes rouges indiquent les éléments non-texte (i.e., des lignes, des rectangles, etc.).

La méthode utilisée par notre solution s'inspire des algorithmes d'analyse d'image et est similaire à la transformée de Hough utilisée par **Tabula**. Une transformée de Hough retrouve des segments de droites arbitrairement orientés dans une image. Notre problème, ici, est plus simple, nous nous intéressons uniquement aux formes horizontales et verticales.

Pour trouver ces lignes verticales et ces colonnes, nous projetons les boîtes bleues (textuelles) sur l'axe horizontal (pour retrouver les colonnes¹⁰) et sur l'axe vertical (pour retrouver les lignes¹¹). La projection consiste en l'énumération du nombre de boîtes bleues le long d'une ligne horizontale ou verticale. Nous concluons ainsi que les frontières entre les colonnes et les lignes sont marquées par les creux des valeurs des projections dans le graphe de projection ; alors que les colonnes et les lignes en elles-même représentent les pics de la courbe. La figure 3.8 montre le résultat du traitement de la page de la figure 3.2.

Les graphes en haut et à gauche représentent le décompte des projections alors que les points noirs représentent les endroits où nous allons placer les frontières des lignes et des colonnes.

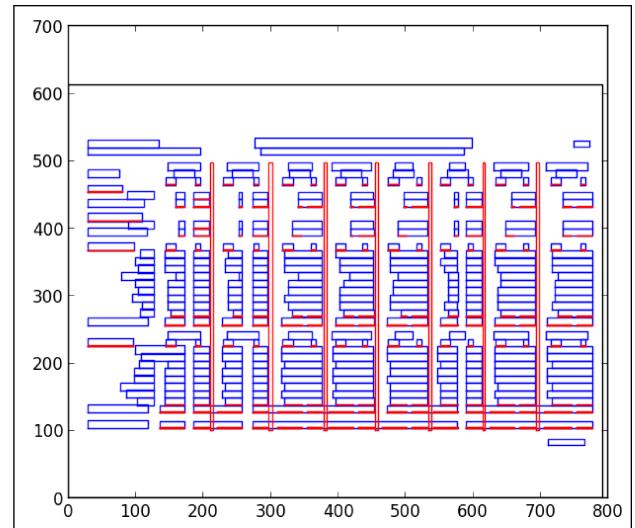


FIGURE 3.7 – Classification de l'information transportée par la page .pdf de la figure 3.2 selon (\in zone de texte, \notin zone de texte)

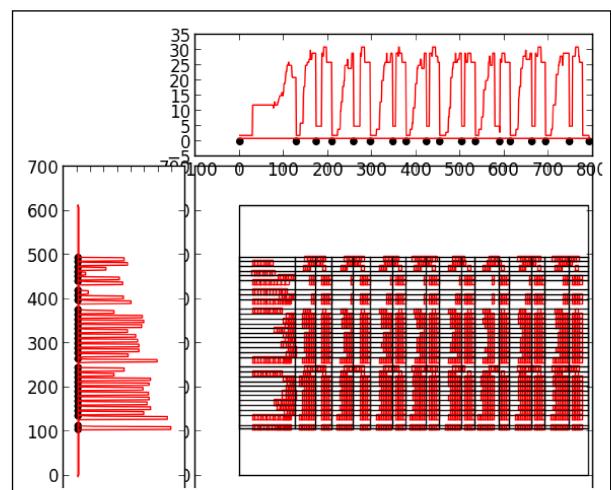


FIGURE 3.8 – Inférence des frontières de la table de la figure 3.2.

9. Zone de texte.

10. Nom du pays exportateur

11. Données des enregistrements-ventes effectués par un pays importateur

3.1.2.3 Conception du flux de données du processus de consolidation

La nature des transformations à faire subir les sources de données en vue de leur consolidation, nous amène à concevoir une solution technique en 2 phases :

1. Une phase de structuration des données¹² :

- Nous parcourons l'arborescence contenant les .pdf source.
- Nous séparons le fichier .pdf en pages individuelles.
- Nous analysons les premiers caractères de la page en vue de labelliser chaque page selon si elle contient un produit, quel trimestre traite-t-elle, quelle normes les chiffres rapportés suivent-ils (cf. figure 3.1).
- Nous faisons appel à notre outil de détection des tables discuté plus haut pour transformer chaque page-produit en une matrice de données.
- Nous compilons les matrices données d'un même produit en un seul fichier "data.csv" que nous confions au gestionnaire de fichier à placer dans une arborescence codifiant les données fonctionnelles de chaque fichier¹³.

2. Une phase d'unification des modèles de données¹⁴ :

- Nous parcourons l'arborescence générée par la phase précédente pour lire l'ensemble des "data.csv".
- Nous interrogeons le système de fichier lors de la récupération de "data.csv" quant aux données fonctionnelles de celui-ci.
- Nous procédons si besoin aux opérations de décumulation et conversion.
- Les champs sont remplis conformément au modèle datamart OCP-CM (cf. figure 3.3) et les données sont ainsi unifiées après audit de leur qualité.

Ces étapes sont résumées par le diagramme de flux de données de la figure 3.9.

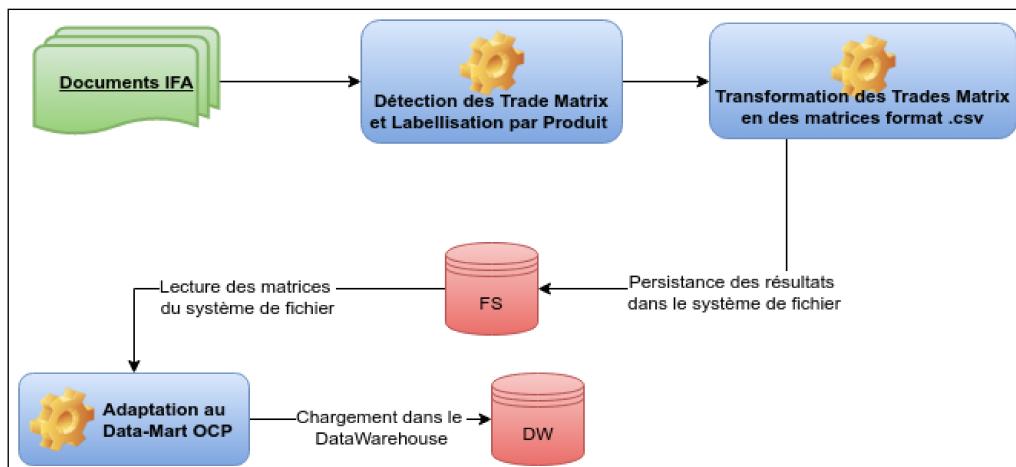


FIGURE 3.9 – Flux de données du processus de consolidation.

12. cf. Annexe B1 pour le code-source

13. cf. Annexe B3 pour plus de détails

14. cf. Annexe B2 pour le code-source

3.1.3 Audit et description des données locales consolidées

3.1.3.1 Audit des données de consolidation

Lors de nos tests de la solution que nous avons mis au point lors de la section 3.1.2.2, nous obtenions un taux de réussite de reconnaissance et extraction des tableau de 100%. Cependant et de par la nature critique des données que nous traitons, nous avons procédé à tes tests rigoureux de deux natures différentes.

Audit assisté par ordinateur :

Notre processus de consolidation a généré en moyenne près de 40800 enregistrements-vente par année. Après une consolidation des années 2014, 2015 et début 2016, le total s'élève à 94240 nouveaux enregistrements-vente dont nous nous devons d'assurer la justesse. Une vérification humaine un par un de chacun de ces enregistrements-vente est une tâche pharamineuse que nous taclons en premier lieu en nous assistons du langage **R**¹⁵.

La figure 3.10a ci-dessous, valide bien le nombre d'enregistrements-vente auxquels on s'attend, soit 94240 lignes. De plus, nous remarquons un premier soucis : Les cases de volumes nuls au sein de la colonne "kT" de la Trade Matrix ont été interprétés par le caractère "-". Ce qui empêche la colonne d'être traitée en tant que vecteur numérique. Nous remédions à ce problème en remplaçant ce caractère par le numérique "0", comme procédé dans la figure 3.10b ci -après.

```

Console ~/ ↵
> f = read.csv('/media/moriarty/YACHAOUI/Final_Database.csv')
> dim(f)
[1] 94240   12
> summary(f)
Importing.countries      Region..IFA.      Region..OCP.      Exporting.cou...
Bulgaria : 1012    Latin America :18458    North Europe & FSU :20712    China :10926
Colombia : 1012    Africa :17574        Latin America :18458    Various: 7126
France : 1012     West Europe :14300       Africa :17574        Morocco: 6706
Germany : 1012    East Asia :11696       South Asia & Middle East:12418    USA : 6584
Indonesia: 1012   West Asia : 8686       East Asia :11696       Tunisia: 6252
Japan : 1012      Central Europe: 7710    South & Central Europe : 7710    Jordan : 5092
(Other) :88168    (Other) :15816       (Other) :5672        (Other):51564
Product      Year      Quarter
DAP :27608    Min. :2014    Q1:19680
MAP :21784    1st Qu.:2014   Q2:23714
PA :23392    Median :2014   Q3:26446
Rock: 8496    Mean :2014    Q4:24400
TSP :12960    3rd Qu.:2014
Max. :2015

Code
Abu Dhabi, UAEWest AsiaSouth Asia & Middle EastAlgeriaRock2014Q2P205 : 2 -
Abu Dhabi, UAEWest AsiaSouth Asia & Middle EastAlgeriaRock2014Q3P205 : 2 0
Abu Dhabi, UAEWest AsiaSouth Asia & Middle EastAlgeriaRock2014Q4P205 : 2 0.1
Abu Dhabi, UAEWest AsiaSouth Asia & Middle EastAustraliaDAP2014Q1P205 : 2 1
Abu Dhabi, UAEWest AsiaSouth Asia & Middle EastAustraliaDAP2014Q1Product: 2 0.2
Abu Dhabi, UAEWest AsiaSouth Asia & Middle EastAustraliaDAP2014Q2P205 : 2 2
(Other) :94228  (Other) :51368
P205.Product Cumulated.Not.Cumulated AGG.DET.ANN
P205 :51368 Cumulated :47120 AGG:16848
Product:42872 Not Cumulated:47120 ANN:21340
DET:56052

```

	0	0.1	1	0.2	2	3	0.3	0.5	5	0.4	4	0.7
80786	2163	903	640	434	351	311	279	233	209	206	197	
6.7	7	6	8	0.6	10	0.8	0.9	1.5	1.1	11	9	
157	150	139	130	128	120	104	100	92	85	83	81	
12	1.2	1.4	17	1.7	2.3	15	16	1.8	1.9	1.6	1.3	
77	77	76	72	72	72	69	68	67	63	61	59	
2.8	2.7	14	2.5	2.4	2.6	13	3.2	18	22	3.3	20	
56	53	52	52	51	50	49	49	48	46	46	42	
3.7	2.2	25	3.8	5.1	4.4	7.6	2.6	27	4.6	2.1	33	
42	40	40	40	39	38	38	37	37	36	35	35	
3.4	3.5	19	32	2.9	5.5	28	4.7	21	35	6.2	7.2	
33	32	31	31	30	30	29	29	28	28	28	28	
28	28	27	26	25	25	25	25	24	24	23	22	
3.9	5.7	36	42	4.2	50	5.2	8.3	29	30	3.6	40	
22	22	21	21	21	21	21	21	20	20	20	20	
7.3	8.1	34	(Other)	20	20	19	3307					
82949	993	640	434	351	311	279	233	209	206	197	157	
150	139	130	128	120	104	100	92	85	83	81	77	
1.2	1.4	17	2.3	15	16	1.8	1.9	1.6	1.3	2.8		
77	76	72	72	69	68	67	63	61	59	56		
2.7	14	2.5	2.6	13	3.2	18	22	23	3.3	20	3.7	
53	52	52	51	49	49	48	46	46	42	42		
2.2	25	3.8	5.1	4.4	7.6	26	27	4.6	2.1	33	3.4	
40	40	39	38	38	37	37	37	36	35	35	33	
3.5	19	32	2.9	5.5	28	4.7	21	35	6.2	7.2	7.5	
32	31	31	30	29	29	28	28	28	28	28		
9.9	5.6	6.6	23	4.5	5.9	7.4	31	38	4.3	3.1	3.9	
28	27	26	25	25	25	24	24	23	22	22		
5.7	36	42	4.2	50	5.2	8.3	29	30	3.6	40	7.3	
22	21	21	21	21	21	20	20	20	20	20		
8.1	34	10.1	(Other)	20	20	19	3789					

(a) Vérification de la dimension des données consolidées

(b) Traitement des cases vides des données consolidées

FIGURE 3.10 – Audit : Dimensions et données manquantes.

15. **R** est un langage de programmation, de traitement des données et d'analyse statistique mettant en œuvre le langage de programmation **S**, avec la sémantique dérivée du langage **Scheme**

Dans les figures 3.11a et 3.11b, Nous observons deux interprétation différentes de **Dubai** en ("*Dubai, UAE*", "*Dubai/ UAE*") et de **Taiwan** en ("*Taiwan, China*", "*Taiwan/ China*"). Ces discrépances seraient analysées en tant que 4 pays différents alors qu'en réalité, ceux-ci ne sont que deux.

Nous procérons à la correction de ces deux ambiguïtés dans les figures 3.11c et 3.11d respectivement.

Un malfonctionnement similaire est à noter pour les drapeaux de la colonne *P2O5.Product*¹⁶ où des caractères "espace" se sont introduits en début de chaîne pour proposer ici aussi 4 niveaux qualitatifs au lieu de 2 uniquement.

Nous diagnostiquons ce soucis et le corrigeons dans la figure 3.11e

Console ~/ ↵	> sort(unique(f\$Importing.countries))		
	[1] Abu Dhabi, UAE	Afghanistan	Switzerland
	[4] Algeria	Angola	Taiwan/ China
	[7] Australia	Austria	Thailand
	[10] Bahamas	Bangladesh	Tunisia
	[13] Belgium	Belize	Ukraine
	[16] Bolivia	Brazil	USA
	[19] Cameroon	Canada	Various Central Europe
	[22] China	Colombia	Various Latin America
	[25] Costa Rica	Cote d'Ivoire	Various South Asia
	[28] Cuba	Cyprus	Various West Europe
	[31] Denmark	Djibouti	Yemen
	[34] Dominican Rep.	Dubai, UAE	Zimbabwe
		163 Levels: Abu Dhabi, UAE Afghanistan Albania Algeria Angola Argentina Australia Austria Azerbaijan Bahamas Bangladesh ... Zimbabwe

Console ~/ ↵	> sort(unique(f\$Importing.countries))		
	[1] Taiwan, China	Tajikistan	Taiwan, China
	[133] Taiwan/ China	Togo	Tanzania
	[136] Thailand	Turkey	Trinidad and Tobago
	[139] Tunisia	United Kingdom	Turkmenistan
	[142] Ukraine	Uzbekistan	Uruguay
	[145] USA	Various Africa	Various E. Europe & I.
	[148] Various Central Europe	Various East Asia	Various Others
	[151] Various Latin America	Various Oceania	Various West Asia
	[154] Various South Asia	Various Subtotal	Vietnam
	[157] Various West Europe	Venezuela	Zambia
	[160] Yemen	Zaire	
	[163] Zimbabwe		
	163 Levels: Abu Dhabi, UAE Afghanistan Albania Algeria Angola Argentina Australia Austria Azerbaijan Bahamas Bangladesh ... Zimbabwe		

(a) Ambiguïté Dubai, UAE

(b) Ambiguïté Taiwan, China

```

Console ~/ ↵
> sort(unique(f$Importing.countries))[36]
[1] Dubai/ UAE
163 Levels: Abu Dhabi, UAE Afghanistan Albania Algeria Angola Argentina Australia Austria Azerbaijan Bahamas Bangladesh ... Zimbabwe
> sort(unique(f$Importing.countries))[35]
[1] Dubai, UAE
163 Levels: Abu Dhabi, UAE Afghanistan Albania Algeria Angola Argentina Australia Austria Azerbaijan Bahamas Bangladesh ... Zimbabwe
> f[which(f$Importing.countries == sort(unique(f$Importing.countries))[35]),]$Importing.countries=sort(unique(f$Importing.countries))[36]
> sort(unique(f$Importing.countries))[35]
[1] Dubai/ UAE
163 Levels: Abu Dhabi, UAE Afghanistan Albania Algeria Angola Argentina Australia Austria Azerbaijan Bahamas Bangladesh ... Zimbabwe
> sort(unique(f$Importing.countries))[36]
[1] Ecuador
163 Levels: Abu Dhabi, UAE Afghanistan Albania Algeria Angola Argentina Australia Austria Azerbaijan Bahamas Bangladesh ... Zimbabwe
> |

```

(c) Correction Dubai, UAE

```

Console ~/ ↵
> sort(unique(f$Importing.countries))[131]
[1] Taiwan, China
163 Levels: Abu Dhabi, UAE Afghanistan Albania Algeria Angola Argentina Australia Austria Azerbaijan Bahamas Bangladesh Belarus Belize ... Zimbabwe
> sort(unique(f$Importing.countries))[132]
[1] Taiwan/ China
163 Levels: Abu Dhabi, UAE Afghanistan Albania Algeria Angola Argentina Australia Austria Azerbaijan Bahamas Bangladesh Belarus Belize ... Zimbabwe
> f[which(f$Importing.countries == sort(unique(f$Importing.countries))[131]),]$Importing.countries=sort(unique(f$Importing.countries))[132]
> sort(unique(f$Importing.countries))[131]
[1] Taiwan/ China
163 Levels: Abu Dhabi, UAE Afghanistan Albania Algeria Angola Argentina Australia Austria Azerbaijan Bahamas Bangladesh Belarus Belize ... Zimbabwe
> sort(unique(f$Importing.countries))[132]
[1] Tajikistan
163 Levels: Abu Dhabi, UAE Afghanistan Albania Algeria Angola Argentina Australia Austria Azerbaijan Bahamas Bangladesh Belarus Belize ... Zimbabwe
> |

```

(d) Correction Taiwan, China

```

Console ~/ ↵
> v15=read.csv('/media/moriarty/YACHAOUI/V15.csv')
> levels(v15$P2O5.Product)
[1] "P2O5"      "P2O5"      "Product"    "Product"
> v15$which(v15$P2O5.Product == ' P2O5 '),]$P2O5.Product = 'P2O5'
> v15$which(v15$P2O5.Product == ' Product '),]$P2O5.Product = 'Product'

```

(e) Correction des drapeaux de la colonne *P2O5.Product*

FIGURE 3.11 – Audit : Interprétation des chaînes de caractères.

16. Drapeau indiquant si le poids indiqué à la colonne *kTest* net en P2O5 ou en poids brut.

X	Importing.countries	Region..IFA.	Region..OCP.	Exporting.countries	Product	Year	Quarter
Min. : 1	Bulgaria : 1012	Latin America :18458	North Europe & FSU :28712	China :10920	DAP :27608	Min. :2014	Q1:19680
1st Qu.:23561	Colombia : 1012	Africa :17574	Latin America :18458	Various: 7120	MAP :21784	1st Qu.:2014	Q2:23714
Median :47120	France : 1012	West Europe :14300	Africa :17574	Morocco: 6708	PA :23392	Median :2014	Q3:26446
Mean :47120	Germany : 1012	East Asia :11696	South Asia & Middle East:12418	USA : 6584	Rock: 8496	Mean :2014	Q4:24400
3rd Qu.:70680	Indonesia: 1012	West Asia : 8686	East Asia :11696	Tunisia: 6252	TSP :12960	3rd Qu.:2014	
Max. : 94240	Japan : 1012	Central Europe: 7710	South & Central Europe : 7710	Jordan : 5092		Max. :2015	
(Other) :88168	(Other)	:15816	(Other)	: 5672	(Other):51564		
Code	KT	P205.Product	Cumulated	Not.Cumulated	AGG.DET.ANN		
Abu Dhabi, UAEWest AsiaSouth Asia & Middle EastAlgeriaRock2014Q2P205	: 2	Min. :-631.100	P205 :51368	Cumulated :47120	AGG:16848		
Abu Dhabi, UAEWest AsiaSouth Asia & Middle EastAlgeriaRock2014Q3P205	: 2	1st Qu.: 0.000	Product:42872	Not Cumulated:47120	ANN:21340		
Abu Dhabi, UAEWest AsiaSouth Asia & Middle EastAlgeriaRock2014Q4P205	: 2	Median : 0.000			DET:56052		
Abu Dhabi, UAEWest AsiaSouth Asia & Middle EastAustraliaDAP2014Q1P205	: 2	Mean : 3.227					
Abu Dhabi, UAEWest AsiaSouth Asia & Middle EastAustraliaDAP2014Q1Product:	: 2	3rd Qu.: 0.000					
Abu Dhabi, UAEWest AsiaSouth Asia & Middle EastAustraliaDAP2014Q2P205	: 2	Max. :3285.000					
(Other) :94228		:94228					

FIGURE 3.12 – Audit : Conversion des unités de mesures.

La figure 3.12 ci-dessous, relève un écart de $51368 - 42872 = 8496$ enregistrements-vente de plus possédant le drapeau *Product* et pas le drapeau *P205*. Ce chiffre nous interpelle puisque celui-ci est justement celui d'enregistrements-vente **Product == "Rock"**. Dans la figure 3.13, nous investiguons cette fausse coïncidence en nous nous intéressons uniquement aux enregistrements-vente **Product ≠ "Rock"** où il s'avère que cet écart n'est plus. Ceci nous amène à ajuster notre code de conversion présenté dans la phase d'unification à la section 3.1.2.3.

X	Importing.countries	Region..IFA.	Region..OCP.	Exporting.countries	Product	Year	Quarter
Min. : 1	Bulgaria : 896	Latin America :17356	North Europe & FSU :18356	China :10344	DAP :27608	Min. :2014	Q1:18224
1st Qu.:22893	Colombia : 896	Africa :15984	Latin America :17356	Various: 6940	MAP :21784	1st Qu.:2014	Q2:21724
Median :47388	France : 896	West Europe :12668	Africa :15984	USA : 6584	PA :23392	Median :2014	Q3:24456
Mean :47034	Germany : 896	East Asia :10588	South Asia & Middle East:11276	Morocco: 6132	Rock: 0	Mean :2014	Q4:21340
3rd Qu.:70814	Indonesia: 896	West Asia : 7892	East Asia :10588	Tunisia: 5676	TSP :12960	3rd Qu.:2014	
Max. : 94240	Japan : 896	Central Europe: 7096	South & Central Europe : 7096	Brazil : 5016		Max. :2015	
(Other) :80368	(Other)	:14160	(Other)	: 5088	(Other):45052		
Code	KT	P205.Product	Cumulated	Not.Cumulated	AGG.DET.ANN		
Abu Dhabi, UAEWest AsiaSouth Asia & Middle EastAustraliaDAP2014Q1P205	: 2	Min. :-631.100	P205 :42872	Cumulated :42872	AGG:16848		
Abu Dhabi, UAEWest AsiaSouth Asia & Middle EastAustraliaDAP2014Q1Product:	: 2	1st Qu.: 0.000	Product:42872	Not Cumulated:42872	ANN:21340		
Abu Dhabi, UAEWest AsiaSouth Asia & Middle EastAustraliaDAP2014Q2P205	: 2	Median : 0.000			DET:47556		
Abu Dhabi, UAEWest AsiaSouth Asia & Middle EastAustraliaDAP2014Q2Product:	: 2	Mean : 2.358					
Abu Dhabi, UAEWest AsiaSouth Asia & Middle EastAustraliaDAP2014Q3P205	: 2	3rd Qu.: 0.000					
Abu Dhabi, UAEWest AsiaSouth Asia & Middle EastAustraliaDAP2014Q3Product:	: 2	Max. :1563.000					
(Other) :		:85732					

FIGURE 3.13 – Audit : Correction des unités de mesures.

Audit manuel d'échantillons aléatoires :

Une vérification supplémentaire a été entreprise par nos soins, en choisissant aléatoirement :

- Une année d'exercice,
- Un semestre d'exercice,
- Un produit phosphaté,
- Un pays importateur.

Nous listons l'ensemble des imports de celui-ci que nous avons consolidé qu'on compare par la suite visuellement aux fichier .pdf de départ en termes de :

- Poids rapporté,
- Drapeaux et normes de l'enregistrement-vente,
- Adéquation des chaînes de synthèse de la colonne *Code*.

Pour un échantillon de 20 tirages, nous n'avons observé aucun écart.

3.1.3.2 Description des données de consolidation

Nous procédons dans la partie suivante à la visualisation exploratoire de nos données consolidées. En statistiques, l'analyse exploratoire des données (AED) est une approche de l'analyse des ensembles de données pour résumer leurs principales caractéristiques, souvent avec des méthodes visuelles. Un modèle statistique peut être utilisé ou non, mais surtout l'AED interrogent les données sur ce qu'elles peuvent nous dire au-delà de la modélisation formelle ou des tâches de vérification d'hypothèses. Le script en langage **R** de l'AED à suivre peut être retrouvé dans le dossier *Source* du dépôt **GitHub** de notre mémoire de projet de fin d'études[15].

Description des quantités exportées

La figure 3.14 ci-dessous est la représentation graphique des distributions des échanges mondiaux des produits. Nous nous intéressons ainsi à la colonne kT du datamart de la figure 3.3. Nous procédons en un premier lieu à la transformation logarithmique de cette colonne en vue d'atténuer les écarts entre les valeurs les plus extrêmes et produire des diagramme visuellement plus plaisants offrant plus de flexibilité à l'interprétation. Les diagramme de la figure 3.14 sont communément appelés des *Boites à Moustaches* et résument l'information de dispersion des valeurs tracées par une boite (en rouge) délimitée par l'écart interquartile $Q_3 - Q_1$ représentant les enregistrements séparant les 25% des valeurs les plus extrêmes et dont la ligne centrale représente la moyenne de l'échantillon. Les moustaches de la boite permettent de se prononcer quant aux valeurs aberrantes : sont coloriées en Orange, les valeurs dépassant les limites des moustaches dont la longueur est $(1.5 \times Q_3 - Q_1)$.

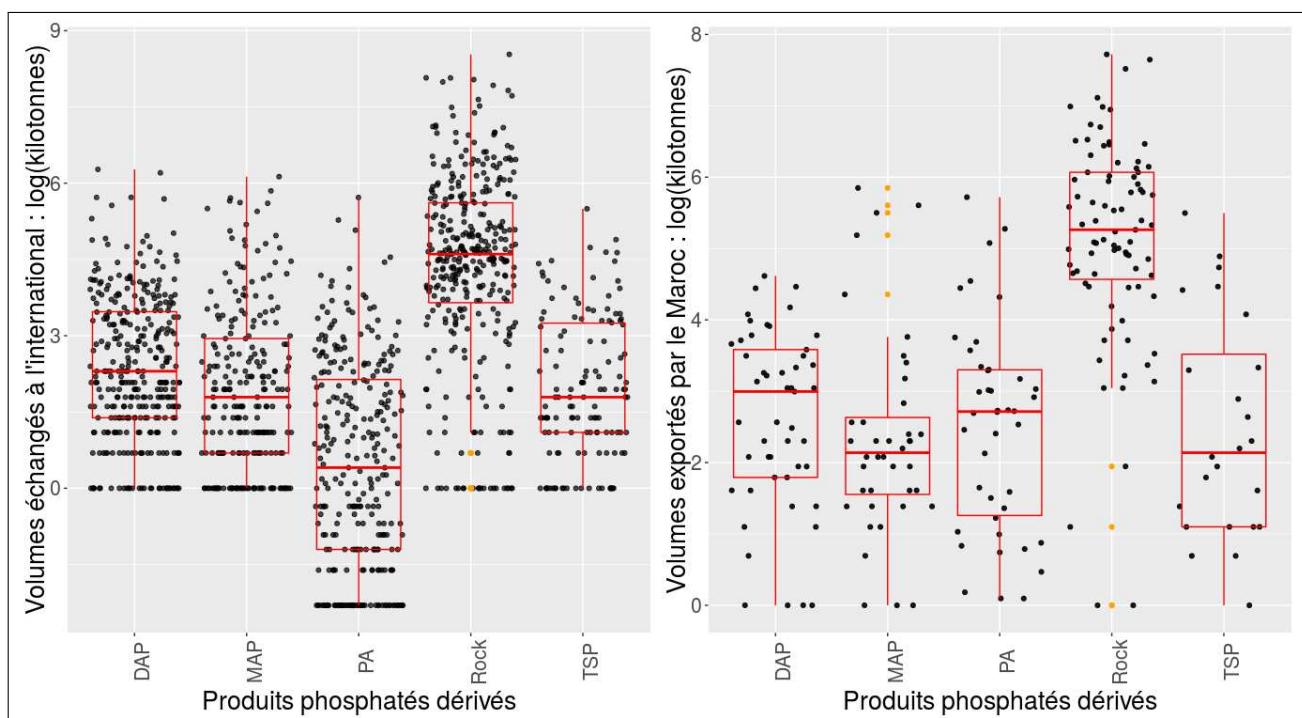
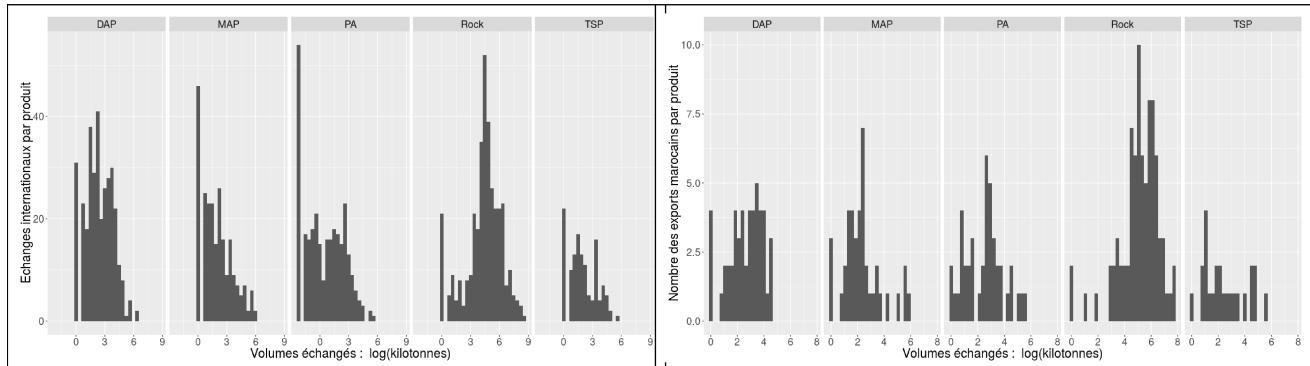


FIGURE 3.14 – Boîtes à moustaches des quantités logarithmiques exportées par produit phosphaté. Gauche : Quantités mondiales. Droite : Exports Marocains.

De par les lignes centrales des boîtes, il s'avère que le Maroc exporte nettement plus que la moyenne mondiale en acide phosphorique (PA), en roche (Rock) ; dépasse sensiblement les moyennes DAP et MAP, et sous-performe dans les exports des TSP. Nous prêterons plus d'attention aux histogrammes de la figure 3.15 ci-dessous en les contextualisant au sein des rôles des exportateurs majeurs de produits phosphatés. Nous nous suffirons ici de relever que ces histogrammes confirment bien la rareté des valeurs aberrantes après notre transformation logarithmique :



(a) Histogramme des quantités logarithmiques mondiales exportées par produit.
(b) Histogramme des exports logarithmiques Marocains par produit.

FIGURE 3.15 – Histogramme des quantités logarithmiques exportées par produit.

Description des exportateurs majeurs

Nous empilons les histogrammes des produits de la figure 3.15 en coloriant les barres selon le produit concerné (figure 3.16a) et les exportateurs les ayant émis (figure 3.16b). La figure 3.16 permet les observations suivantes :

Observations relatives au type de produits exportés :

- Les volumes d'acide phosphorique sont échangés par lot de petites quantités. Ceci s'explique d'une part par la spécificité des besoins en acide qui ne couvrent qu'un ensemble réduit d'industries ; et d'une autre part par les précautions contraignantes à observer durant le fret de l'acide [3]. La figure 3.16c montre bien un pic au niveau des plus faibles valeurs des quantités exportées.
- Les fertilisants DAP, MAP et TSP sont similairement centrés. L'explication est à corrélérer avec la généricité des applications agricoles de ceux-ci ainsi que le volume transportable par les cargo maritimes de type *Panamax*¹⁷.
- Les volumes des lot de roche de phosphate s'articule largement au-dessus du reste des produits et est possiblement imputable au format de transport de la roche¹⁸ qui ne nécessite pas de dispositions particulières lors du transport maritime ainsi que la nécessité pour

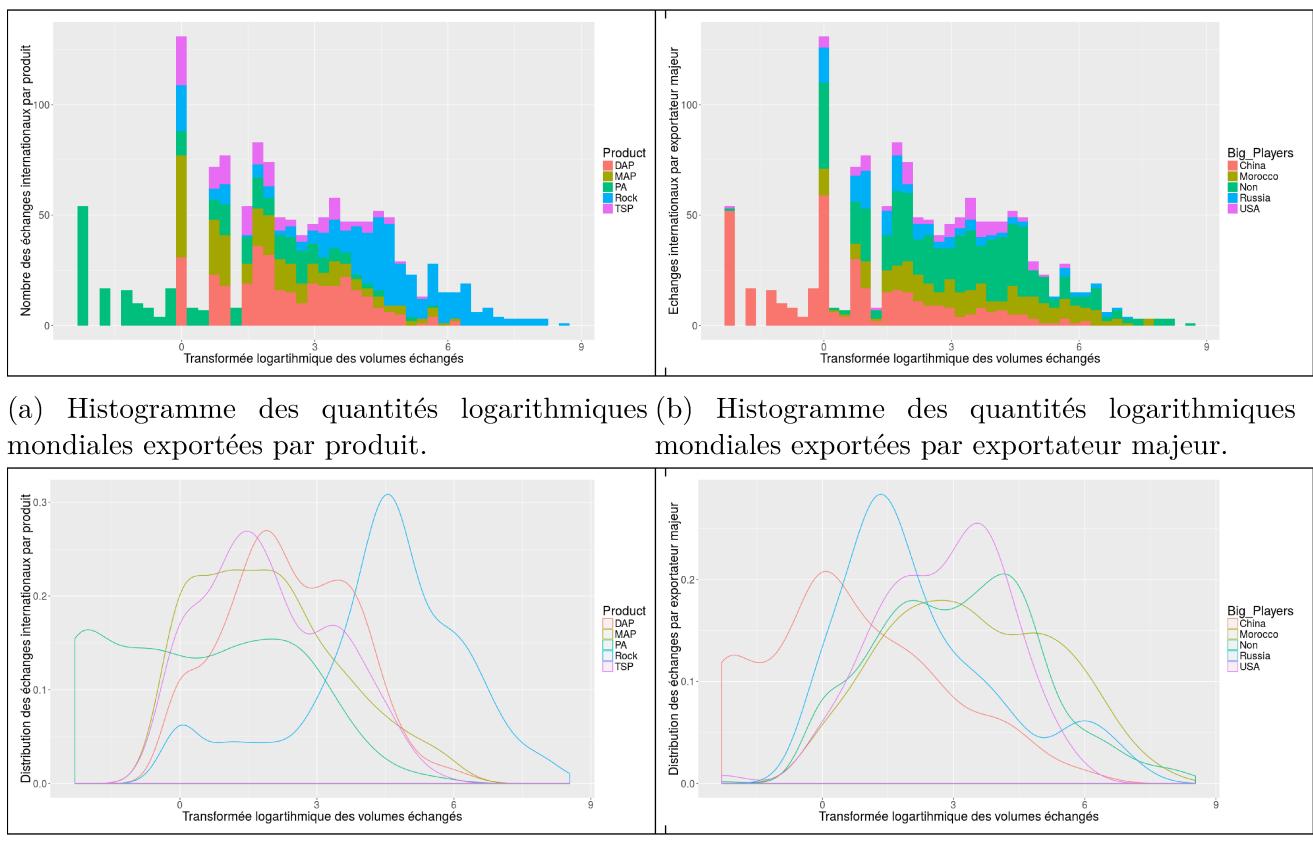
17. Les navires classés comme Panamax ont les dimensions maximum pour rentrer dans les écluses du canal de Panama. La majorité des navires sont conçus à la limite de cette taille. Pour le transport des fertilisants phosphatés, ceux-ci sont chargés à la hauteur de $55kT = e^{1.74}kT$ [21]

18. Le format de transport de la roche en vrac permet d'atteindre des chargements Panamax de $56000 kT = e^{4.75}kT$ [23]

les pays importateurs de disposer d'usine de traitement de la roche, exemple d'industrie qui acquiert sa matière première en contrat annuels[21].

Observations relatives aux positionnement des exportateurs majeurs :

- La Chine s'est forgée un marché de lots d'export de faibles quantités mais plus grand en fréquence. Ce positionnement est représenté par la courbe chinoise sur la figure 3.16b et éclairci par la figure 3.17 qui présente une Chine dominant le marché asiatique caractérisé par un nombre élevé de pays-clients différents.
- Dans les livraisons à volumes intermédiaires, les USA adoptent une politique unimodale caractérisée par des exports en forme quasi-gaussienne, indicateur d'une stabilité des volumes de production.
- La Russie présente le même positionnement que la moyenne du marché avec des courbes calquées sur la figure 3.16b avec une tendance des Russes à décaler leur exports vers des volumes légèrement supérieurs à la moyenne. C'est la position du *Market Follower*¹⁹
- Dans les livraisons à grand volume, le Maroc est incontestablement favori, surclassant ainsi la Russie et les USA dans la figure 3.16b. C'est la position du *Price Setter* qui s'affirme.



(c) Densités des quantités logarithmiques mondiales exportées par produit.

(d) Densités des quantités logarithmiques mondiales exportées par exportateur majeur.

FIGURE 3.16 – Distributions des quantités logarithmiques mondiales exportées par produit et par exportateur majeur.

19. Market Follower : se dit d'une entreprise qui permet à des firmes plus dominantes de décider des prix du marché.

Description des régions importatrices

Deux facteurs semblent peser sensiblement dans l'acquisition des parts de marchés par les différents exportateurs. D'abord la distance de ceux-ci par rapport aux régions de destinations. Ensuite les atouts politiques dont jouissent les exportateurs majeurs sur le marché. Nous dressons à l'appréciation du lecteur la figure 3.17 ci-dessous en attirons son attention sur les quatres exemples suivants :

- La Chine s'accapare les marchés est et sud asiatique.
- La Russie domine les pays FSU²⁰ bien que ceux-ci soit proches des exportateurs européens.
- Le marché américain latin et nord privilégie les USA.
- Le Maroc présent sur l'ensemble du marché dans les tranches supérieurs des exports tous produits confondus.

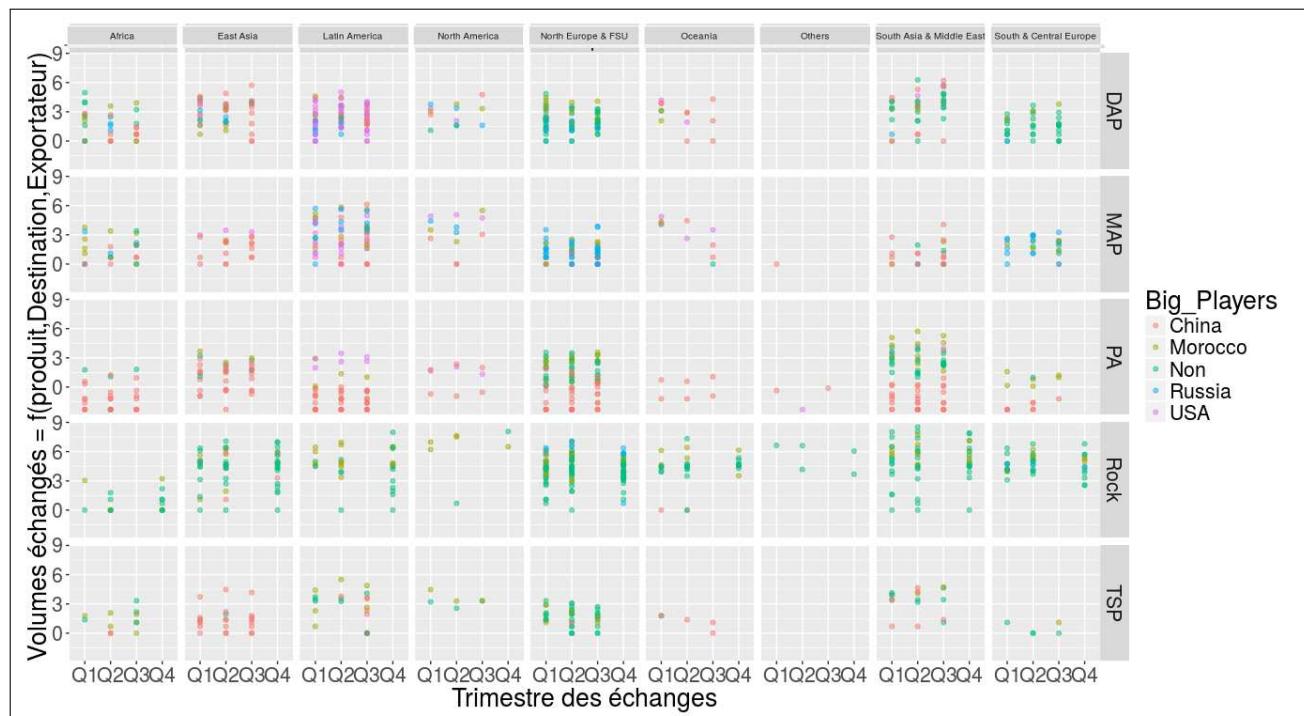


FIGURE 3.17 – Parts de marché des produits phosphatés par région de destination et exportateur majeur

20. Ex-URSS

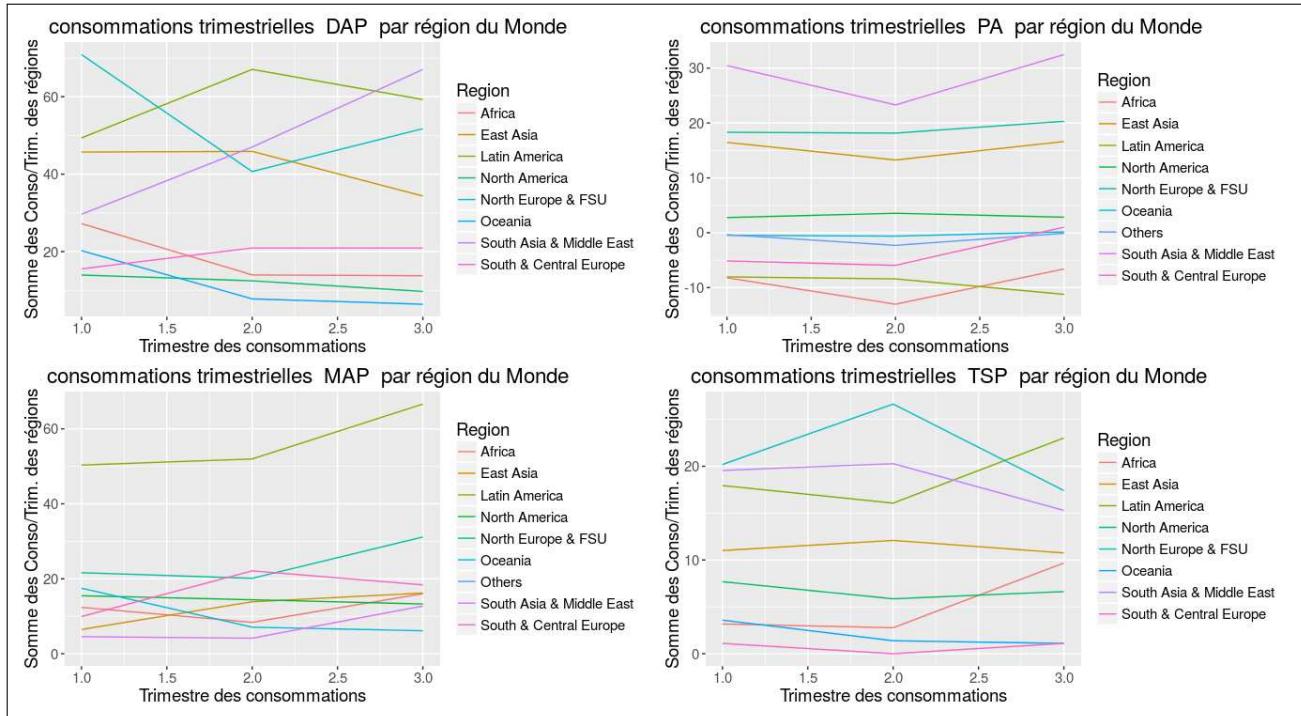


FIGURE 3.18 – Sommes trimestrielles des exports par région de destination

3.2 Collecte et préparation des données externes

L'étendue des données disponibles au sein du système d'information de l'OCP ne pouvant donner qu'une vue réduite de la situation du marché et ne peuvent de par leur définition révéler les structures socio-économiques, et de politiques agraires sous-jacentes aux demandes en produits phosphatés, nous nous attellerons à la tâche d'enrichir cette base de données historiques par des données quantitatives concernant les pays du monde.

Nous décrivons dans ce qui suit ce processus.

3.2.1 Énumération et collecte des données externes

Guidés par nos lectures bibliographiques résumées dans la section 2.2.2, nous utilisons l'API de la banque de données publiques WorldBank²¹ qui présente une interface RESTful et nous permet de récolter les indicateurs de politiques agraires et socioéconomiques suivants :

- **Access to electricity, rural (% of rural population)** : Fraction de paysans ayant accès à l'électricité.
- **Access to non-solid fuel, rural (% of rural population)** : Fraction de paysans ayant accès aux fuels non solides.
- **Account at a financial institution (% age 15+)** : Fraction de personnes âgées de +15 ans ayant un compte chez une institution bancaire.

21. <http://data.worldbank.org/developers?display=>

- **Net enrollment rate, primary (% of primary school age children)** : Fraction d'enfants scolarisés.
- **Net national income per capita (constant 2005 US\$)** : PIB²² par habitant, inflation ajustée au dollar US fin 2005.
- **Literacy rate, adult total (% of people ages 15 and above)** : Fraction de personnes alphabétisées âgées de +15ans.
- **Agricultural irrigated land (% of total agricultural land)** : Fraction de terres irriguées parmi les terres exploitées pour l'agriculture.
- **Agricultural land (% of land area)** : Fraction des terres agricoles de la surface totale du pays.
- **Agricultural tractors per 100 sq. km of arable land** : Nombre de tracteurs par 100 km² de terres arables.
- **Agriculture, value added (% of GDP)** : Fraction de la valeur ajoutée agricole du PIB.
- **Agriculture value added per worker (constant 2005 US\$)** : Valeur ajoutée agricole par ouvrier agricole, inflation ajustée au dollar US fin 2005.
- **All education staff compensation, total (% of total expenditure in public institutions)** : Fraction des dépenses en éducation des dépenses en institutions publiques.
- **Annual freshwater withdrawals, agriculture (% of total freshwater withdrawal)** : Fraction du volume d'eau utilisée à des fins agricoles de la totalité de l'eau consommée.
- **Arable land (% of land area)** : Fraction de terres arables de la surface totale du pays.
- **Arable land (hectares per person)** : Nombre d'hectares de terres arables par personne.
- **Birth rate, crude (per 1,000 people)** : Nombre de naissances par 1000 personnes.
- **Cereal yield (kg per hectare)** : Rendement des céréales en kilogramme par hectare.
- **Commercial bank branches (per 100,000 adults)** : Nombre d'agences bancaires par 100,000 adultes.
- **Consumer price index (2010 = 100)** : L'indice des prix à la consommation (IPC) mesure l'évolution du niveau moyen des prix des biens et services consommés par les ménages, pondérés par leur part dans la consommation moyenne des ménages. Harmonisé pour permettre une comparaison entre les pays à fin 2010.
- **Cost to export (US\$ per container)** : Coût en Dollars US de l'export d'un conteneur de marchandises.

22. Produit intérieur brut, un des agrégats majeurs des comptes nationaux, il vise à quantifier — pour un pays et une année donnés — la valeur totale de la « production de richesse » effectuée par les agents économiques résidant à l'intérieur de ce territoire (ménages, entreprises, administrations publiques).

- **Cost to import (US\$ per container)** : Coût en Dollars US de l'import d'un conteneur de marchandises.
- **Crop production index (2004-2006 = 100)** : L'indice de production des cultures montre la production agricole pour chaque année par rapport à la période de base de 2004 à 2006. Cet indice porte sur l'ensemble des cultures à l'exception des cultures fourragères. Les regroupements par région et par revenu des indices de production de la FAO sont calculés à partir des valeurs sous-jacentes en dollars US et normalisés par rapport à la période de référence de 2004 à 2006.
- **Droughts, floods, extreme temperatures (% of population, average 1990-2009)** : Pourcentage moyen annuel entre 1990 et 2009 de la population affectée par les catastrophes naturelles classifiées comme sécheresses, inondations et événements climatiques extrêmes.
- **Employment in agriculture (% of total employment)** : Fraction des ouvriers agricoles de l'ensemble des employés.
- **Food production index (2004-2006 = 100)** : L'indice de production alimentaire porte sur les cultures vivrières qui sont considérées comme comestibles et qui contiennent des nutriments et normalisées par rapport à la période de référence de 2004 à 2006.
- **GDP per capita (constant 2005 US\$)** : PIB par habitant. Inflation ajustée à fin 2005.
- **Household final consumption expenditure (constant 2005 US\$)** : La consommation privée désigne la valeur marchande de tous les biens et services, y compris les produits durables achetés par les ménages.
- **Lending interest rate (%)** : Le taux d'intérêt perçu par les banques sur les prêts accordés aux clients.
- **Life expectancy at birth, total (years)** : L'espérance de vie à la naissance indique le nombre d'années qu'un nouveau-né devrait vivre si les règles générales de mortalité au moment de sa naissance devaient rester les mêmes tout au long de sa vie.
- **Livestock production index (2004-2006 = 100)** : L'indice de production animale comprend la production de viande et de lait de toutes sources, les produits laitiers tels que le fromage, les œufs, le miel, la soie brute, la laine ainsi que les peaux et les cuirs.
- **Logistics performance index** : La note globale de l'indice de performance de la logistique reflète les perceptions relatives à la logistique d'un pays basées sur l'efficacité des processus de dédouanement, la qualité des infrastructures commerciales et des infrastructures de transports connexes, la facilité de l'organisation des expéditions à des prix concurrentiels, la qualité des services d'infrastructure, la capacité de suivi et de traçabilité des consignations et la fréquence avec laquelle les expéditions arrivent au destinataire dans les délais prévus. L'indice varie continuellement de 1 à 5 et la note la plus élevée représente la meilleure performance.

- **Low-birthweight babies (% of births)** : Fraction des nouveau-nés pesant moins de 2 500 grammes des naissances totales.
- **Net migration** : Nombre d'immigrants total moins le nombre d'émigrants annuel, comprenant à la fois les citoyens et les non citoyens.
- **Permanent cropland (% of land area)** : Fraction des terres occupées par des cultures pour de longues périodes et qui doivent être replantées après chaque récolte de la surface totale du pays.
- **Population density (people per sq. km of land area)** : Densité des habitants en personne par km².
- **Population growth (annual %)** : Croissance relative annuelle de la population.
- **Rural population (% of total population)** : Fraction rurale de la population.
- **Rural poverty gap at national poverty lines (%)** : L'écart de pauvreté par rapport au seuil national de la pauvreté en milieu rural est le manque à gagner pour remonter au-dessus du seuil de la pauvreté (en considérant que les non pauvres ont un manque à gagner de zéro) exprimé en pourcentage du seuil national de la pauvreté en milieu urbain. Cette mesure témoigne à la fois de l'ampleur de la pauvreté et de sa fréquence.
- **Unemployment, total (% of total labor force)** : Fraction de la population active qui est sans emploi mais qui est disponible pour et à la recherche d'un emploi.

3.2.2 Préparation des données externes

Cette section s'intéresse à l'application des forêts de décisions aléatoires à des fins de sélection de variables. Le but ici est double : d'abord introduire le comportement de l'indexation de l'importance des variables en utilisant les forêts aléatoires et l'utiliser pour proposer un algorithme à deux phases pour la sélection de variables à la base de leur importance.

La stratégie générale se résume en un classement des variables exogènes²³. en utilisant le score d'importance de ces variables introduit par les forêts aléatoires puis une sélection ascendante itérative des variables.

3.2.2.1 Introduction à la selection de variables via forêts de décision aléatoires

Les FA est un algorithme populaire et très efficient appartenant aux méthodes d'agrégation pour les problèmes de régression et de classification, introduit par Breiman[22], et apparaît dans les application de 'Machine Learning'²⁴ à la fin du dernier millénaire[25]. Les FA deviennent de plus en plus populaires et semblent être très robustes dans beaucoup d'applications bien qu'ils ne soient pas clairement théorisés mathématiquement[8]. Une introduction sommaire des FA est donnée en annexe A1.

23. à savoir les variables externes énumérées dans la section 3.2.1

24. L'apprentissage automatique ou apprentissage statistique (machine learning en anglais), champ d'étude de l'intelligence artificielle, concerne la conception, l'analyse, le développement et l'implémentation de méthodes permettant à une machine (au sens large) d'évoluer par un processus systématique, et ainsi de remplir des tâches difficiles ou impossibles à remplir par des moyens algorithmiques plus classiques.

Le principe des FA est de combiner plusieurs arbres (\hat{e}_i) de décision CART[11] en utilisant plusieurs échantillons bootstrap de l'ensemble d'apprentissage L_n et de choisir aléatoirement à chaque nœud un sous-ensemble de k variables exogènes X_i .

L'agrégation à laquelle procèdent les FA est d'autant plus performante que la corrélation entre les prédicteurs agrégés (arbres CART) est faible. Afin de diminuer cette corrélation, Breiman[22] propose de rajouter une couche d'aléa dans la construction des prédicteurs. Nous attirons l'attention du lecteur vers l'annexe A2 pour une note sur CART et son utilisation au sein des FA. Sommairement, à chaque étape de CART, k variables sont sélectionnées aléatoirement parmi les p et la meilleure coupure est sélectionnée uniquement sur ces k variables :

Algorithme Forêts aléatoires

Entrées :

- x , une nouvelle observation à prévoir.
- L_n , l'échantillon.
- B , le nombre d'arbres.
- $k \in \mathbb{N}^*$, le nombre de variables candidates pour découper un nœud.

Pour $i = 1, \dots, B$:

- Tirer un échantillon bootstrap dans L_n .
- Construire un arbre CART sur cet échantillon bootstrap, chaque coupure est sélectionnée en minimisant la fonction de coût de CART sur un ensemble de k variables choisies au hasard parmi les p . On note $\hat{e}(., \theta_i)$ l'arbre construit.

Sortie : L'estimateur $\hat{e}(x) = \frac{1}{B} \sum_{i=1}^B \hat{e}_i(x, \theta_i)$

3.2.2.2 Procédure de sélection des variables et élagage des données externes

Parmi les nombreuses sorties proposées par la fonction randomForest, deux se révèlent particulièrement intéressantes. **L'erreur Out-of-Bag** et **Le score FA d'importance des variables**. Nous définissons celles-ci rigoureusement dans l'annexe A3. Ces deux sorties nous permettent de proposer la procédure suivante à deux phases pour l'élimination itérative des variables les moins pertinentes :

Phase 1. Classement et élimination initiale :

- Nous modélisons par FA notre ensemble de données initial, comprenant la totalité des variables exogènes listées dans la section 3.2.1.
- Nous calculons les score FA d'importance des variables et nous éliminons les variables d'importance triviale.
- Nous classons les m variables restantes par ordre descendant des score FA d'importance

Phase 2. Sélection des variables :

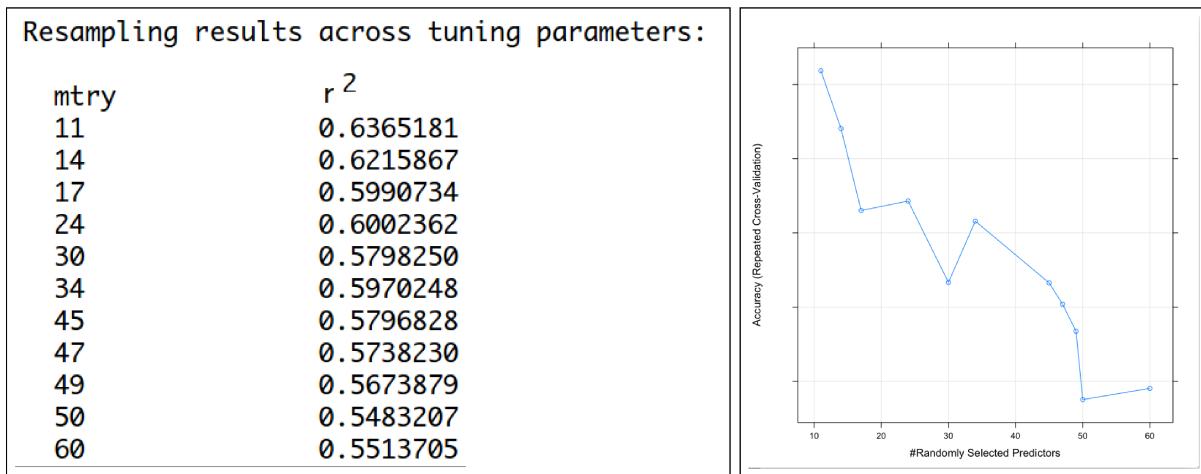
- Nous construisons la collection itérative des modèles de FA employant les α premières variables selon leur classement de score d'Importance. Pour $\alpha \in [1, m]$.
- Nous retenons les variables du modèle ayant réussi la plus faible erreur OOB.

La mise en œuvre de cette procédure de sélection des variables peut être retrouvée écrite en langage **R** dans le dossier *Source* du dépôt **GitHub** de notre mémoire de projet de fin d'études[15].

3.2.2.3 Mise en œuvre de la sélection des variables et élagage

Comme discuté en Annexe A2, la nature de l'algorithme des FA fait que le paramètre k est plus critique que le nombre B d'arbres qui lui devrait être choisi aussi grand que les limites calculatoires le permettent. Nous commençons d'abord par trouver la valeur k maximisant le coefficient d'ajustement R^2 pour une prédiction via FA de la **Consommation de kilogrammes de fertilisants phosphatés par hectare de terre arable** par la liste des **39 variables exogènes** listées à la section 3.2.1.

Recherche aléatoire de k : Puisque nous n'avons pas une idée de l'intervalle sur lequel le meilleur k pourrait se trouver, nous commençons d'abord par essayer des valeurs aléatoires sur un intervalle. De par la nature du faible nombre de pays, 214, pour 52 caractères quantitatifs, nous ne pouvons pas nous permettre de retenir un ensemble de test. Nous utilisons ainsi 3 répétitions d'une validation-croisée sur 10 dossiers pour minimiser les risques de sur-apprentissage.



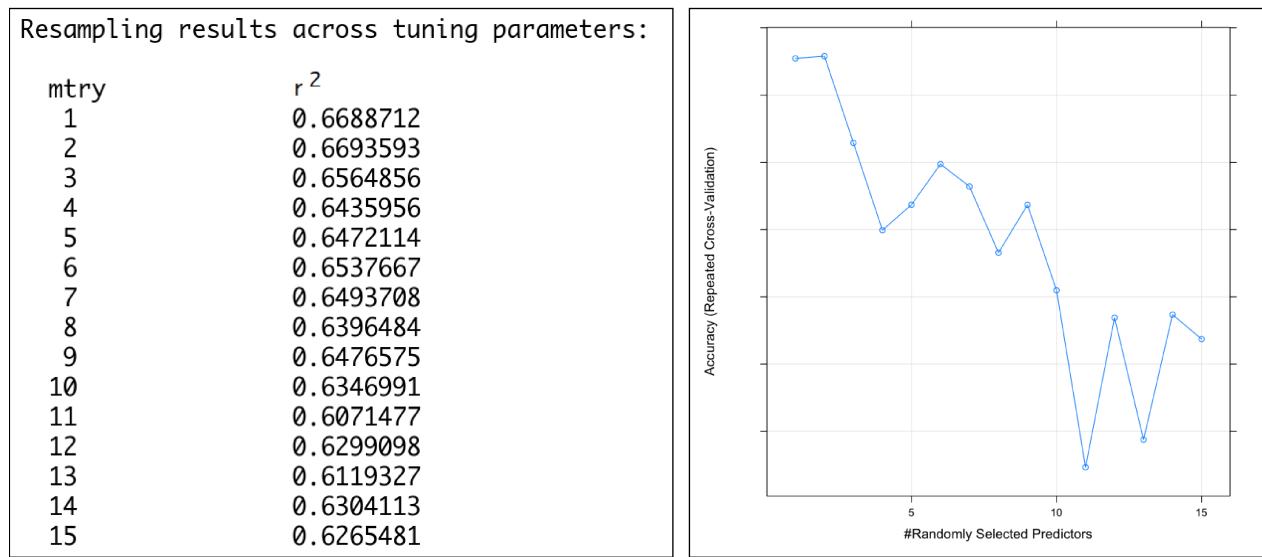
(a) Valeurs du R^2 selon k

(b) Graphe de R^2 selon k

FIGURE 3.19 – Recherche aléatoire du paramètre k (*mtry*)

Recherche séquentielle de k : D'après les valeurs prises par R^2 sur la figure 3.19, il nous paraît que le meilleur k se trouve quelque part dans les plus petites valeurs. Nous définissons une grille des 15 premières valeurs que nous parcourons à la recherche du k optimal sur la figure 3.20.

Il s'avère ainsi que la meilleure valeur de k est 2. Nous utiliserons ce paramètre lors des deux phases de notre procédure de sélection de variables.



(a) Valeurs du R^2 selon k

(b) Graphe de R^2 selon k

FIGURE 3.20 – Recherche séquentielle du paramètre k (*mtry*)

Phase 1. Classement et élimination initiale : Le graphe de la figure 3.21 suivant trace les variations de l'importance des variables de l'ajustement FA de la *Consommation de kilogrammes de fertilisants phosphatés par hectare de terre arable* par la liste des **39 variables exogènes** listées à la section 3.2.1 et classées par ordre décroissant de score d'importance.

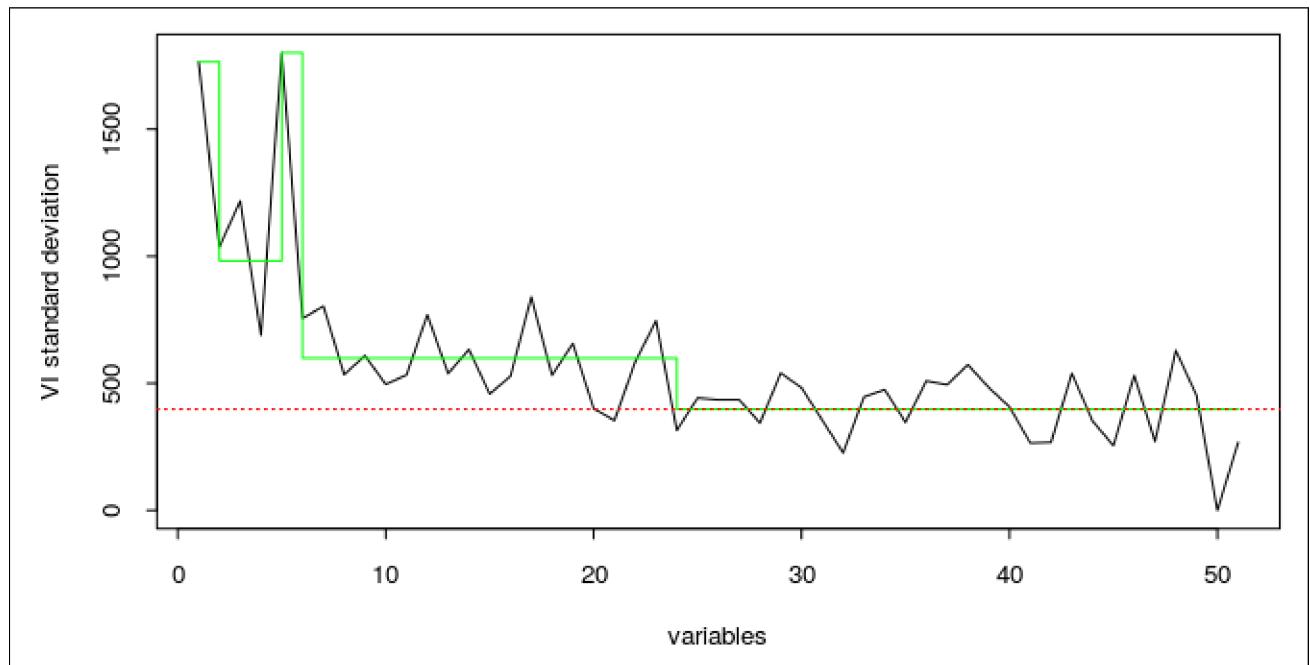


FIGURE 3.21 – Variation de l'importance des variables exogènes.

Nous avons tracé en vert le profil de la variation des score d'importance pour décider quant aux variables dont l'importance triviale nous pousse à éliminer. Nous nous arrêtons ainsi à la

$m = 24$ ème variable la plus importante et éliminons $15 = 39 - 24$ variables triviales d'embûche. La liste classée par ordre décroissant de score d'importance est donnée par la figure 3.22.

variable	relative_importance	scaled_importance
1 Population.growth..annual...	687827779584	
2 Agriculture..value.added....of.GDP.	341940600832	0.497131129305081
3 Unemployment..total....of.total.labor.force.	208326852608	0.302876474012138
4 Rural.population....of.total.population.	121729777664	0.17697711734996
5 Arable.land..hectares.per.person.	107707949056	0.156591449564805
6 Cost.to.import..US..per.container.	95595347968	0.138981516602624
7 Cost.to.export..US..per.container.	69183307776	0.100582311196914
8 GDP.per.capita..constant.2005.US..	67876417536	0.098682285814411
9 Logistics.performance.index..Overall..1.low.to.5.high.	58593656832	0.085186522805807
10 Population.density..people.per.sq..km.of.land.area.	56164171776	0.0816544103670373
11 Cereal.yield..kg.per.hectare.	55453200384	0.0806207629728742
12 Life.expectancy.at.birth..total..years.	37920976896	0.0551314995142166
13 Agricultural.land....of.land.area.	34348466176	0.0499375965838048
14 Arable.land....of.land.area.	32613332992	0.0474149692118056
15 Permanent.cropland....of.land.area.	31620321280	0.0459712768494521
16 Lending.interest.rate....	29810638848	0.0433402658817147
17 Agriculture.value.added.per.worker..constant.2005.US..	29585426432	0.0430128402925124
18 Birth.rate..crude..per.1.000.people.	29241864192	0.0425133515393716
19 Crop.production.index..2004.2006...100.	27676928000	0.0402381654558051
20 Livestock.production.index..2004.2006...100.	27027683328	0.0392942595955435
21 Food.production.index..2004.2006...100.	24201277440	0.0351850828918788
22 Commercial.bank.branches..per.100.000.adults.	21092128768	0.030664839941121
23 Rural.land.area..sq..km.	9223297024	0.0134093116006136
24 Consumer.price.index..2010...100.	231193.15625	3.36120702176097e-07

FIGURE 3.22 – Classement par ordre d'importance des 24 premières variables

Phase 2. Sélection des variables : Durant cette phase nous procéderons à l'ajustement de 24 modèles de FA. Chaque modèle FA_i , $i \in [1, 24]$ intégrera les i variables les plus importantes comme classées sur la figure 3.22. Nous choisissons le modèle j ayant réalisé la plus faible erreur OOB. Ce modèle déterminera le nombre j de variables que nous conserverons de nos données externes réalisant ainsi l'élagage des variables non importantes.

La figure 3.23 ci-après trace l'évolution de l'erreur OOB pour chaque modèle FA_i . Nous traçons en rouge l'abscisse du modèle $j = 12$.

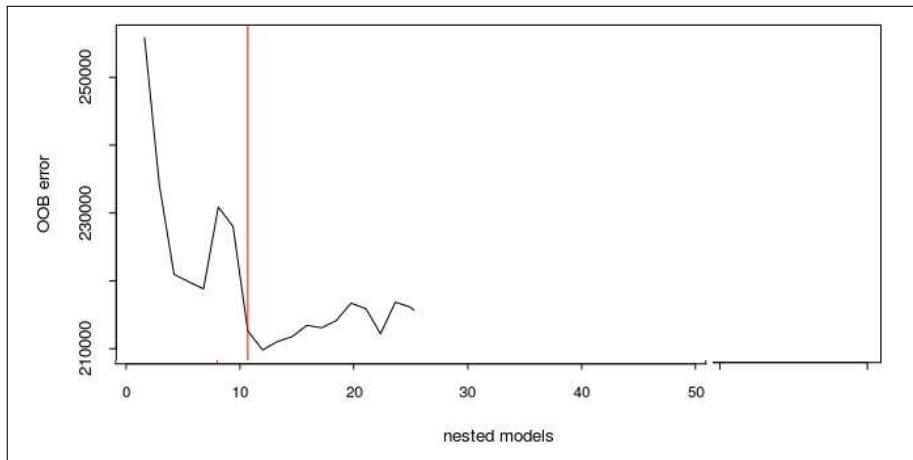


FIGURE 3.23 – Graphe de l'évolution de l'erreur OOB dépendamment du nombre de variables exogènes intégrées.

CHAPITRE 4

ANALYSE CAUSALE

Economists and agronomists are locked in debate about likely future yields.

Since the method of the economists is to predict future outcomes from past performance, economists expect success to continue. And since for the scientists future success depends on discoveries they will have to make and do not now know how to make, the scientists are doubtful. At its core, this is a disagreement about the pace of technical change.

Robert Socolow

4.1 Introduction à la régression en composantes principales

Les méthodes de régression multivariée comme la RCP¹ jouissent d'une large popularité dans divers domaines d'études, y compris les sciences naturelles. La raison principale de ce succès étant qu'elles ont été conçues dans l'optique de faire face à des situations où on dispose de plusieurs variables exogènes, généralement hautement corrélées, et relativement peu d'échantillons. Une situation qui est justement notre problématique. Ainsi une forme de réduction de dimensions dans l'espace des exogènes peut grandement simplifier le problème de régression. L'analyse en composantes principales est une technique de réduction de dimensions où une combinaison linéaire de p variables indépendantes/orthogonales des paramètres intrants X_1, X_2, \dots, X_p sont créées de façon à ce que la première combinaison linéaire Z_1 (composante principale) capture autant de la variance totale des données d'origine. La 2^{de} Z_2 capture autant que possible du reste de la variance sous la réserve que celle-ci soit orthogonale (i.e. non corrélée) à la première. En général toutes les variables intrantes sont standardisées de façon à posséder une moyenne nulle et une variance unitaire que l'on notera X_j^* . Ainsi la variance totale des p intrants standardisés est donnée par :

$$\sum_{j=1}^p V(X_j^*) = p = \sum_{j=1}^p V(Z_j)$$

avec $\forall i \in [1, p]$:

$$Z_i = \sum_{j=1}^p a_{ij} X_j^*$$

et $\forall i \neq j$

$$Cov(Z_i, Z_j) = Corr(Z_i, Z_j) = 0$$

Les composantes principales sont déterminées par l'analyse spectrale (i.e la recherche des valeurs propres et vecteurs propres) de la matrice des corrélations des variables intrantes. La variance de la j^{eme} composante principale Z_j est λ_j la j^{eme} valeur propre la plus grande. Les coefficients a_{ij} de la combinaison linéaire sont les vecteurs propres correspondants.

Idéalement les k premières composantes principales comprendront une part consistante de la variance totale des données. Nous pouvons par la suite utiliser ces k principales composantes Z_1, Z_2, \dots, Z_k comme variables exogènes dans le modèle de régression linéaire multiple :

$$Y = \beta_0 + \sum_{\alpha=1}^k \beta_\alpha Z_\alpha + \epsilon$$

Dans ce qui suit nous mettons en œuvre une RCP sur nos données élaguées obtenus après le traitement du chapitre précédent. Le script en langage **R**, avec lequel la RCP est automatisée, est disponible dans le dossier *Source* de notre dépôt PFE **GitHub**[15].

1. Régression en composantes principales

4.2 Mise en œuvre : corrélations croisées des variables exogènes

Nous commençons d'abord par charger les données quantitatives des pays. Il s'agit de l'élagage des données externes présentées à la section 3.2.1. Ces données représentent, de la totalité des 39 variables initiales, les 12 variables retenues en sortie de la procédure introduite par les forets aléatoires en section 3.2.2.3 avant de créer deux matrices :

- DF.actifs = Cette matrice accueillera les 12 variables retenues dans la figure 3.23.
- DF.illus = Cette matrice-vecteur accueillera notre variable de sortie qui est la **Consommation de kilogrammes de fertilisants phosphatés par hectare de terre arable**

Commençons d'abord à chercher les variables relativement colinéaires. Nous donnons dans la figure 4.1 les diagrammes de dispersions deux-à-deux croisés entre les 12 variables de notre modélisation. Nous joignons à ceux-ci des courbes splines facilitant la détection de structure ainsi qu'en gras les valeurs des coefficients de corrélations les plus importants. Plusieurs couples

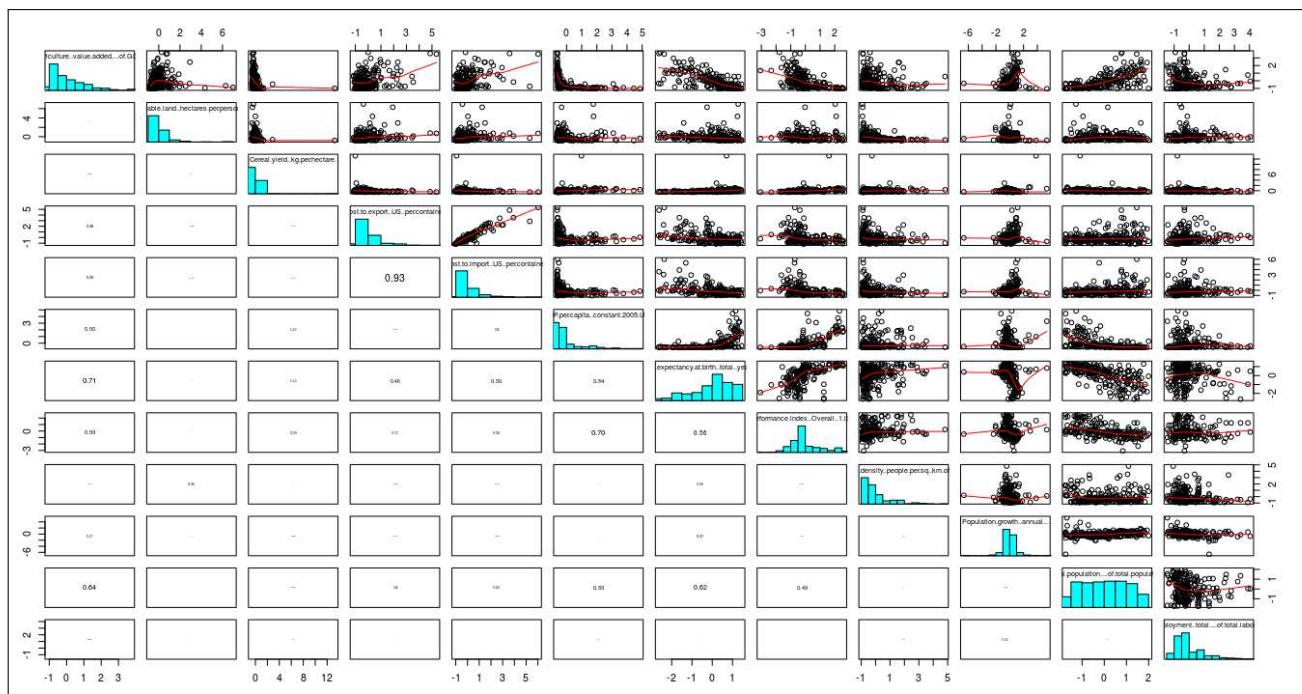


FIGURE 4.1 – Diagrammes des dispersions croisées, courbes splines et corrélations des variables exogènes

de valeurs présentent des corrélations sensiblement élevées en valeurs absolues :

- {Coût d'import conteneur Vs Coût d'export conteneur}
- {%(Agriculture) du PIB Vs Espérance de vie}
- {PIB par habitant Vs Indice de performance logistique}
- {%(Agriculture) du PIB Vs %(Population rurale) de la population totale}
- {Indice de performance logistique Vs Espérance de vie}

- Certaines de ces variables entre elles par transition des couple ci-dessus.

Une autre façon d'examiner la structure des corrélations est la création de carte coloriée dressant les corrélations deux à deux entre les variables numériques. Ceci sont *clusturisés* et ordonnés aux seins de groupes de variables corrélées. Sur la figure 4.2 nous pouvons distinguer 4 à 5 groupes de variables exogènes corrélées. Ainsi une ACP devrait produire 4 à 5 composante principales intégrant la majorité de l'information contenue dans les données.

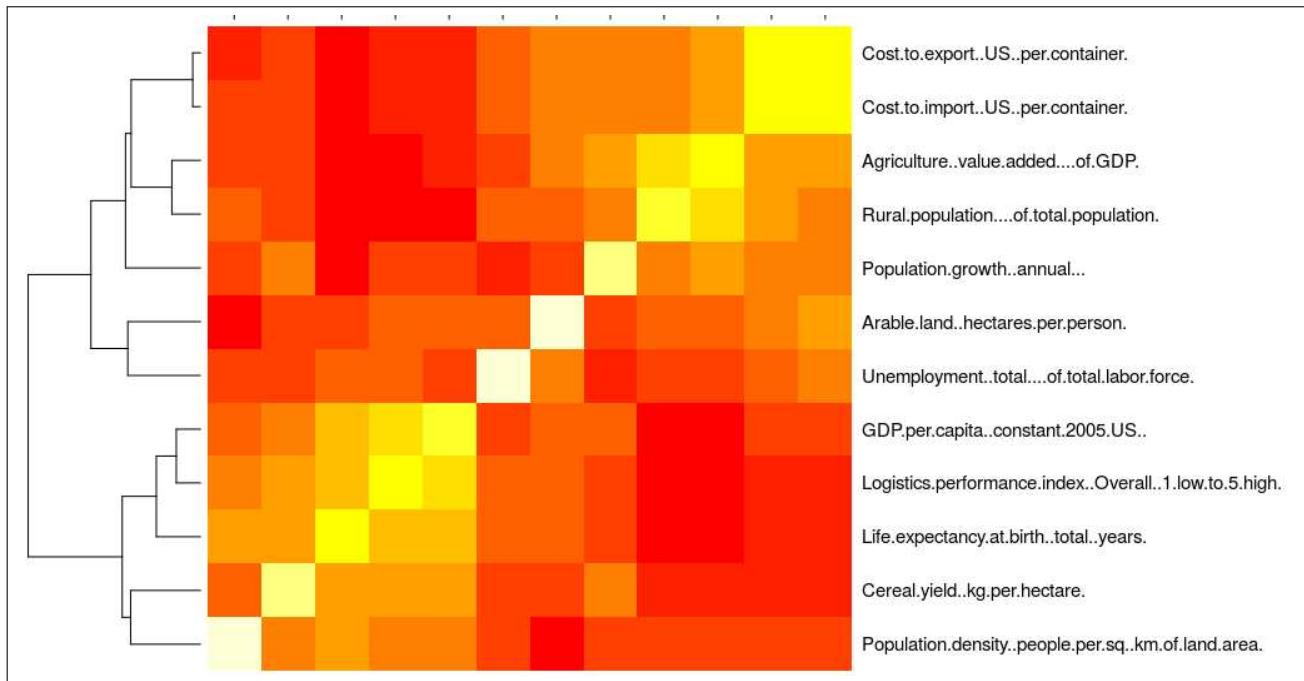


FIGURE 4.2 – Carte coloriée des regroupement des variables corrélées

Les groupes créés par le dendrogramme ci-dessous donne un avant-goût des choses à venir. Il nous tente de deviner 3 familles de variables :

- **Des indicateurs de pressions démographiques** : Les variables *Densité de population*, *PIB par habitant* et *Espérance de vie* communiquent sur la concentration, le pouvoir d'achat et la qualité de vie (et donc de nourriture) des gens dans un pays en question. Puisque le pouvoir d'achat ainsi que l'approvisionnement continu en matières premières est possible (*Indice de la performance logistique*), des pays réalisant de hauts scores dans les variables de cette famille, devraient en toute intuition, mener à une culture intensive des terres à dispositions-*Rendement des céréales en kg par hectare*. Ces 5 variables sont rassemblées par l'arbre inférieur de niveau 1 sur le dendrogramme de la figure 4.2
- **Des indicateurs de la structure industrielle** : La tendance d'un pays à importer de la marchandise dépend du coût des imports relativement à leur production chez soi. Plus la part de *L'agriculture dans le PIB* augmente, moins le pays est industrialisé. De pair avec la *Fraction des populations rurales des habitants*, ceci informerait potentiellement sur la prépondérance de l'agriculture comme secteur clé ou secondaires au sein du pays. Ces 3 variables sont rassemblées par l'arbre supérieur de niveau 3 sur le dendrogramme de la figure 4.2

- **Des indicateurs de pressions écologiques :** Les *surface arables par personne* est un rapport essentiel dans notre analyse. Il est évident que pour un pays disposant de peu de terres pour subvenir aux besoins alimentaires d'une grande population, une utilisation intensive d'engrais est à prévoir. L'arbre intermédiaire de niveau 2 du dendrogramme de la figure 4.2 qui fait intervenir *le taux de chômage* nous incite à nous poser la question de la capacité du consommateur à payer les frais d'une utilisation massive d'engrais pour sa nourriture.

Ces hypothèses ne sont pas forcément significatives, mais nous aurons dépoussiéré des pistes de réflexions pour élucider éventuellement les contributions de chaque variable au sein des CP en section 4.6 que nous calculerons dans le paragraphe suivant :

4.3 Mise en œuvre : calcul des composantes principales des données élaguées

Nous commençons par effectuer une décomposition spectrale de la matrice de corrélation. La figure 4.3 ci-dessous liste les valeurs propres résultant de cette décomposition. Alors que la figure 4.4 résume leur Scree-plot.

```
> eigenDF$values
[1] 4.17502264 1.63986342 1.49900122 1.09320225 0.96593529 0.66743517 0.58671614 0.48559214
[9] 0.34768265 0.27163069 0.20500463 0.06291375
>
```

FIGURE 4.3 – Valeurs propres de la décomposition spectrale de la matrice de corrélation

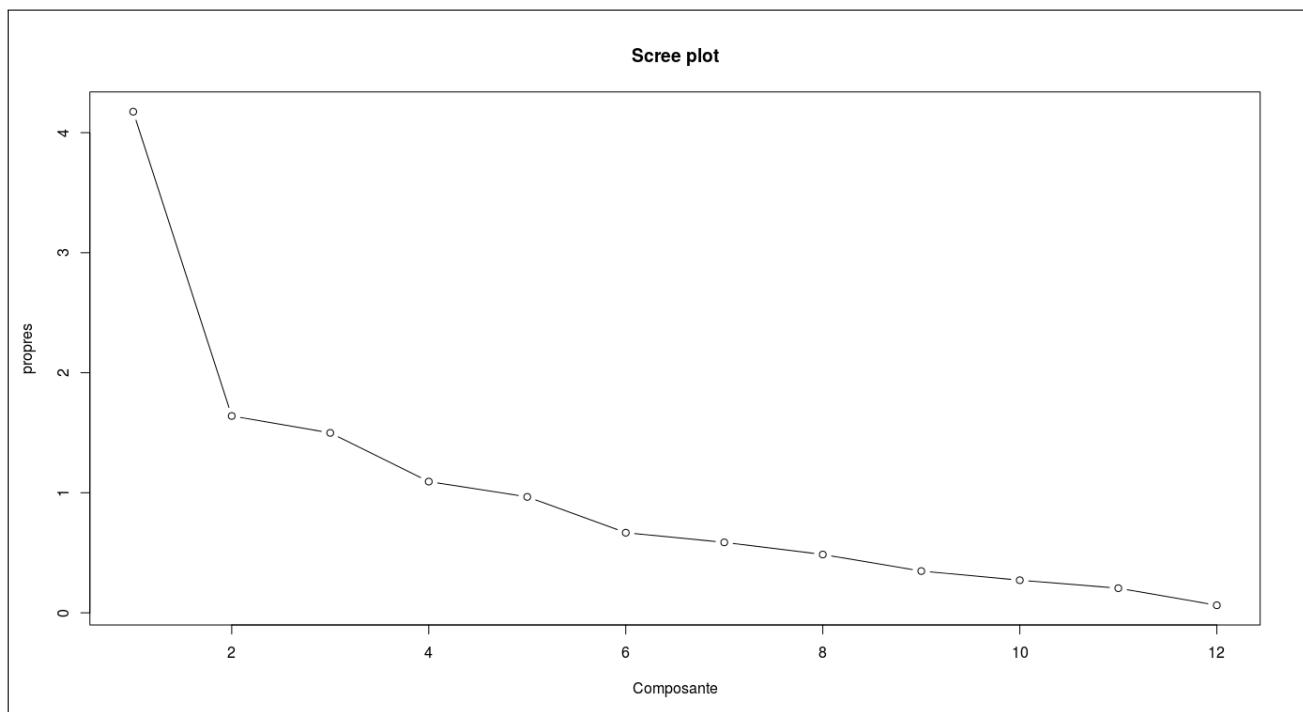


FIGURE 4.4 – Graphe des éboulis de la décomposition spectrale de la matrice de corrélation

La figure 4.5 suivante montre que 70% et 78% de l'inertie de l'ensemble des données est expliquée par les 4 et 5 premières CP respectivement.

```
> sum(eigenDF$values)
[1] 12
> sum(eigenDF$values[1:4]/12)
[1] 0.7005908
> sum(eigenDF$values[1:5]/12)
[1] 0.7810854
```

FIGURE 4.5 – Inertie expliquée par les 5 premières CP

Puisque seulement les 4 premières CP possèdent des variances supérieures à une variable normée, i.e. il y'a 4 valeurs propres > 1 . Nous procédons à la création des termes des combinaisons linéaires Z_1, Z_2, Z_3, Z_4 en utilisant les vecteurs propres correspondant. Les 4 facteurs ainsi créés, nous établissons leur corrélation à l'égard des variables intrantes de bases. La figure 4.6 donne le résultat des corrélations variables-facteurs.

	Z1	Z2	Z3	Z4
Agriculture..value.added....of.GDP.	-0.79	-0.2355	0.1344	0.155
Arable.land..hectares.per.person.	-0.16	0.4022	-0.4912	0.519
Cereal.yield..kg.per.hectare.	0.36	-0.3916	-0.1640	0.184
Cost.to.export..US..per.container.	-0.65	0.0513	-0.5990	-0.401
Cost.to.import..US..per.container.	-0.68	-0.0016	-0.5572	-0.400
GDP.per.capita..constant.2005.US..	0.70	-0.1081	-0.4583	-0.021
Life.expectancy.at.birth..total..years.	0.86	0.1351	-0.0141	-0.077
Logistics.performance.index..Overall..1.low.to.5.high..	0.74	0.0066	-0.3190	0.065
Population.density..people.per.sq..km.of.land.area.	0.31	-0.2227	0.3987	-0.570
Population.growth..annual...	-0.27	-0.7599	-0.2517	0.119
Rural.population....of.total.population.	-0.72	-0.1520	0.3233	0.145
Unemployment..total....of.total.labor.force.	-0.01	0.5970	-0.0044	-0.258
Fertilizer.consumption..kilograms.per.hectare.of.arable.land.	0.40	-0.5726	-0.1721	-0.117
>				

FIGURE 4.6 – Corrélations variables-facteurs

1. Le facteur Z1 est :

- Négativement corrélé avec :
 - La part de l'agriculture dans le PIB
 - Les coûts d'import et exports
 - La fraction rurale de la population
 - La part de l'agriculture dans le PIB
- Positivement corrélé avec :
 - L'espérance de vie
 - L'indice de la performance logistique
 - Le PIB par habitant

Z1 varie le plus entre les pays les plus riches, avec les meilleurs standards de la vie et les plus industrialisés d'une part, et d'une autre part les pays les plus pauvres, les plus dépendant de la culture agricole vivrière pour la nourriture et où les conditions de vie sont très durs. Comme le montre la carte des individus 4.7, sur les valeurs positives, nous retrouvons les pays européens et sur les valeurs négatives, nous retrouvons les pays africains subsahariens pour la plupart. La consommation de fertilisants est positivement corrélée indiquant ainsi une tendance des pays riches, et non dépendants sur l'agriculture économiquement à utiliser plus d'engrais contrairement aux pays pauvres agricoles. Les fertilisants nécessitent des moyens financiers pour encourager la consommation. Le besoin environnemental ne suffit pas.

2. Le facteur Z2 est :

- Négativement corrélé avec :
 - La croissance démographique annuelle.
 - La consommation de fertilisants.
 - Le rendement de la production céréalière.
- Positivement corrélé avec :
 - Le nombre d'hectares arables par personne
 - Taux de chômage

Z2 varie le plus entre les pays les moins peuplés par rapport à la surface du pays et disposant d'assez de terres arables pour leur population en émigration à cause des perspectives professionnels non prometteuses d'une part, et d'autre part entre les pays où le moins de terres arables est disponible par habitant pour une économie en plein boom. Comme le montre la carte des individus 4.7, sur les valeurs positives, nous retrouvons les pays de l'ex URSS et les Balkans, et sur les valeurs négatives, nous retrouvons les pays souffrant de peu de surfaces praticables pour l'agriculture disposés en archipels d'îles, des pays désertiques et qui connaissent une croissance démographique majeure. L'économie florissante de ces derniers avec des taux de chômage marginaux pour une population grandissante, favorise largement la consommation de fertilisants puisque ceux-ci se voient renforcés par deux facteurs qui s'imposent souvent : Les moyens financiers et la pression démographique sur les ressources naturelles.

Nous nous suffirons dans notre discussion uniquement du premier plan factoriel puisque Z3 et Z4 ne sont pas corrélés à notre variable endogène : **Consommation de kilogrammes de fertilisants phosphatés par hectare de terre arable** comme présenté sur la figure 4.6.

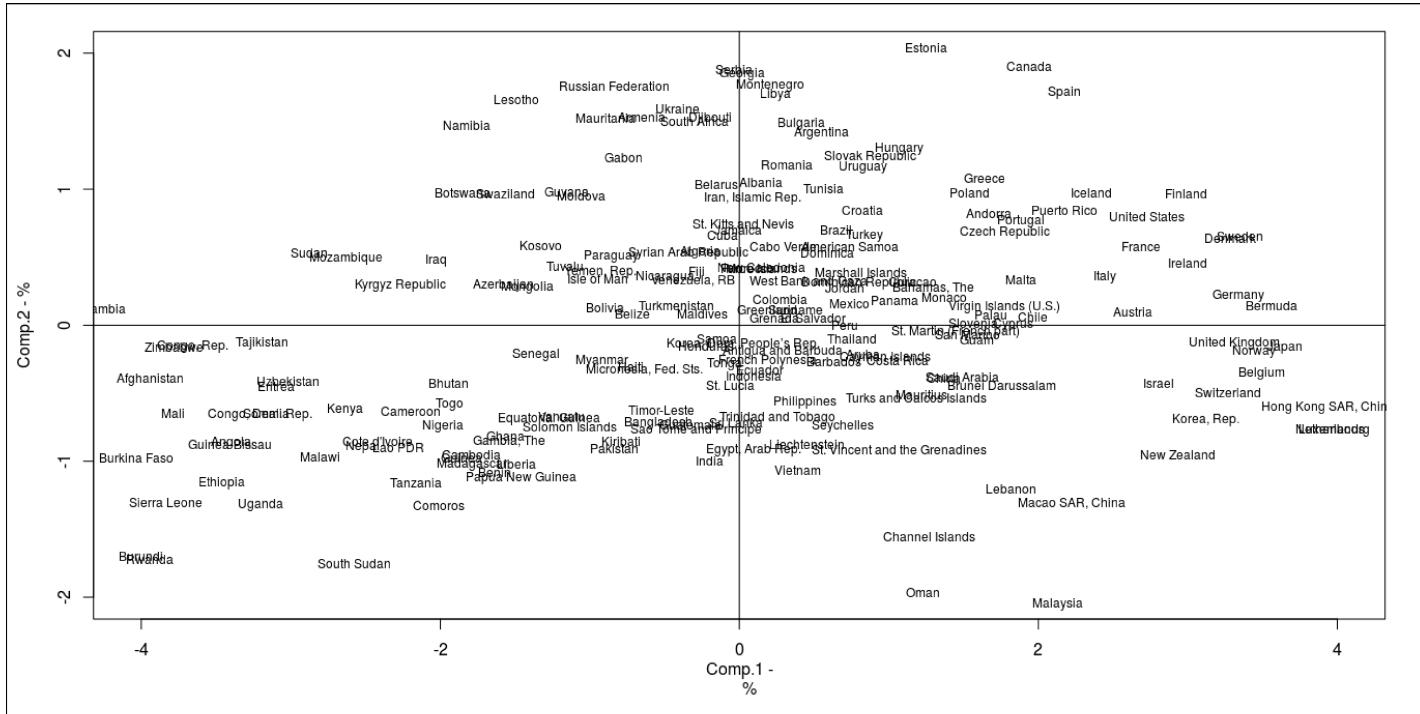


FIGURE 4.7 – Carte des individus

4.4 Mise en œuvre : régression en composantes principales

Nous créons une nouvelle matrice des données composée de la variable endogène, **Consommation de kilogrammes de fertilisants phosphatés par hectare de terre arable** et des composantes principales. La figure 4.8 ci-après présente la sortie de la matrice des corrélations croisée mais cette fois-ci entre la variable endogène et les CP.

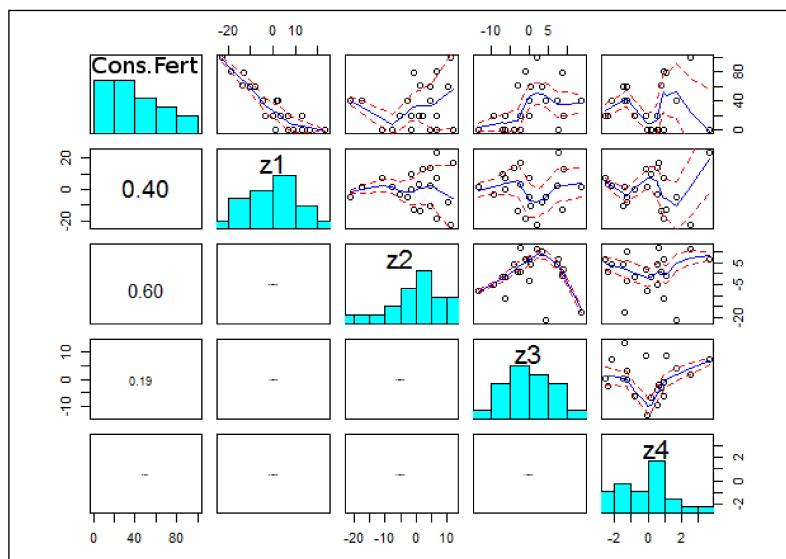


FIGURE 4.8 – Diagrammes des dispersions croisées, courbes splines et corrélations des CPs

Nous procémons par la suite à une régression linéaire multiple en utilisant les 4 CPs en

tant que variables explicatives de la **Consommation de kilogrammes de fertilisants phosphatés par hectare de terre arable**. La procédure et la sortie de cette régression est donnée par la figure 4.9 suivante :

```

> Matrice_CP = data.frame(Cons.Fert,z1,z2,z3,z4)
> Matrice_rcp = lm(Cons.Fert~.,data=Matrice_CP)
> summary(Matrice_rcp)

Call:
lm(formula = Cons.Fert ~ ., data = Matrice_CP)
eigen
Residuals:
    Min      1Q  Median      3Q     Max 
-2.106 -0.522   0.246   0.632   1.219 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 33.6200    0.2249  149.5 < 2e-16 ***
z1          -2.3062    0.0194 -118.7 < 2e-16 ***
z2           0.6602    0.0261   25.3  2.5e-14 ***
z3           1.8044    0.0341   52.8 < 2e-16 ***
z4           0.9759    0.1434    6.8  4.2e-06 ***
---
Signif. codes:  0 '****' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.03 on 16 degrees of freedom
Multiple R-squared: 0.701,    Adjusted R-squared: 0.675 
F-statistic: 4.39e+03 on 4 and 16 DF,  p-value: <2e-16

```

FIGURE 4.9 – Régression linéaire de Cons.Fert sur les CP

Les résidus que nous traçons sur la figure 4.11 se montrent remarquables compte tenu des relations non linéaires sur les graphes de la figure 4.8.

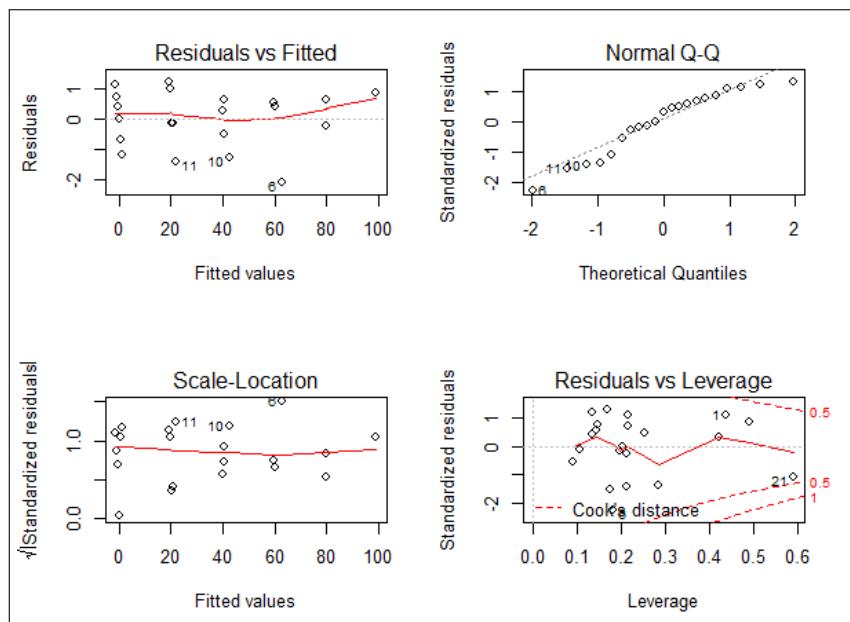


FIGURE 4.10 – Graphiques des résidus de la régression linéaire de Cons.Fert sur les CP

Nous avons retenu 7 échantillons des données originales que nous utilisons ici aux fins de validation et d'évaluation du modèle sur lequel nous réalisons un RMSE = 4.48.

```

> MatriceTest = Test[1:7,]
> ypred = predict(Matrice.rcp, newdata=MatriceTest)
> ypred
, , 4 comps
  Cons.Fert
110 50.48652
22  50.12368
31  31.82468
41  34.76234
51  30.84568
61  19.92650
71  19.78648

> MatriceTest$Cons.Fert
[1] 51.84 50.30 32.94 34.06 30.30 20.45 17.06
> sqrt(mean(ypred-MatriceTest$Cons.Fert)^2)
[1] 4.480886

```

FIGURE 4.11 – Validation du modèle et calcul du RMSE

CONCLUSION

ANNEXE A

SÉLECTION DE VARIABLES VIA FORETS ALÉATOIRE

A.1 Introduction aux forêts de décision aléatoires

Considérons un ensemble d'apprentissage $L_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$ de n observations i.i.d.¹ d'un vecteur aléatoire (X, Y) . Le vecteur $X_i = (X_i^1, \dots, X_i^p)$ contient les variables exogènes, $X_i \in \mathbb{R}^p$ et $Y_i \in \mathbb{R}$, une réponse numérique. Pour les problèmes de regression, nous supposons que $\exists S \forall i Y_i = S(X_i) + \varepsilon$ où $E[\varepsilon|X] = 0$ et S est appellée fonction de regression. Les FA est une stratégie de construction d'un modèle $\hat{e}(x)$ estimant la fonction de regression.

Les méthodes d'agrégation consistent à agréger un nombre B d'estimateurs $\hat{e}_1, \dots, \hat{e}_B$: $\hat{e}(x) = \hat{e}_B(x) = \frac{1}{B} \sum_{i=1}^B \hat{e}_i$. On considère l'erreur quadratique moyenne d'un estimateur \hat{e} et sa décomposition biais-variance :

$$E[(\hat{e}(x) - S(x))^2] = (E[\hat{e}(x)] - S(x))^2 + Var(\hat{e}(x))$$

Si on suppose les régressseurs $\hat{e}_1, \dots, \hat{e}_B$ i.i.d on a :

$$E[\hat{e}(x)] = E[\hat{e}_1(x)] \text{ et } Var(\hat{e}(x)) = \frac{1}{B} Var(\hat{e}_1(x))$$

Le biais de l'estimateur agrégé est donc le même que celui des $\hat{e}_k(x)$ mais la variance diminue. Bien entendu, en pratique il est quasiment impossible de considérer des estimateurs $\hat{e}_k(x)$ indépendants dans la mesure où ils dépendent tous du même échantillon L_n . L'approche des FA consiste à tenter d'atténuer la dépendance entre les estimateurs que l'on agrège en les construisant sur des échantillons bootstrap². Nous référerons le lecteur à l'annexe A4, pour une démonstration.

1. indépendantes identiquement distribuées
2. Ré-échantillonnage

A.2 Utilisation de CART dans la construction des Forêts aléatoires

En comparaison avec le modèle CART qui lui procède à une phase de construction de l'arbre suivie d'une phase d'élagage, deux différences sont à relever. D'abord, à chaque noeud, un nombre paramètre (dénoté k^3) de variables exogènes sont choisies aléatoirement et la meilleure variable pour la subdivision du noeud est choisie parmi celles-ci seulement. Ensuite, aucun élagage n'est introduit : tous les arbres de la forêt sont maximaux.

Le principe de CART est de partitionner récursivement l'espace engendré par les variables explicatives (ici \mathbb{R}^p) de façon dyadique. Plus précisément, à chaque étape du partitionnement, on découpe une partie de l'espace en deux sous parties selon une variable X_j comme le montre la figure A.1.

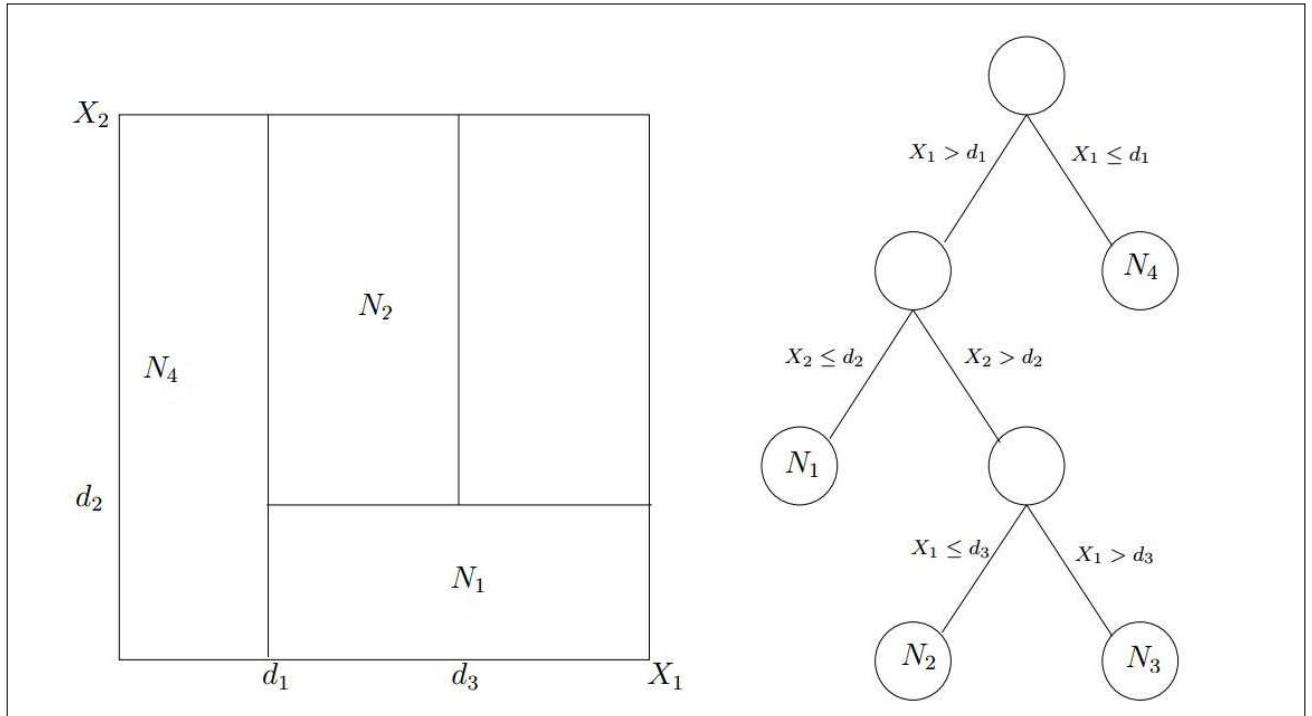


FIGURE A.1 – Arbre CART

Les coupures sont choisies de manière à minimiser une fonction de coût particulière. A chaque étape, on cherche la variable X_j et le réel d qui minimisent la variance des noeuds fils dans les problèmes de régression. Les arbres sont ainsi construits jusqu'à atteindre une règle d'arrêt. Par exemple, on ne découpe pas un noeud qui contient moins de 5 observations comme y procède le package ***randomForest*** du langage **R**.

L'agrégation à laquelle procèdent les FA est d'autant plus performante que la corrélation entre les prédicteurs agrégés (arbres CART) est faible. Afin de diminuer cette corrélation, Breiman[22] propose de rajouter une couche d'aléa dans la construction des prédicteurs. Plus précisément, à chaque étape de CART, k variables sont sélectionnées aléatoirement parmi les

3. L'autre paramètre des FA étant B , le nombre d'arbres dont la forêt.

p et la meilleure coupure est sélectionnée uniquement sur ces k variables.

On retrouve un compromis biais-variance dans le choix de m :

- lorsque k diminue, la tendance est à se rapprocher d'un choix aléatoire des variables de découpe des arbres. Dans le cas extrême où $k = 1$, les axes de la partition des arbres sont choisies au hasard, seuls les points de coupure utiliseront l'échantillon. Ainsi, si k diminue, la corrélation entre les arbres va avoir tendance à diminuer également, ce qui entraînera une baisse de la variance de l'estimateur agrégé. En revanche, choisir les axes de découpe des arbres de manière aléatoire va se traduire par une moins bonne qualité d'ajustement des arbres sur l'échantillon d'apprentissage, d'où une augmentation du biais pour chaque arbre ainsi que pour l'estimateur agrégé.
- lorsque k augmente, les phénomènes inverses se produisent.

On déduit de cette remarque que le choix de k est lié aux choix des paramètres de l'arbre, notamment au choix du nombre d'observations dans ses nœuds terminaux. En effet, si ce nombre est petit, chaque arbre aura un biais faible mais une forte variance. Il faudra dans ce cas là s'attacher à diminuer cette variance et on aura donc plutôt tendance à choisir une valeur de k relativement faible. À l'inverse, si les arbres ont un grand nombre d'observations dans leurs nœuds terminaux, ils posséderont moins de variance mais un biais plus élevé. Dans ce cas, la procédure d'agrégation se révélera moins efficace. C'est pourquoi, en pratique, le nombre maximum d'observations dans les nœuds est par défaut pris relativement petit (5). Concernant le choix de k , **randomForest** propose par défaut $k = \frac{p}{3}$ en régression. Ce paramètre peut également être sélectionné via des procédures apprentissage-validation ou validation croisée.

A.3 L'erreur Out-Of-Bag et le score FA d'importance des variables

L'erreur Out Of Bag : Il s'agit d'une procédure permettant de fournir un estimateur de l'erreur $E[(\hat{e}(x) - Y)^2]$ en régression. De tels estimateurs sont souvent construits à l'aide de méthode apprentissage-validation ou validation croisée. L'avantage de la procédure Out Of Bag (OOB) est qu'elle ne nécessite pas de découper l'échantillon. Elle utilise le fait que les arbres sont construits sur des estimateurs agrégés et que, par conséquent, ils n'utilisent pas toutes les observations de l'échantillon d'apprentissage. Étant donné une observation (X_w, Y_w) de L_n , on désigne par ω_B l'ensemble des arbres de la forêt qui ne contiennent pas cette observation dans leur échantillon bootstrap. Pour estimer, la prévision de la forêt sur Y_w on agrège uniquement ces arbres là :

$$\hat{Y}_w = \frac{1}{|\omega_B|} \sum_{i \in \omega_B} \hat{e}(X_w, \theta_i)$$

L'importance des variables : L'inconvénient des méthodes d'agrégation est que le modèle construit est difficilement interprétable, on parle souvent d'aspect boîte noire. Pour ce type de méthodes. Pour le modèle de forêts aléatoires que nous venons de présenter, Breiman[22]

propose une mesure qui permet de quantifier l'importance des variables X_j , $j=1,\dots,p$ dans le modèle. On désigne par OOB_k l'échantillon Out Of Bag associé au $k^{\text{ème}}$ arbre de la forêt. Cet échantillon est formé par les observations qui ne figurent pas dans le $k^{\text{ème}}$ échantillon bootstrap. On note E_{OOB_k} l'erreur de prédiction de l'arbre $\hat{e}(\cdot, \theta_k)$ mesurée sur cet échantillon :

$$E_{OOB_k} = \frac{1}{|OOB_k|} \sum_{i \in OOB_k} (\hat{e}(X_i, \theta_k) - Y_i)^2.$$

On désigne maintenant par OOB_k^j l'échantillon OOB_k dans lequel on a perturbé aléatoirement les valeurs de la variable j et par $E_{OOB_k^j}$ l'erreur de prédiction de l'arbre $\hat{e}(\cdot, \theta_k)$ mesurée sur cet échantillon :

$$E_{OOB_k^j} = \frac{1}{|OOB_k^j|} \sum_{i \in OOB_k^j} (\hat{e}(X_i^j, \theta_k) - Y_i)^2.$$

où les X_i^j désignent les observations perturbées de OOB_k^j . Empiriquement, si la $j^{\text{ème}}$ variable joue un rôle déterminant dans la construction de l'arbre $\hat{e}(\cdot, \theta_k)$, alors une permutation de ces valeurs j dégradera fortement l'erreur. La différence d'erreur $E_{OOB_k^j} - E_{OOB_k}$ sera alors élevée. L'importance de la $j^{\text{ème}}$ variable sur la forêt est mesurée en moyennant ces différences d'erreurs sur tous les arbres :

$$Imp(X_j) = \frac{1}{B} \sum_{k=1}^B (E_{OOB_k^j} - E_{OOB_k})$$

A.4 Étude du biais et de variance d'un estimateur agrégé [14]

Comparons le biais et la variance de l'estimateur agrégé à ceux des estimateurs que l'on agrège. Le fait de considérer des échantillons bootstrap introduit un aléa supplémentaire dans l'estimateur. Afin de prendre en compte cette nouvelle source d'aléatoire, on note $\theta_k = \theta_k(L_n)$ l'échantillon bootstrap de k variables exogènes à l'étape i et $\hat{e}(\theta_k)$ l'estimateur construit. On écrira l'estimateur final $\hat{e}_B(x) = \frac{1}{B} \sum_{i=1}^B \hat{e}_i(x, \theta_k)$

Les tirages bootstrap sont effectués de la même manière et indépendamment les uns des autres. Ainsi, conditionnellement à L_n , les variables $\theta_1, \dots, \theta_B$ sont i.i.d. et de même loi que θ (qui représentera la loi de la variable de tirage de l'échantillon bootstrap). Ainsi, d'après la loi des grands nombres :

$$\hat{e}(x) = \lim_{B \rightarrow \infty} \hat{e}_B(x) = \lim_{B \rightarrow \infty} \frac{1}{B} \sum_{i=1}^B \hat{e}_i(x, \theta_k) = E_\theta[\hat{e}(x, \theta) | L_n]$$

L'espérance est ici calculée par rapport à la loi de θ . Prendre B trop grand ne va pas surajuster l'échantillon, autrement dit prendre la limite en B revient à considérer un estimateur "moyen" calculé sur tous les échantillons bootstrap. Le choix de B n'est donc pas crucial pour la performance de l'estimateur, il est recommandé de le prendre le plus grand possible. On note :

- $T_k = \hat{e}(x, \theta_k)$, les estimateurs que l'on agrège, ceux-ci étant identiquement distribués.
- $\sigma^2(x) = Var(T_k)$, la variance des estimateurs que l'on agrège.
- $\rho(x) = corr[T_1, T_2]$, le coefficient de corrélation entre deux estimateurs que l'on agrège (calculés sur deux échantillons bootstrap).

La variance $\sigma^2(x)$ et la corrélation $\rho(x)$ sont calculées par rapport aux lois de L_n et de θ . On suppose que les estimateurs sont identiquement distribués. Il est alors facile de voir que le biais de l'estimateur agrégé est le même que le biais des estimateurs que l'on agrège. Par conséquent, agréger ne modifie pas le biais. Pour la variance, on a le résultat suivant :

$$\begin{aligned} Var(\hat{e}_B(x)) &= Var\left[\frac{1}{B} \sum_{i=1}^B T_i\right] = \frac{1}{B^2} \left[\sum_{i=1}^B Var(T_i) + \sum_{1 \leq i \neq i' \leq B} cov(T_i, T_{i'}) \right] \\ &= \frac{1}{B} \sigma^2(x) + \frac{1}{B^2} [B^2 - B] \rho(x) \sigma^2(x) = \rho(x) \sigma^2(x) + \frac{1 - \rho(x)}{B} \sigma^2(x) \end{aligned}$$

Ainsi, si $\rho(x) < 1$, l'estimateur agrégé a une variance plus petite que celle des estimateurs que l'on agrège (pour B suffisamment grand). On déduit que c'est la corrélation $\rho(x)$ entre les estimateurs que l'on agrège qui quantifie le gain de la procédure d'agrégation : la variance diminuera d'autant plus que les estimateurs que l'on agrège seront décorrélatés. Le fait de construire les estimateurs sur des échantillons bootstrap va dans ce sens.

BIBLIOGRAPHIE

- [1] NACER A. Structuration, extraction et analyse de l'information qualitative et quantitative externe reçue par la direction commerciale, ENSIAS, Juin 2015. PFE BI.
- [2] Vance A. Climate corp. updates crop insurance via high tech. <http://goo.gl/6CzvaA>, consulté le 09 Avril 2016.
- [3] NAVIOS MARITIME ACQUISITION CORP. Form f-1. <http://goo.gl/BCStQc>, consulté le 09 Juin 2016.
- [4] IBM Corporation. Guide crisp-dm de ibm spss modeler. <ftp://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/15.0/fr/CRISP-DM.pdf>.
- [5] Pyle D. *Data Preparation for Data Mining*. 1999.
- [6] Vail E. Prices of fertilizer materials and factors affecting the fertilizer tonnage. Master's thesis, Cornell University), 1927. Thèse de doctorat.
- [7] McAfee A. et Brynjolfsson E. Big data : The management revolution. *Harvard Business Review*, Octobre 2012.
- [8] Biau G. et Devroye L. et Lugosi G. *Consistency of random forests and other averaging classifiers*. Journal of Machine Learning Research. 9, 2039- 2057.
- [9] Smith D.L et Dhavala S.P. Using big data for decisions in agricultural supply chain. Master's thesis, Massachusetts Institute of Technology (MIT), 2013. Master of Engineering in Logistics.
- [10] Piatetsky-Shapiro G. et Frawley W.J. Knowledge discovery in databases. *AAAI Press*, 1991.
- [11] Breiman L. et Friedman J.H. et Olshen R.A. et Stone C.J. *Classification And Regression Trees*. 1984.
- [12] Brynjolfsson E. et Hitt L.M. et Kim H.H. Strength in numbers : How does data-driven decisionmaking affect firm performance ? http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1819486.

- [13] Mehring A. et Shaw B. Relationships between farm income and farmers' expenditures for fertilizer and a forecast of the commercial demand for fertilizer in 1944 and 1945, by states. *American Fertilizer*, 1944.
- [14] Hastie T. et Tibshirani R. et Friedman J. *The Elements of Statistical Learning : Data Mining, Inference, and Prediction*.
- [15] Sefiane H. et Yachoui A. Analyse de la concurrence du groupe ocp sur le marché des phosphates et produits dérivés. <http://github.com/sanxore/lpr>, ENSIAS, Juin 2016. Rapport de PFE BI et livrables code-source.
- [16] Tenkorang F. Projecting world fertilizer demand in 2015 and 2030. <http://goo.gl/r7CgbL>.
- [17] FAO. Fertilizer requirements in 2015 and 2030. *Food and Agriculture Organization of the United Nations Outlook*, 2000.
- [18] Piatetsky-Shapiro G. Kdnuggets methodology poll. <http://www.kdnuggets.com/polls/2014/Analytics-data-mining-data-science-methodology.html>.
- [19] OCP Group. Discover our subsidiaries and joint ventures. <http://www.ocpgroup.ma/group/group-overview/main-companies>.
- [20] CHEMLAL I.E. Mise en place d'un portail décisionnel, à base de fichiers pdf, dédié à la direction commerciale, ENSIAS, Juin 2014. PFE BI.
- [21] Phosphate Chemicals Export Association Inc. Dap contract. <http://goo.gl/yHt9UB>, consulté le 09 Juin 2016.
- [22] Breiman L. *Random Forests. Machine Learning*. 45, 5-32. 2001.
- [23] LOTS Shipping Limited. Gearless panamax lighterage in cochin. <https://goo.gl/i03EZL>, consulté le 09 Juin 2016.
- [24] Parthasarathy N. S. Demand forecasting for fertilizer marketing. *Food and Agriculture Organization of the United Nations Outlook*, 1994.
- [25] Dietterich T. *An experimental comparison of three methods for constructing ensembles of decision trees : Bagging, Boosting and randomization*. 1999.
- [26] Shinyama Y. Pdfminer, python pdf parser and analyzer. <http://www.unixuser.org/~euske/python/pdfminer/>, consulté le 22 Mars 2016.
- [27] Griliches Z. The demand for fertilizer : An economic interpretation of a technical change. *Journal of Farm Economics*, 1958.